UNIVERSITÉ DE STRASBOURG

Observatoire astronomique de Strasbourg

Thèse pour l'obtention du grade de
Docteur ingénieur
de l'Université de Strasbourg

Spécialité Astrophysique

Par François-Xavier PINEAU

Analyse statistique du catalogue de sources X cosmiques du satellite ESA XMM-Newton

Soutenue publiquement le 23 octobre 2009

Membres du jury

Directeur de thèse :	M. Christian MOTCH, Directeur de recherche, Observatoire de Strasbourg
Rapporteur Interne :	M. Mark ALLEN, Chargé de recherche CNRS, Observatoire de Strasbourg
Rapporteur Externe :	M. Jean BALLET, Physicien, CEA Saclay
Rapporteur Externe :	M. Didier PELAT, Professeur, Observatoire de Paris Meudon
Examinateur :	M. Mike WATSON, Professeur, Université de Leicester (Angleterre)
Co-encadrant et Examinateur :	M. Sébastien DERRIERE, Astronome adjoint, Observatoire de Strasbourg
Co-encadrant et Membre Invité :	M. Laurent MICHEL, Ingénieur de recherche, Observatoire de Strasbourg

Résumé

XMM Newton est un satellite de l'ESA lancé en 1999. Il observe le ciel dans le domaine des rayons X. Le *Survey Science Center* (SSC) d'XMM est en charge de l'exploitation scientifique du télescope spatial et a créé à partir de ses observation le plus grand catalogue de sources X à ce jour. La classification des sources de ce catalogue est une des tâches attribuées au SSC. C'est dans ce cadre que s'insère ce travail de thèse.
Nous avons développé et implémenté un algorithme de corrélation croisée permettant d'obtenir des informations multi-longueur d'onde pour un certain nombres de sources du catalogue 2XMMi. Cet algorithme utilise une approche Bayesienne pour fournir des probabiltés d'identification. Il utilise également une approche originale permettant de se passer de simulations de Monte Carlo pour l'estimation du nombre de fausses associations. Nous avons implémenté et utilisé une analyse en composantes principale (ACP) qui prend en compte les erreurs de mesure de chaque observations. Cette ACP nous a permis d'explorer l'espace des paramètres multi-longueur d'onde et de choisir un nombre restreint de dimensions en entrée d'algorithme de classification. Nous avons utilisé les données de l'ACDS et les résultats de la corrélation entre le 2XMMi et le SDSS DR7 pour créér un échantillon d'identification. Cette échantillon est utilisé pour la création d'échantillons d'apprentissage qui sont indispensables pour la classification supervisée. Nous avons classé les sources X de plusieurs échantillons : celui issu des résultats de la corrélation 2XMMi/SDSS DR7, celui issu de la corrélation 2XMMi/GSC2/2MASS et un échantillon de sources purement X. L'algorithme de classification a été choisi après avoir comparé différentes méthodes. Enfin, nos résultats ont été insérés dans la XCAT-DB pour une large diffusion.

Abstract

XMM Newton is a European satellite designed by ESA and launched in 1999. It scans the sky in the X-ray wavelengths. The XMM Survey Science Centre (SSC) is in charge of the scientific exploitation of the spatial telescope and has built from its observations the larger catalogue of X-ray sources ever published so far. The work presented here has been carried out in this framework.
We designed and implemented a cross-correlation algorithm that allows to retrieve multi-wavelength data for a fraction of 2XMMi catalogue sources. This algorithm is based on a Bayesian approach providing probabilities of identification. It also uses a new way to compute the rates of spurious associations without resorting to Monte Carlo simulations. We implemented and applied a principal component analysis (PCA) taking into account the measurement errors of each source. That modified PCA has been used to explore the multi-wavelength parameter space and to restrict the number of dimensions in input of the classification algorithms. ACDS data has been used together with the results of the cross-correlation of the 2XMMi with the SDSS DR7 to build an identification sample. We derived learning samples from this sample. Learning samples are essential to perform supervised classification. X-ray sources of several samples have been classified. Those samples are the 2XMMi/SDSS DR7, coming from the cross-correlation of the two catalogues, the 2XMMi/GSC2/2MASS and one only containing XMM X-ray sources. The classification algorithm used has been chosen from the results of tests performed and several methods such as decision trees, Kohonen neural networks, meanshifts or kernel density classification. Finally, we put our results into the XCAT-DB so that they are available for the rest of astronomy community.

Table des matières

Introduction **1**

I L'astronomie X et XMM-Newton **3**

1 Les différents émetteurs X 4
1.1 Origine des rayons X 4
1.1.1 Rayonnement du corps noir 4
1.1.2 Émission thermique de gaz chaud : le bremsstrahlung thermique . . . 5
1.1.3 Émission synchrotron : émission en loi de puissance 6
1.1.4 Effet Compton Inverse : émission en loi de puissance 6
1.1.5 Caractéristiques observationnelles 7
1.2 Des sources X dans le système solaire 8
1.3 L'émission X des étoiles 8
1.3.1 Rappel sur la classification spectrale des étoiles 8
1.3.2 Les étoiles pré-séquence principale 9
1.3.3 Les étoiles de la séquence principale 10
1.3.4 Les étoiles évoluées 13
1.3.5 Les étoiles binaires serrées non-accrétantes 13
1.3.6 Les étoiles post-séquence principale 14
1.4 Les binaires X (XRB) 16
1.4.1 Pulsars accrétants 20
1.4.2 Étoiles à neutrons faiblement magnétisées et trous noirs 20
1.4.3 Sursaut X de type I 21
1.4.4 Sources X ultra-lumineuses (ULX) 21
1.5 Variables cataclysmiques (CV) 22
1.5.1 Généralités sur les CV non magnétisées 22
1.5.2 Les CV magnétisées : Polars et Polars intermédiaires 23
1.5.3 Sources X super-soft (SSS) 24
1.6 Restes de supernovae 25
1.7 Galaxies, AGN et Amas de galaxies 26
1.7.1 Les galaxies 26
1.7.2 Noyaux actifs de galaxies (AGN) 28
1.7.3 Les Amas de galaxies 30
1.8 Le fond diffus X 30
1.8.1 Le fond diffus X galactique 30
1.8.2 Le fond diffus X extragalactique 30
2 De XMM-Newton au catalogue 2XMMi 31
2.1 L'observatoire spatial XMM-Newton 31
2.2 Du satellite au catalogue 2XMMi 32
2.3 Le 2XMMi en chiffres 34
2.3.1 Statistiques globales 34
2.3.2 Les grandeurs physiques 34
2.3.3 Les filtres 36
2.3.4 Correction des positions 36

Références 38

II Corrélation croisée des sources XMM **43**

1 Cadre général du problème et simplifications 43
1.1 Problèmes de positionnement 44
1.2 Des origines de la différence des sources observées 44
1.3 Superposition X de sources résolues en optique 45
2 Algorithme de corrélation croisée mis en place 45
2.1 Critère de sélection des candidats optiques 45
2.2 Le rapport de vraisemblance, ou *likelihood ratio* 46
2.3 Estimation de la densité locale 48
2.4 Calcul des fiabilités 48
2.4.1 Définition de la fiabilité 48
2.4.2 Estimation du taux de fausses associations 49
2.4.3 Calculs en pratique 50
2.4.4 Ajustement de la fonction de fiabilité 51
2.5 Calcul des probabilités d'identification dans le cas général 52
2.6 Tests et implémentation 53
2.6.1 Groupement d'observations 53
2.6.2 Tests sur des données simulées 53
2.6.3 Implémentation 53
3 Approfondissement de quelques aspects techniques 54
3.1 Référentiel pour la convolution des ellipses d'erreur 54
3.1.1 Passage d'un repère local sphérique à un repère local plan 55
3.1.2 Erreurs associées au passage à un repère plan 58
3.2 Convolution des ellipses d'erreur et distance en σ 59
3.2.1 Calcul sur les ellipses 59
3.2.2 Ellipse, matrice de variance-covariance et changement de repère 60
3.2.3 Approche matricielle 63
3.2.4 Distance en σ de deux sources 64
3.3 Origines du rapport de vraisemblance et probabilité d'identification 64
3.3.1 Origines du rapport de vraisemblance 64
3.3.2 Généralisation des probabilité Bayesiennes 65
3.3.3 Probabilité d'identification de Rutledge 66
3.4 Estimation du paramètre de la loi de poisson par k-NN 67
4 Recalibration des positions des sources XMM 67
4.1 Méthode générale de recalibration de positions 68
4.1.1 Le problème 68
4.1.2 Le système 68
4.2 Correction du positionnement des sources XMM 69
4.2.1 Erreur de positionnement des sources XMM 69
4.2.2 Généralisation de la méthode de Levenberg-Marquardt 70
5 Estimation de densités locales 72
5.1 Histogrammes et boites glissantes 72
5.1.1 Généralités 72
5.1.2 Image 3D 72
5.2 Les k plus proches voisins 75
5.3 La tesselation de Voronoï 75
5.4 Les ondelettes et les SOM 76
5.5 Le kernel smoothing 76
5.5.1 Description 76
5.5.2 Détermination de la largeur de bande optimale par méthode de Monte Carlo Markov Chain 80
Références 83

III Analyse de données et classification de sources X **87**

1 L'Analyse en Composantes Principales (ACP) . . . 87
1.1 L'ACP classique . . . 87
1.1.1 Principe . . . 88
1.1.2 Algorithme . . . 88
1.1.3 Nombre de composantes principales à conserver . . . 89
1.1.4 Lien entre vecteurs propres et vecteurs d'origine . . . 89
1.2 L'ACP modifiée . . . 89
1.2.1 ACP robustes . . . 90
1.2.2 Bases statistiques de l'ACP modifiée . . . 92
1.2.3 Estimation de la moyenne et de la variance . . . 93
1.2.4 Hétéroscédasticité . . . 95
1.2.5 Problèmes liés à l'ACP en vue d'une classification . . . 96
1.3 La "Kernel PCA" . . . 96
1.3.1 Principe . . . 96
1.3.2 Algorithme . . . 97
1.3.3 Implémentation . . . 98
1.3.4 Validation de notre implémentation . . . 99
1.3.5 Résultats et analyse . . . 99
2 Réseaux de neurones . . . 100
2.1 Introduction . . . 100
2.2 Les cartes auto-organisatrices de Kohonen . . . 101
2.2.1 Description de l'algorithme . . . 102
2.2.2 Les traceurs T1 et T2 . . . 102
2.2.3 Les fonctions d'apprentissage et de voisinage . . . 103
2.2.4 Exemple test . . . 104
2.2.5 Quelques résultats et exemples d'utilisation . . . 104
3 Les k-means . . . 106
4 Le *mean shift* . . . 107
5 Les arbres de décision . . . 108
6 Les Support Vector Machines . . . 108
6.1 Cas idéal . . . 109
6.2 Cas moins idéal . . . 110
6.3 En pratique . . . 110
6.4 Remarques . . . 110
7 Kernel density classification . . . 111
7.1 Principe . . . 111
7.2 Paramètres à fixer . . . 111
7.2.1 Les paramètres du kernel smoothing . . . 111
7.2.2 Le nombre d'éléments de chaque classe . . . 111
8 Conclusion . . . 112
Références . . . 114

IV Résultats **117**

1 Corrélation du 2XMMi avec le SDSS DR7 . . . 117
1.1 Résultats de la corrélation . . . 117
1.1.1 Quelques chiffres . . . 117
1.1.2 Ajustement d'une fonction de Rayleigh . . . 118
1.1.3 Calcul de la luminosité X . . . 119
1.2 Constitution d'un échantillon d'apprentissage . . . 119
1.2.1 Constitution globlale . . . 121
1.2.2 *Kernel density classification* pour les étoiles du SDSS . . . 123
1.2.3 Composition de l'échantillon d'apprentissage pour les données XMM/DR7126
1.3 Discussion scientifique . . . 127
1.3.1 Mise en évidence des principales classes des sources X . . . 127

1.3.2 Étoiles actives en X . . . 130
1.3.3 Galaxies actives et quasars . . . 134
1.4 ACP sur l'échantillon XMM/SDSS . . . 142
1.5 Classification de l'échantillon XMM/DR7 . . . 143
2 Corrélations du 2XMMi, du 2MASS et GSC2.2.1 . . . 144
2.1 Création de l'échantillon 2XMMi-GSC2-2MASS . . . 144
2.1.1 Corrélation du 2MASS . . . 144
2.1.2 Corrélation du GSC2 . . . 145
2.1.3 Échantillon 2XMMi-GSC2-2MASS . . . 145
2.2 Analyse sommaire de l'échantillon . . . 146
2.3 Résultat de la Classification . . . 148
3 Classification sur les propriétés X uniquement . . . 150
4 Comparaison de différents algorithmes de classification . . . 153
4.1 Jeux de données en entrée . . . 153
4.2 Grandeurs mesurées . . . 153
4.3 Algorithmes testés . . . 155
4.4 Résultats . . . 156
4.4.1 Analyse globale des résultats sur le premier jeu de données . . . 156
4.4.2 Analyse algorithme par algorithme . . . 156
4.5 Conclusions . . . 157
4.6 Perspectives : proposition de classification . . . 166
Références . . . 167

V Interactions avec la XCAT-DB 171

1 Présentation de la XCAT-DB et de SAADA . . . 171
1.1 La XCAT-DB . . . 171
1.2 SAADA . . . 172
1.3 Genèse, modèle de données de SAADA et unicité des sources . . . 173
1.3.1 La genèse . . . 173
1.3.2 Le modèle de données de SAADA . . . 173
1.3.3 Unicité des sources d'archive . . . 174
2 Module de corrélation et XCAT-DB . . . 176
2.1 Module de corrélation . . . 176
2.1.1 La « base de connaissances » . . . 176
2.1.2 Structure de la base de données . . . 178
2.1.3 Actions et entrées-sorties . . . 179
2.1.4 Séquence des opérations de corrélation . . . 180
2.2 Interactions entre module de corrélation et XCAT-DB . . . 181
2.2.1 Calcul des corrélations . . . 182
2.2.2 Mise à jour des relations . . . 182
2.2.3 Insertion de nouveaux catalogues . . . 184
2.2.4 Gestion des erreurs . . . 184
2.2.5 Schéma d'ensemble . . . 185
3 Classification et XCAT-DB . . . 185
4 Accès aux données hebergé dans la XCAT-DB . . . 186
Références . . . 189

Conclusion & Perspectives 191

Références . . . 195

Annexes 197

A Les Quaternions 199

B Logiciel de création de SOM **201**

C Liste des travaux publiés ou en cours de publication **203**

D Proceeding ADASS 2007 **205**

E Proceeding ADASS 2008 **211**

F Proceeding Ringberg Castle 2008 **217**

G Proceeding EURO-VO 2008 **227**

H Article en préparation sur la corrélation 2XMMi/DR7 **237**

I Demandes de temps télescope **261**

1 Demande de temps sur le Calar Alto 261
2 Demande de temps sur l'OHP 270

Index **279**

Table des figures

I.1 Spectres de corps noirs . . . 5
I.2 Spectres d'émissions Bremsstrahlung . . . 6
I.3 Spectres en lois de puissances . . . 7
I.4 Diagramme de Hertzsprung-Russell . . . 9
I.5 Classification spectrale de Morgan-Keenan . . . 11
I.6 Spectre EPIC d'une étoile de type G0V . . . 11
I.7 Spectre EPIC de deux naines blanches isolées . . . 15
I.8 Spectres EPIC d'une étoile à neutron isolée . . . 16
I.9 Schéma synthétique d'un système binaire X . . . 17
I.10 Binaires X à trous noirs . . . 18
I.11 Spectres EPIC de 4 XRB . . . 19
I.12 Spectre EPIC de l'ULX NGC 1313 X-1 . . . 22
I.13 Spectres EPIC de 2 CV. . . . 24
I.14 Spectres EPIC d'un polar et d'un polar intermédiaire . . . 24
I.15 Spectres EPIC de deux SSS. . . . 25
I.16 Spectre EPIC de 2 SNR . . . 26
I.17 Spectre SDSS d'une galaxie elliptique . . . 27
I.18 Spectres EPIC et SDSS d'une galaxie *starburst* . . . 27
I.19 Spectres EPIC et SDSS d'une galaxie *star forming* . . . 28
I.20 Spectres EPIC et SDSS d'un QSO . . . 29
I.21 Spectres EPIC de 2 LINER . . . 30
I.22 Le satellite XMM-Newton et ses trois miroirs . . . 31
I.23 Capteurs CCD des caméras EPIC MOS et PN . . . 32
I.24 Segment sol d'XMM-Newton . . . 33
I.25 Interface XCAT-DB . . . 34
I.26 Erreurs de position des sources du 2XMMi . . . 35
I.27 Flux soft et hard des sources du 2XMMi . . . 35

II.1 Plan de projection choisi pour la convolution des erreurs sur les positions . . . 47
II.2 k-NN et *kernel smoothing* sur des données simulées . . . 49
II.3 Histogramme et courbe de fiabilité . . . 52
II.4 Fiabilités réelles et estimées à partir de données simulées . . . 54
II.5 Triangle quelconque sur une surface sphérique . . . 55
II.6 Plan tangent à la sphère unité . . . 56
II.7 Plan tangent à la sphère unité (2) . . . 57
II.8 Repère $(X, \vec{Nord}, \vec{Est})$ . . . 58
II.9 Schéma représentant la correction du référentiel d'un FOV . . . 69
II.10 Image 3D et des boîtes glissantes . . . 73
II.11 Masque sur une image 3D . . . 73
II.12 *Kernel smoothing* 1D . . . 77
II.13 Les trois fonctions noyaux usuels en 1D . . . 79
II.14 Exemples de résultats d'une MCMC . . . 83

III.1 Corrélations entre axes principaux d'une ACP et paramètres d'entrée 90
III.2 Deux exemples d'hétéroscédasticité 95
III.3 KPCA sur un échantillon synthétique 99
III.4 KPCA sur l'échantillon XMM/DR7 100
III.5 Neuronnes d'une SOM dans l'espace des paramètres 104
III.6 Carte de Kohonen sur données synthétiques 105
III.7 Carte de Kohonen sur données synthétiques (2) 105
III.8 Carte de Kohonen ayant pour entrée les spectres EPIC « binés ». Gauche : la couleur code la densité locale dans l'espace des paramètres. Centre : la couleur code le nombre d'entrées associé à chaque neurone. Droite : la couleur code la valeur selon le 11^{e} bin. 105
III.9 Exemple de plan séparateur d'une SVM 109

IV.1 Probabilités d'identification et distances 2XMMi/DR7 118
IV.2 Complétude et fiabilité du 2XMMi/DR7 119
IV.3 Histogrammes des distances en sigma des couples XMM/DR7 120
IV.4 Étoiles et QSO SDSS dans le diagramme $u-g$ / $g-r$: densité de probabilité . . . 124
IV.5 Étoiles et QSO SDSS dans le diagramme $u-g$ / $g-r$: probabilité d'être une étoile 125
IV.6 Résultat de la KDC sur les sources 2XMMi/DR7 126
IV.7 Diagrammes $log(f_x/f_r)$ vs $g-i$ de l'échantillon XMM/DR7 (1) 128
IV.8 Diagrammes $log(f_x/f_r)$ vs $g-i$ de l'échantillon XMM/DR7 (2) 129
IV.9 Diagrammes $log(f_x/f_r)$ vs $g-i$ de l'échantillon XMM/DR7 (3) 130
IV.10 Diagrammes $u-g$ vs $g-r$ et $g-r$ vs $r-i$ de l'échantillon XMM/DR7 131
IV.11 Diagramme $g-r$ vs $r-i$ de étoiles de l'échnatillon XMM/DR7 133
IV.12 Luminosité X VS HR2 pour les étoiles de l'échnatillon XMM/DR7 134
IV.13 Histogramme en $r-i$ des étoiles de l'échantillon XMM/DR7 135
IV.14 Erreurs sur les magnitudes optiques de l'échantillon XMM/DR7 136
IV.15 Log(Lx) vs HR pour les galaxies et QSO de l'échantillon XMM/DR7 138
IV.16 HR3 VS z pour les galaxies et QSO de l'échantillon XMM/DR7 139
IV.17 Distribution des corrélations XMM/DR7 dans le diagramme $u-g$ vs $g-i$ 140
IV.18 Diagramme HR3 vs $u-g$ pour les quasars spectroscopiques XMM/DR7 141
IV.19 Diagramme O III/H_β vs N II/H_α 142
IV.20 Résultats de l'ACP sur l'échantillon 2XMMi/DR6 142
IV.21 Diagrammes $log(f_x/f_r)$ vs $g-i$ de l'échantillon XMM/DR7 classé 144
IV.22 Complétude et fiabilité du 2XMMi/2MASS 145
IV.23 Données 2XMMi-GSC2-2MASS 146
IV.24 Données 2XMMi-GSC2-2MASS 147
IV.25 Données 2XMMi-GSC2-2MASS classées (1) 148
IV.26 Données 2XMMi-GSC2-2MASS classées (2) 149
IV.27 Diagramme HR1 vs HR2+HR3 des sources du 2XMMi 151
IV.28 Histogrammes des probabilités d'appartenance à chaque classe 152

V.1 Modèle de données de SAADA 174
V.2 Problème d'unicité des sources d'archive 175
V.3 Traitement des sources d'archive avec doublons 175
V.4 Schéma de la base de données du modules de corrélation 179
V.5 Entrées-sorties du modules de corrélation 180
V.6 Processus d'insertion des corrélations dans la XCAT-DB 186
V.7 Interface XCAT-DB avec probabilités d'identification 187

Liste des tableaux

I.1 Bandes d'énergies du 2XMMi 36
I.2 Rapports de dureté d'XMM-Newton 36
I.3 Nombre de détections par filtre XMM 37

II.1 Nombre de pixel en fonction de la taille d'une boîte glissante 74

IV.1 Nombre de candidats DR7 par source 2XMMi unique 118
IV.2 Fiabilité et complétude des associations 2XMMi/DR7 118
IV.3 Composition des étoiles de l'échantillon d'apprentissage 122
IV.4 Composition des binaires X de l'échantillon d'apprentissage 122
IV.5 Composition des galaxies de l'échantillon d'apprentissage 123
IV.6 Composition des AGN et des QSO de l'échantillon d'apprentissage 123
IV.7 Composition de l'échantillon d'apprentissage 123
IV.8 Résultat de l'auto-classification du LS SDSS par KDC 124
IV.9 Composition de l'échantillon d'apprentissage 127
IV.10 Candidats quasar de type II 137
IV.11 Matrice de confusion pour l'éhantillon 2XMMi/DR7 143
IV.12 Points de *meanshift* pour l'échantillon 2XMMi/DR7 144
IV.13 Matrice de confusion pour l'éhantillon 2XMMi-GSC2-2MASS (1) 148
IV.14 Matrice de confusion pour l'éhantillon 2XMMi-GSC2-2MASS (2) 149
IV.15 Matrice de confusion pour l'éhantillon 2XMMi seul 151
IV.16 Échantillons pour comparaison des algorithmes de classification 154
IV.17 Matrice de confusion 154
IV.18 Résultats des algorithmes de classification sur XS_3D_1 158
IV.19 Résultats des algorithmes de classification sur XS_3D_2 159
IV.20 Résultats des algorithmes de classification sur XS_3D_3 160
IV.21 Résultats des algorithmes de classification sur XS_3D_4 161
IV.22 Résultats des algorithmes de classification sur XS_9D_1 162
IV.23 Résultats des algorithmes de classification sur XS_9D_2 163
IV.24 Résultats des algorithmes de classification sur XS_9D_3 164
IV.25 Résultats des algorithmes de classification sur XS_9D_4 165

V.1 Quelques chiffres sur le contenu de la XCAT-DB 172

Introduction

Comme pour de nombreuses disciplines, les techniques d'investigation en astronomie ont évolué avec le développement de l'informatique, du Web, et, plus largement, des technologies modernes. Différents types d'instruments permettent aujourd'hui la collecte de données dans des longueurs d'ondes allant des rayons gamma au domaine radio, alors qu'elle était autrefois limitée à l'étroite bande optique. L'observation minutieuse du ciel à l'œil nu a été remplacée par les plaques photographiques puis, depuis environ 25 ans, par les caméras CCDs. Le traitement des images numériques ou numérisées, associé au développement de pipelines de traitement de données automatiques, a permis la construction, depuis une quinzaine d'année, de grands ensembles de données cohérents à partir de campagnes d'observations. On mentionnera à titre d'exemple la construction des catalogues optiques Tycho (1997), USNO A2 et B1 (1998 et 2003), du catalogue X Rosat (1999) ou encore du catalogue infrarouge 2MASS (2003). Plus récemment la 7^{e} version du Sloan Digital Sky Survey, qui est accessible depuis fin 2008, recense plus de trois cent millions de sources optiques. Face à l'explosion des informations disponibles et à la multiplication des points d'accès, les astronomes développent des protocoles et des normes par soucis d'interopérablilité entre les différents services. Aussi l'objectif à terme de l'Observatoire Virtuel (OV) est de permettre à n'importe quel astronome de chercher et d'accéder de façon transparente à n'importe quelles données instrumentales mises en ligne n'importe où sur la surface du globe.

Dans ce contexte général, l'astronomie X prend une place de plus en plus importante. C'est une discipline relativement récente. L'atmosphère terrestre étant opaque aux rayons X, il a fallu attendre le développement des technologies spatiales et le lancement, en 1978, du satellite Einstein pour qu'elle connaisse un réel essor. Depuis, d'autres satellites ont suivi. Un nouveau pas à été franchi avec les satellites américain et européen Chandra et XMM-Newton, tous deux lancés en 1999. L'observatoire spatial de l'ESA est à l'origine d'un nombre considérable de publications scientifiques pour un instrument en astronomie. L'analyse systématique de ses observations à permis au SSC, le consortium en charge de l'exploitation scientifique d'XMM-Newton, de créer le plus grand catalogue de sources X disponible à ce jour. Il contient près de trois cents mille détections. C'est sur l'analyse statistique de ces sources X que porte le présent travail de thèse. Un des objectifs est de les classer en fonction des grands types d'objets que sont les étoiles, les quasars et noyaux actifs, les binaires X et les galaxies.

Les enjeux potentiels sont de plusieurs natures. Premièrement, la mise à disposition de telles informations pour la communauté représente un intérêt certain, que ce soit par exemple pour la recherche d'objets particuliers ou encore pour la création d'échantillons de référence. L'estimation des populations d'objets de différentes natures, à différents endroits dans l'espace et le temps, permet de mieux contraindre certains paramètres tel que le taux de formation de systèmes binaires. De telles estimations permettent d'affiner, voire de discriminer certains scénarios d'évolution des galaxies, et contribuent de cette façon à notre compréhension de l'univers. Une analyse poussée de certaines sources Chandra à récemment montré, par exemple, que les astronomes avaient jusque récemment largement sous-estimé le nombre de système binaires X contenant une naine blanche, expliquant ainsi le mystérieux rayonnement diffus X galactique au voisinage du bulbe central. D'autres questions restent ouvertes. On observe par exemple un fond diffus X extragalactique dont l'origine est toujours incertaine. Il pourrait être dû à une classe particulière de noyaux actifs de galaxies, les AGN de type 2, prédits par les modèles standards, en partie trouvés grâce à

l'OV, mais qui, globalement, restent encore peu identifiés.

Les rayons X proviennent de phénomènes astrophysiques à hautes énergies. Nous parvient essentiellement l'émission coronale des étoiles tardives (G-K) jeunes (moins d'un milliard d'années), les phénomènes de chocs dans les restes de supernovae et les étoiles chaudes (O, B, A), les phénomènes d'accrétion pour les AGNs ou les systèmes binaires, et l'émission thermique du gaz chaud en équilibre hydrostatique au coeur des amas de galaxies. Les données X seules ne permettent pas d'associer sans ambiguïté à une source X une classe d'objet donnée. Il faut donc récupérer, pour chaque source, des informations dans différentes longueurs d'ondes.
Pour récupérer ces informations multi-longueur d'onde, il est nécessaire de disposer d'un algorithme de corrélation croisée et d'accès à différents catalogues. Les résultats d'une corrélation croisée comportent toujours des fausses associations. Il est important de pouvoir estimer le taux de fausses associations qu'elles comportent et de disposer d'outils statistiques permettant de fixer ce taux pour différents échantillons. À partir de ces échantillons – dont on maitrise le nombre d'entrées erronées – et d'identifications spectroscopiques (par exemple), on peut créer des échantillons d'identification qui serviront eux-même à construire des échantillons d'apprentissage. Ces échantillons d'apprentissage pourront alors être utilisés par des algorithmes de classification supervisée pour classer les sources multi-longueur d'onde obtenues grâce à la corrélation croisée. L'analyse en composantes principales (ACP) est une méthode d'analyse de données multi-dimentionnelles. Elle permet d'extraire d'un ensemble de paramètres ceux qui sont le plus pertinents. Elle permet également de réduire la dimensionnalité des données en cherchant les informations redondantes (les paramètres qui sont corrélés entre eux). Ce dernier point est particulièrement intéressant avant d'effectuer une classification de ces données. Mais l'algorithme classique d'ACP est peu robuste.

Nous commencerons ce manuscrit par une introduction générale sur les différentes sources qui peuvent être détectées en X. Ce premier chapitre comportera également une présentation d'XMM Newton et du catalogue 2XMMi construit par le SSC (Survey Science Center) à partir de ses observations.

Le chapitre II sera dédié à l'algorithme de corrélations croisées que nous avons développé et implémenté pendant cette thèse. Il approfondira certains points techniques et présentera en détail le *kernel smoothing* utilisé pour les calculs de densités locales.

Le chapitre III traitera de l'analyse en composantes principales et de méthodes de classification. Nous avons modifié l'algorithme classique d'ACP pour qu'il prenne en compte les erreurs de mesures. Nous présenterons ensuite les différents algorithmes de classification : réseaux de neurones de Kohonen, arbres de décisions, *meanshift*, k plus proches voisins, *kernel density classification*, séparateurs à vaste marges (SVM).

Le chapitre IV abordera les résultats obtenus grâce aux différents algorithmes décrits dans les chapitres précédents. Nous présenterons notamment les résultats de la corrélation entre les catalogues 2XMMi et SDSS DR7, suivis d'une discussion scientifique sur l'échantillon obtenu. Nous verrons l'utilisation de cet échantillon pour la construction d'un échantillon d'identifications. Cet échantillon d'identifications sera utilisé pour créer les échantillons d'apprentissage qui servent dans les différents exemples de classification que comporte le chapitre. Ce chapitre contient également une comparaison de différents algorithmes de classification, supervisée et non-supervisée.

Enfin, le chapitre V abordera l'insertion de nos travaux dans la XCAT-DB. Nos résultats sont ainsi rendus publics et accessibles par l'ensemble de la communauté.



Chapitre Premier

L'astronomie X et XMM-Newton

L'astronomie X est une discipline relativement récente qui a pour origine la découverte fortuite, en 1948, de l'émission X solaire. Les instruments qui firent cette première mesure n'étaient autres que de simples compteurs Geiger embarqués dans un V2 lancé par les militaires américains. Le but, dans le contexte de la guerre froide, était la détection d'éventuels essais nucléaires soviétiques. Il faudra attendre 1962 et le lancement d'une fusée Aerobee équipée de trois compteurs Geiger pour une première étude scientifique. Là encore, le résultat fut une découverte fortuite qui valut à Riccardo Giacconi, 40 ans plus tard, le prix Nobel de physique 2002. À défaut d'observer d'hypothétiques émissions X lunaires, Giacconi et son équipe détectaient la première source X extra-solaire, baptisée Scorpius X-1. C'est également Giacconi qui est à l'origine du satellite Uhuru avec lequel fut produit le premier relevé de sources X.

Le développement de l'astronomie X, relativement tardif, est dû aux contraintes observationnelles : comme les rayons X sont arrêtés par l'atmosphère (une couche de seulement 10 cm d'air suffit à stopper un photon X de 3 keV), il faut recourir à des ballons-sondes ou à des satellites pour les observer. Ce n'est donc qu'avec le développement des technologies spatiales, et le lancement du satellite Einstein en 1978, que cette discipline a connu un réel essor. Les rayons X sont difficile à collimater et les collimateurs (de simples tuyaux empêchant le passage des photons en dehors de l'axe) ont été remplacé par des « miroirs » particuliers : les miroirs de type Wolter. Einstein a été le premier satellite à disposer de miroirs de type Wolter, à offrir des images de sources X étendues et d'émissions diffuses, et le premier a détecter des sources X faibles.

Avant Einstein, le satellite Uhuru[1] (1970 – 1973) mentioné ci-dessus, aussi connu sous le nom de SAS-1, avait découvert 339 sources X. On lui doit notamment la découverte de l'émission X diffuse en provenance des amas de galaxies. Depuis Einstein (1978 – 1981), de nombreux satellites ont suivi. On peut citer à titre d'exemple Exosat (1983 – 1985), Rosat (1990 – 1998), ASCA (japonais, 1993–2000), l'italo-néerlandais BeppoSAX (1996 - 2002) ou encore le japonais Suzaku lancé en 2005.
Une nouvelle étape a été franchie avec le lancement en 1999 de Chandra par la NASA et de XMM-Newton par l'ESA. Ces deux satellites, complémentaires, offrent une nouvelle vision de l'univers en X grâce à leur précision, leur sensibilité et la qualité de leurs spectres X. La quantité de sources détectées, associée à de faibles erreurs sur les positions, ouvre des possibilités d'analyses statistiques.

Ce chapitre commence par une revue sommaire des différents émetteurs X et, de façon simplifiée, de leurs mécanismes d'émission. On abordera ensuite les caractéristiques générale du satellite XMM-Newton et la construction du plus grand catalogue des sources X publié à ce jour.

[1] http://heasarc.gsfc.nasa.gov/docs/uhuru/uhuru.html

1 Les différents émetteurs X

1.1 Origine des rayons X

Il y a principalement cinq processus qui génèrent la majorité de l'émission X qu'on observe. Il s'agit des émissions :

- de corps noir, qui provient de milieux optiquement épais ;
- de type bremsstrahlung thermique avec des transitions *bound-free*[2] et *bound-bound*[3], qui provient de gaz chaud optiquement mince et qui est à l'origine de raies d'émission ;
- Compton inverse, qui est due au freinage d'électrons en mouvement rapide par des photons ;
- cyclotron/synchrotron, liée à des particules chargées qui spiralent dans un champ magnétique macroscopique.

De façon très schématique, on trouve essentiellement l'émission de corps noir dans les disques d'accrétion et les astres compacts très chauds (naines blanches et étoiles à neutrons). Le rayonnement mince thermique intervient essentiellement dans les couronnes, les gaz choqués et les gaz irradiés dans les vents des étoiles massives, les restes de supernovae, …. L'émission Compton inverse a un rôle important dans le rayonnement X des jets et couronnes associés aux disques d'accrétion, notamment au sein des AGN. Enfin, les pulsars les plus puissants présentent des raies cyclotron jusque dans la gamme des rayons X.

1.1.1 Rayonnement du corps noir

Le corps noir est un modèle théorique d'objet (un gaz de photons). Pour connaître son rayonnement, on cherche à déterminer la distribution d'énergie des photons à l'intérieur. On fait l'hypothèse qu'il est en équilibre thermique et qu'aucun rayonnement n'y rentre ni n'en sort. Par des considérations thermodynamiques, et sachant que la masse des photons est nulle, on peut alors caractériser son rayonnement (voir par exemple Rybicki & Lightman 1979).

L'intensité lumineuse du rayonnement de corps noir par unité de longueur d'onde, exprimée en erg.s^{-1}.cm^{-2}.sr^{-1}.Hz^{-1}, ne dépend que de la température et s'exprime par la fonction de Planck (Eq. I.1) suivante :

$$B(E,T) = \frac{2hc^2/\lambda^4}{\exp\left(hc/\lambda kT\right) - 1} \tag{I.1}$$

avec h la constante de Planck (6.626x10^{-27} erg.s) et k la constante de Boltzmann (1.36x10^{-16} erg.K^{-1}). Ou encore, en fonction de l'énergie des photons :

$$B(E,T) = \frac{2E^3}{h^2c^2\exp\left(E/kT\right) - 1} \tag{I.2}$$

Une émission de corps noir se produit quand un rayonnement électromagnétique est piégé dans un gaz thermique. Dans les étoiles, on sait par exemple qu'un photon produit dans le cœur peut mettre plusieurs millions d'années pour arriver jusqu'à la photosphère d'où provient la plus grande partie de son émission. Le rayonnement provenant de cette région est généralement proche d'une émission de corps noir à laquelle s'ajoutent des raies d'absorption caractéristiques des atomes présents dans la photosphère.

Pour les objets astrophysiques dont le spectre est proche de celui du corps noir, on peut estimer leurs températures via la loi de déplacement de Wien (Eq. I.3).

$$\lambda_{\max} T = \sigma_w = 2.898\,768\,651 \text{ m.K} = 2.898\,768\,651\ 10^2 \text{ cm.K} \tag{I.3}$$

Il suffit de déterminer la longueur d'onde pour laquelle l'intensité lumineuse est maximale.

[2] Émission d'un photon suite à la capture d'un électron par un ion (recombinaison). Ce type de transition est en général à l'origine d'un spectre continu.

[3] Émission d'un photon suite au transfert d'un électron d'une couche d'énergie à une autre dans un atome.

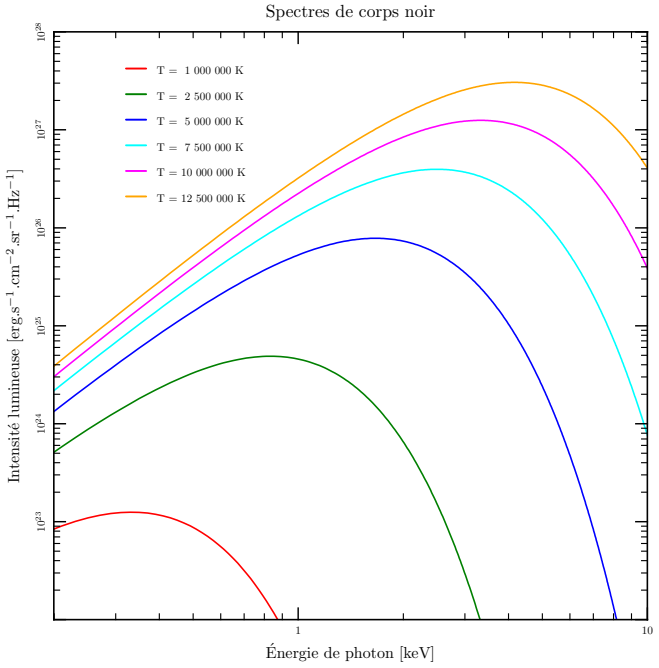


FIG. I.1 – Intensité lumineuse émise par des corps noirs de différentes températures en fonction de l'énergie des photons en keV. La bande 0.2-12 keV correspond à la plage d'énergie couverte par XMM-Newton. La longueur d'onde d'un photon en fonction de son énergie en keV est donnée par la formule $\lambda = hc/(1.6\ 10^{-9}E)$.

Mais en général on utilise plutôt le rayon R et la luminosité L de l'étoile pour déterminer sa température effective via la relation :

$$T_{\text{eff}}^4 = \frac{L}{4\pi R^2 \sigma_{sb}} \tag{I.4}$$

Où σ_{sb} est la constante de Stefan-Boltzmann qui vaut 5.670 400 10^{-8} W.m^{-2}.K^{-4}, soit 5.670 400 10^{-5} erg.s^{-1}.cm^{-2}.K^{-4}.

1.1.2 Émission thermique de gaz chaud : le bremsstrahlung thermique

Cette émission provient de gaz peu denses, optiquement minces, ayant des températures T>100 000 K. A ces températures, le gaz consiste en un plasma d'électrons libres et d'ions positifs. Dans ces conditions les rayons X proviennent essentiellement de la superposition de trois phénomènes : le Bremsstrahlung thermique (encore appelé rayonnement de freinage ou rayonnement *free-free*), les transitions électroniques (rayonnement *bound-bound*) et les recombinaisons (rayonnement *bound-free*).

- bremsstrahlung thermique : quand un électron passe près d'un ion positif, sa trajectoire est déviée par le champ coulombien de l'ion et l'électron émet de l'énergie sous forme d'un rayonnement électromagnétique ;
- quand un électron rentre en collision avec un ion, il peut céder de l'énergie aux électrons de l'ion qui passent à un état d'excitation supérieur. La désexcitation de l'ion entraîne un rayonnement qui génère des raies d'émission.

La composante continue du bremsstrahlung thermique s'exprime par la formule I.5 suivante :

$$B(E,T) = AG(E,T)Z^2 n_e n_i (kT)^{-1/2} \exp(-E/kT) \tag{I.5}$$

Avec :

- A : une constante

- $G(E,T) = \frac{\sqrt{3}}{\pi} \log(\frac{kT}{E})$: facteur de Gaunt
- Z : charge électrique des ions positifs
- n_e : densité d'électrons
- n_i : densité d'ions positifs

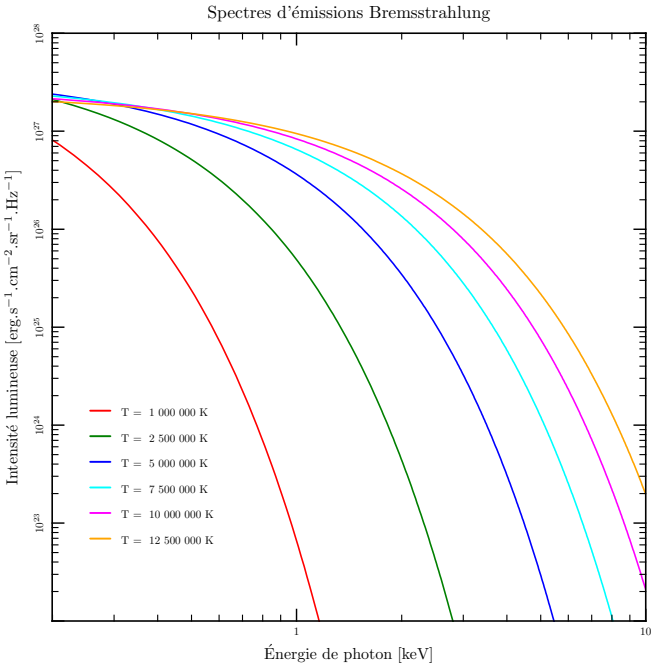


FIG. I.2 – Intensité lumineuse émise par des émissions Bremsstrahlung de plasma de différentes températures en fonction de l'énergie des photons en keV. La bande 0.2-12 keV correspond à la plage d'énergie couverte par XMM-Newton. La longueur d'onde d'un photon en fonction de son énergie en keV est donnée par la formule $\lambda = hc/(1.6\ 10^{-9}E)$.

1.1.3 Émission synchrotron : émission en loi de puissance

L'émission synchrotron, parfois appelée bremsstrahlung magnétique, à pour origine le freinage d'électrons plongés dans un champ magnétique macroscopique. C'est un rayonnement de freinage magnétique (cf. Théorie des Champs, Landau et Lifchitz). Cette émission, souvent polarisée, présente fréquemment un spectre en loi de puissance (quand le pectre des électrons est lui-même en loi de puissance) :

$$B(\lambda) = A\left(\frac{hc}{\lambda}\right)^{-\alpha} \tag{I.6}$$

Ou encore, en fonction de l'énergie :

$$B(E) = AE^{-\alpha} \tag{I.7}$$

avec A une constante et α l'indice spectral, encore appelé indice de photon. Plus la valeur de α est petite, plus l'émission est dure.

1.1.4 Effet Compton Inverse : émission en loi de puissance

L'effet Compton inverse est dû au freinage d'électrons rapides par des photons. Quand un plasma d'électrons allant à des vitesses relativistes rencontre des photons, il y a des collisions. Lors d'une collision, l'électron cède de l'énergie au photon qui, par conséquent voit sa fréquence augmenter.
Considérons le spectre de corps noir correspondant à la température des photons avant collision. Les différentes énergies cédées aux photons lors des collisions sont à l'origine d'un décalage en

longueur d'onde du spectre de corps noir d'origine ayant une intensité de plus en plus faible au fur et à mesure que $\Delta\lambda$ croît. La somme de ces spectres de corps noir décalés forme un spectre en loi de puissance.
Dans le cas des QSO, l'émission X est dominée par l'effet Compton inverse. Leurs spectres s'apparentent donc à des lois de puissances dont l'indice spectral, ou indice de photon, est souvent compris entre 1.6 et 1.9.

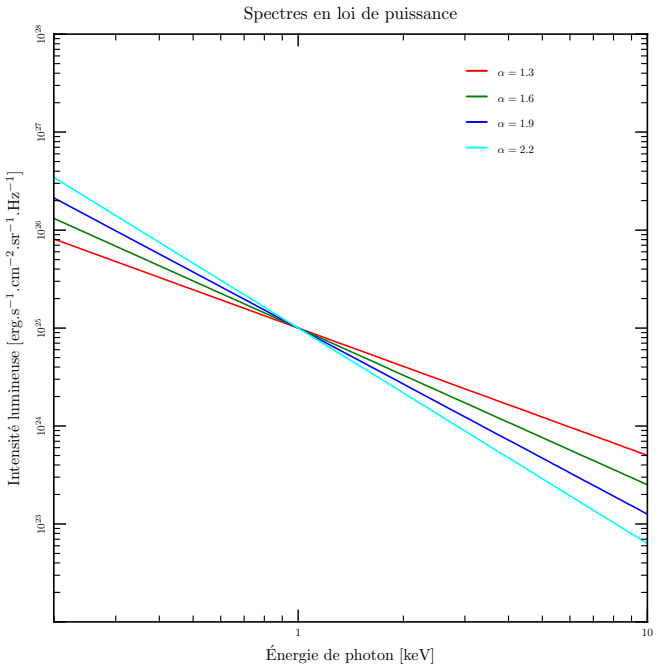


FIG. I.3 – Intensité lumineuse émise par des sources d'émission dont le spectre est en forme de loi de puissance en fonction de l'énergie des photons en keV.

1.1.5 Caractéristiques observationnelles

On définit ici quelques termes qui seront utilisés dans la suite.

a Type spectral

Le type spectral d'une étoile consiste en une lettre, un chiffre et éventuellement un chiffre romain et permet de situer une étoile sur le diagramme de Hertzsprung-Russell (Fig. I.4). On peut en déduire la masse et la température de l'étoile ainsi que son état évolutif. La première lettre désigne la classe principale, le chiffre la classe secondaire et le chiffre romain la classe de luminosité. Pour plus d'informations voir la section 1.3.1 ci-après.

b Reprocessing

Le *reprocessing* de rayons X consiste en l'absorption et la réémission à plus faible énergie de ces rayons X par de la matière. Ce terme, générique, peut par exemple désigner de la diffusion par effet Compton.

c Limite d'Eddington

La limite d'Eddington représente la luminosité limite pour laquelle la pression de gravité arrive encore à contre-balancer la pression de radiation. Cette limite est définie pour les étoiles par la formule suivante :

$$L_{\text{edd}} \sim 1.3\,10^{38}\,\frac{M}{M_\odot}\ \text{erg.s}^{-1} \tag{I.8}$$

Au delà de cette limite, la pression de radiation l'emporte sur la gravité et l'objet va expulser de la matière en quantité importante. En l'abscence de mécanismes spécifiques autres que la gravité, qui permet soit d'augmenter la luminosité sans augmenter la pression de radiation, soit de contre-balancer cette dernière, une source super-Eddington (dont la luminosité dépasse la limite d'Eddington) est dans un état nécessairement transitoire.

d Quelques grandeurs mesurées

Nous utiliseront souvent le rapport Fx/Fopt. Cette grandeur représente le rapport entre le flux X observé d'une source et son flux observé en optique. Le logarithme décimal du rapport Fx/Fopt est directement proportionnel (à une constante multiplicative -2.5 et à une constante additive près) à une couleur X–optique. Sauf mention contraire, le flux X (Fx) sera pris dans la bande 0.2-12 keV d'XMM et le flux optique correspond à la bande r du SDSS.

Nous emploierons également fréquemment les rapports de dureté (HR pour *Hardness ratio* en anglais). Les rapports de dureté représentent la différence normalisée entre le nombre de photons reçus par unité de temps dans deux bandes d'énergies. La différence se fait par rapport à la bande de plus grande énergie, à savoir la bande la plus « dure ». Le lecteur trouvera la définition des différent HR pour XMM-Newton dans la section 2.3.2.

Dans le jargon des astronomes qui travaillent avec des données X, les rayons X durs (*hard*) sont des rayons X dont l'énergie est plus importante que celle des rayons X mous (*soft*). Mais ces notions sont assez relatives. On parle par exemple de la partie molle et de la partie dure d'un spectre pour désigner respectivement les composantes de basses et hautes énergies. De même, on qualifie un spectre de mou ou dur en fonction de l'énergie à laquelle se trouve le maximum d'émission X.

1.2 Des sources X dans le système solaire

On sait depuis les résultats apportés par Rosat en 1996 que, outre le Soleil, la Lune et la Terre, Venus, Jupiter, Saturne, Io, Europa et même peut-être Ganymède émettent en X (Dennerl 2008, §9.1). On peut également ajouter à cette liste les comètes. La quasi-totalité de ces émissions X provient de l'interaction entre les objets et le vent solaire. Par exemple, une des réactions qui prédomine est le captage d'un électron venant souvent d'une molécule d'eau ou d'OH par un ion O^{7+} présent dans ce dernier. Cette transition ($O^{7+} \rightarrow O^{6+}$) par échange de charges provoque une raie d'émission caractéristique à 0.561 keV. Pour les astres possédant des champs magnétiques intenses comme la Terre ou Jupiter un autre phénomène intervient : des électrons énergétiques qui sont accélérés dans la magnétosphère sont à l'origine d'une émission bremsstrahlung quand ils sont freinés par l'atmosphère planétaire.

1.3 L'émission X des étoiles

1.3.1 Rappel sur la classification spectrale des étoiles

Les étoiles sont classées selon leur position dans le diagramme HR représenté Fig. I.4. L'axe des x correspond à la couleur et l'axe des y à la luminosité. La classification spectrale des étoiles, des plus chaudes (bleu) aux plus froides (rouge), suit la séquence de lettres suivante : OBAFGKM. Deux étoiles de même couleur peuvent avoir une luminosité différente. Les classes de luminosité sont :

Classe VII : naines blanches (*white dwarf*) ;

Classe VI : étoiles sous-naines (*sub-dwarf stars*) ;

Classe V : étoiles de la séquence principale (*main sequence stars*), parfois appelées étoiles naines (*dwarf stars*) ;

Classe IV : étoiles sous-géantes (*sub-giant stars*) ;

Classe III : étoiles géantes (*giant stars*) ;

Classe II : étoiles géantes brillantes (*bright-giant stars*) ;

Classe I : étoiles super-géantes (*super-giant stars*) qui se divisent en deux classes, Ia et Ib

Classe O : étoiles hyper-géantes (*hyper-giant stars*)

Pour les classes de luminosité allant de 0 à IV, on parle d'étoiles évoluées (*evolved stars*), c'est-à-dire qui ont quitté la séquence principale après avoir consumé l'essentiel de l'hydrogène de leur cœur.

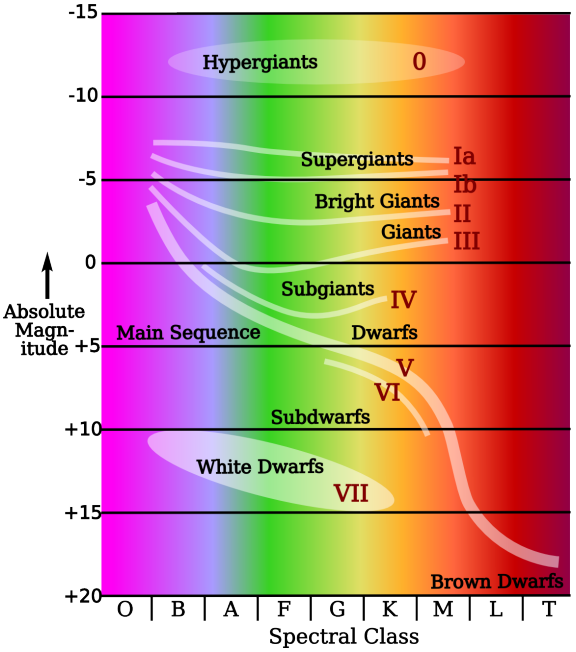


FIG. I.4 – Diagramme de Hertzsprung-Russell (HR) représentant les différents types d'étoiles (crédits : Wikipédia).

1.3.2 Les étoiles pré-séquence principale

a Étoiles T Tauri

La phase pré-séquence principale s'étend de la contraction du nuage de gaz et de poussières primordial, phase proto-étoile, au déclenchement des réactions thermonucléaires de fusion dans le coeur de l'astre. Pour que l'étoile en formation devienne visible en optique et en X, elle doit rayonner suffisamment. Le rayonnement qu'elle émet grâce au mécanisme de Kelvin-Helmholtz vient de l'énergie gravitationnelle libérée par la contraction du nuage. Dans la suite, on utilisera le terme générique « étoile » pour désigner l'étoile en formation.

Les T Tauri sont les plus jeunes des étoiles visibles de type spectral F, G, K ou M ayant une masse inférieure à 2 masses solaires. Elles émergent de leur enveloppe gazeuse après un à 10 millions d'années (Schmitt & Stelzer 2008, §10.4). On détecte autour de ces étoiles des champs magnétiques importants et la présence de *flares* – interprétation de la variabilité X – peut laisser supposer que les mécanismes à l'origine de ces champs doivent être similaires aux étoiles de type solaire ou aux étoiles M (Güdel 2004, §18.1.5). Ne disposant pas de cœur radiatif, ils pourraient par exemple provenir d'un effet dynamo purement convectif dit α^2.
L'émission X des *flares* est associée à une émission thermique de plasma chauffé rapidement à plus de 10^7 K par de violentes reconnexions magnétiques (Feigelson & Montmerle 1999, §1), comme dans la couronne solaire, mais d'intensités supérieures.
Pour certaines T Tauri, on pense que la matière accrétée est chauffée par des chocs près de la surface de l'étoile. Cette région serait à l'origine d'un excès de X mous (<0.5 keV) détecté dans les systèmes où l'absorption photoélectrique par le gaz accrété est faible (Güdel & Telleschi 2007).
Pour en savoir plus sur les T Tauri, voir (Güdel 2004, §18), Feigelson & Montmerle (1999) et Güdel & Telleschi (2007).

b Étoiles de type Herbig Ae/Be

Les étoiles Herbig Ae/Be, ou HAeBe, sont des étoiles pré-séquence principale plus massives que les T Tauri ($M \sim 2 - 10 M_{\odot}$), et qui ont tendance à être plus lumineuses en X. Elles sont purement radiatives et on suppose que comme pour les étoiles chaudes de la séquence principale, leur champ magnétique est d'origine fossile. Selon un autre scénario, le champ magnétique est dû à une rotation différentielle entre le cœur de l'étoile et ses couches externes.
À la place d'une émission coronale classique d'étoile de type solaire, la production de rayons X peut être expliquée par un modèle de chocs entre vents confinés magnétiquement (Schmitt & Stelzer 2008, §10.4.3). Cependant, la nature exacte de la forte émission X de ces objets est toujours sujette à controverse car elle possède à la fois des caractéristiques des étoiles chaudes et des étoiles froides (Güdel 2004, §4.4.3).

c Les étoiles pré-séquence principale massives

Les progéniteurs des étoiles encore plus massives sont assez peu connus. Ces étoiles naissent dans des régions H II très compactes. Ces zones de formation sont souvent distantes, donc difficiles à observer. De plus, du fait de la quantité de gaz importante, ces zones sont très absorbées.
Un progéniteur d'étoile B se trouve dans Orion mais n'a pas été détecté en X malgré des observations faites par le satellite américain Chandra. En revanche, une émission X dure provenant de la protoétoile GGD 27 de type B a été récemment observée (Pravdo *et al.* 2009).

d Les objets de type Herbig-Haro

Les objets de type Herbig-Haro (HH) sont des nébuleuses lumineuses dans des zones de formation d'étoiles. Elles sont issues du choc d'un puissant vent bipolaire émis par une étoile en formation qui rentre en contact avec le milieu interstellaire. Ces zones de chocs émettent en rayons X avec des températures typiques de l'ordre de 1 à 10 millions de degrés K (Schmitt & Stelzer 2008, §10.4.5).

1.3.3 Les étoiles de la séquence principale

On peut séparer les étoiles de la séquence principale en quatres catégories en fonction de leur structure interne :

- la catégorie des étoiles chaudes, ou *early type stars*, qui regroupe les étoiles de type spectral O et B et qui ont un coeur convectif et une couche externe radiative ;
- les étoiles de type spectral A, qui sont purement radiatives ;
- la catégorie des étoiles froides, ou *cool stars*, ou encore *late type stars*, qui regroupe les étoiles de type spectral allant de F à la moitié des M, et qui ont un cœur radiatif et et une couche externe convective, comme notre Soleil ;
- la moitié la plus froide des étoiles M, qui sont purement convectives.

Nous allons voir que cette structure interne joue un rôle dans les mécanismes d'émission de rayons X. Une représentation schématique des différents types d'étoiles de la séquence principale, qui par définition appartiennent toutes à la classe de luminosité V, est présentée Fig. I.5.

a Émission coronale des étoiles de type solaire

Les étoiles de type solaire sont définies ici comme étant les étoiles ayant un coeur radiatif et une couche externe convective. Ces étoiles sont pourvues de champs magnétiques venant d'un effet dynamo turbulent auto-entretenu dit alpha-omega et noté $\alpha\omega - dynamo$ dans la littérature (Pagano *et al.* 2006). Cette dynamo combine les effets de la convection, composante α, et la rotation différentielle, composante ω. Il est actuellement admis que l'origine de cette dynamo se trouve dans les basses couches de la zone convective, avec une bonne probabilité qu'elle se situe à l'interface entre la zone radiative et la zone convective (Paterno 1998).
Le satellite Einstein a découvert l'émission X d'étoiles normales. Mais, depuis Rosat, on sait que toutes les étoiles de la séquence principale de type tardif, donc qui ont une couche externe convective, émettent des rayons X (Schmitt 1997, Schmitt & Liefke 2004). Les champs magnétiques à l'œuvre dans ces étoiles sont responsables de l'émission de ces rayons X d'origine thermique. Ils

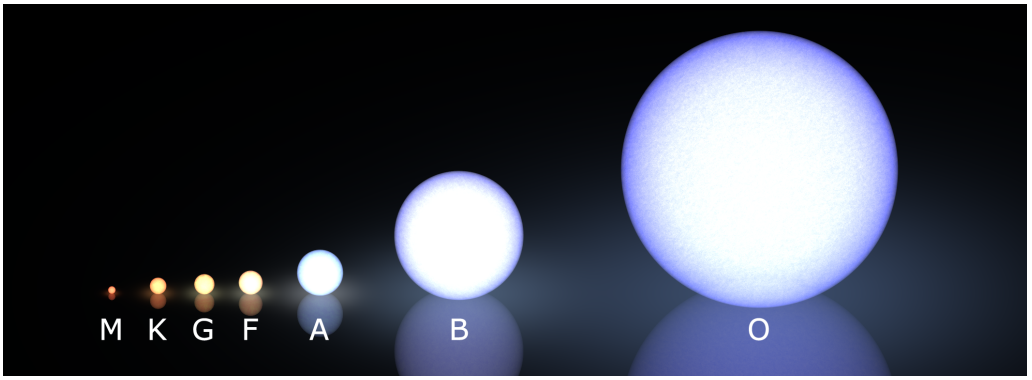


FIG. I.5 – Classification spectrale de Morgan-Keenan des étoiles de la séquence principale (crédits : Wikipédia).

proviennent de la couronne chauffée à plus d'un million de degrés par des phénomènes de reconnexion magnétique responsables des éruptions solaires et peut-être aussi par des phénomènes de dissipation d'ondes magnétohydrodynamiques, à savoir les ondes d'Alfvén (Tomczyk *et al.* 2007). Mis à part les étoiles de type F et G, l'activité coronale est souvent variable, les éruptions (*flares*) pouvant multiplier par 100 la luminosité X (Paerels & Kahn 2003) d'une étoile, notamment pour les naines M et les systèmes binaires de type RS CVn. Les étoiles forment la classe d'émetteurs X la plus variable (Schmitt & Liefke 2004).
On trouvera un exemple de spectre X d'étoile Fig. I.6. Il s'agit en l'occurence du spectre EPIC PN de HD 20367, étoile de type spectral G0V (voir van Belle & von Braun 2009, Table 3).

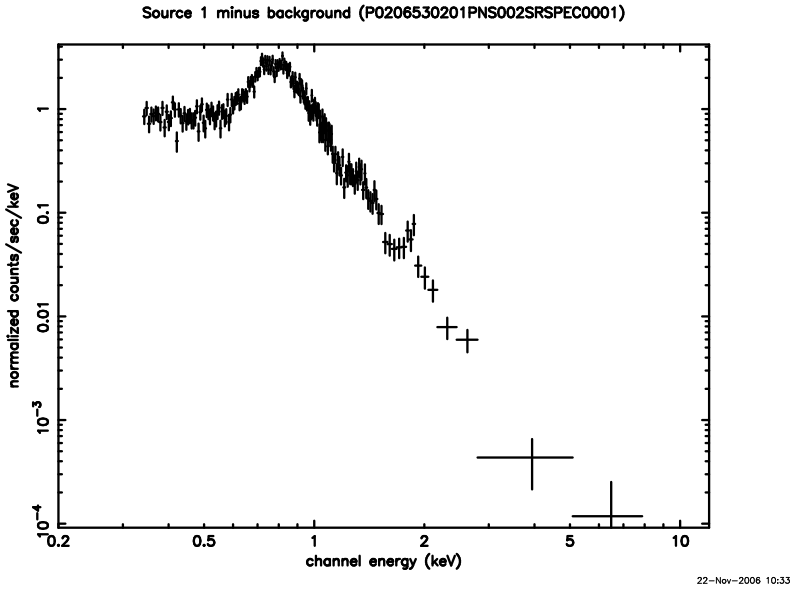


FIG. I.6 – Spectre EPIC PN de l'étoile HD 20367 de type G0V. Avec peu de photons detectés au delà de 1.5 keV, ce spectre est dit « mou ».

b Les étoiles de type M

Dans la moitié la plus froide des étoiles de type M, ces astres sont purement convectifs. Leur niveau d'activité magnétique important en l'absence d'interface entre couche radiative et couche convective où est censé se produire l'effet dynamo pose un problème théorique (Donati & Land-

street 2009, §4.3). Leur faible brillance de surface et leur champ magnétique intense pouvant atteindre 1 000 fois l'intensité du champ magnétique solaire font que le rapport Fx/Fopt de ces étoiles dans la bande 0.1-1 KeV peut être proche de l'unité.

c Les étoiles jeunes

On sait que l'activité magnétique, et donc l'activité X, des étoiles de type solaire est liée à un effet dynamo dû à la rotation différentielle des couches de convection sous la photosphère. Or l'effet dynamo augmente avec le différentiel et donc avec la vitesse de rotation de l'étoile.
Les étoiles jeunes sont des rotateurs rapides. Mais cette rotation est freinée dans le temps : les vents stellaires particulièrement importants à cause de l'activité magnétique sont à l'origine de pertes de moment angulaire ((Bouvier *et al.* 1997)) qui ont donc tendance à réduire l'activité magnétique par diminution de l'effet dynamo. Il faut environ 100 000 ans pour que cet effet se stabilise (Güdel 2004, §5.1). C'est donc tout naturellement que, avant la fin de cette période, les étoiles jeunes présentent une activité magnétique, et donc X, plus importante que celle de leurs sœurs ainées. L'étude d'amas d'étoiles tel que les Hyades et les Pleiades a permis de confirmer cette tendance. On peut en effet connaitre l'âge d'un amas et de ses différentes composantes en ajustant un modèle sur la séquence principale dans le diagramme de Hertzprung-Russell (HR). On a alors pu étudier la relation entre âge et émission X des étoiles de type solaire, confirmant que les étoiles jeunes émettent plus de rayons X (Schmitt & Stelzer 2008, §10.2.4).
Pour plus de détails sur les liens entre vitesse de rotation et activité X des étoiles, voir §5.1 et §17.5.2 de Güdel (2004).

d Les vents des étoiles chaudes

Les étoiles chaudes, ou *early type stars*, sont les étoiles de la séquence principale de type spectral O ou B. Elles émettent un rayonnement X assez mou (kT$\sim$0.5keV), sont significativement moins variables que les étoiles de type solaire et leur rapport Fx/Fopt à tendance à être plus faible que pour ces dernières. Ces différences suggèrent un mécanisme d'émission X différent entre les deux types d'étoiles.
Il est aujourd'hui établi que l'émission X des étoiles chaudes provient de chocs hydrodynamiques résultant des instabilités intrinsèques du puissant vent stellaire mû par la pression radiative de la couche externe de ces astres (Lucy & White 1980). La luminosité X, approximativement constante, provient d'une multitude de petits chocs (Schmitt & Stelzer 2008, §10.5). Dans une petite fraction des étoiles chaudes on détecte également un rayonnement X dur. On explique cette composante par un modèle hybride dans lequel le vent stellaire est perturbé par une magnétosphère : au sein d'arches magnétiques, les températures du plasma peuvent atteindre 10^7 K et sont à l'origine de l'émission X plus dure que celle des étoiles chaudes non magnétiques. En plus de prédire l'émission X dure, ce modèle permet également d'expliquer l'émission radio de ces objets. Cependant, les modèles théoriques actuels ont du mal à expliquer l'origine du champ magnétique de ces étoiles (Schmitt & Stelzer 2008, §10.6). L'hypothèse la plus vraisemblable invoque une origine fossile datant de l'époque où l'étoile s'est formée (voir Donati & Landstreet 2009, §5.3) : quand l'étoile se forme suite à l'effondrement d'un nuage interstellaire, le nuage emporte dans sa chute une parcelle du champ magnétique galactique qu'il amplifie par sa contraction.

e Incertitudes autour des étoiles A

Les étoiles de la séquence principale de type spectral A n'ont pas de couche convective et ne devraient pas montrer de signes d'émission X coronale. D'ailleurs, le satellite Rosat n'a détecté aucune des étoiles A (entre A0 et A6) situées dans le voisinage solaire (Schmitt 1997). Ces étoiles, ainsi que les étoiles de type B tardives (*late-B stars*) n'émettent pas non plus de vents suffisamment puissants pour émettre des rayons X comme c'est le cas pour les étoiles chaudes.
On s'interroge sur la possibilité qu'ont ces étoiles d'émettre des rayons X et sur l'origine de cette émission hypothétique. Un certain nombre de détections X ROSAT ont été associées à des étoiles A. Mais la résolution d'XMM-Newton et de Chandra a permis de prouver que, pour au moins la moitié des cas, l'émission X provient d'un compagnon proche dont la composante optique est noyée dans celle, dominante, de l'étoile A, mais qui apparaît en infrarouge (Schmitt & Stelzer 2008, §10.2.3).
On ne sait toujours pas, aujourd'hui, si pour l'autre moitié de ces cas l'émission X provient de

secondaires très proches ou de l'étoile primaire A ou B : « *absence of evidence is not evidence of absence* » (Schmitt & Stelzer 2008, §10.2.3). On pense en tout cas que les étoiles de type A ne sont pas capables de maintenir une couronne chaude comme le font les étoiles de type solaire (voir Güdel 2004, §4.4.1). Mais peut-être que pour certaines d'entre elles, comme pour certaines étoiles de type spectral O et B, un champ magnétique fossile suffisamment fort existe pour générer des rayons X au sein d'arches magnétiques.
Au final, les rapports Fx/Fopt des étoiles A peuvent avoir des valeurs très différentes, les composantes X et optiques ne provenant pas forcément du même astre.

1.3.4 Les étoiles évoluées

La théorie des structures stellaires prédit que lorsqu'une étoile froide arrive en fin de vie et s'engage sur la branche asymptotique des géantes (AGB pour *Asymptotic Giant Branch*), la partie convective devient plus large, ce qui pourrait augmenter temporairement l'activité X (Güdel 2004, §17.2). Mais, du fait de la dilatation de l'étoile et de la conservation du moment angulaire, cette dernière voit sa vitesse de rotation diminuée. Effet aggravé par le fait qu'en fin de séquence principale, les étoiles froides sont déjà des rotateurs lents. L'implication directe est une diminution de l'effet dynamo, et donc du champ magnétique de l'étoile. On invoque souvent cette raison pour expliquer le peu de géantes rouges trouvées parmi les émetteurs X (Güdel 2004, §4.5). De fait, la population de géantes rouges observée en X est largement dominée par les systèmes RS CVn abordés plus loin dans la section 1.5 (voir graphiques dans la thèse d'Alexis Klutch, p. 180).

Le cas des étoiles chaudes de la séquence principale est différent. À l'origine purement radiative, une zone de convection se forme quand elle s'engage sur l'AGB. Démarre alors un effet dynamo responsable de l'émission d'un vent magnétisé. Ces étoiles se retrouvent parmi les géantes G les plus émettrices en X. Mais comme le passage sur la branche asymptotique des géantes est court, on détecte relativement peu de ces objets.

Enfin, on connait peu de choses sur les étoiles supergéantes en raison du faible nombre d'objets connus.

1.3.5 Les étoiles binaires serrées non-accrétantes

On a vu que l'activité X des étoiles de type solaire est liée à leur vitesse de rotation et que si les étoiles jeunes sont des rotateurs rapides, elles sont freinées par l'éjection de matière et perdent assez vite du moment angulaire (Bouvier *et al.* 1997). Cependant, un mécanisme permet d'entretenir la vitesse de rotation : le couplage gravitationnel qui, par effet de marées, synchronise la rotation des deux astres formant un couple binaire serré détaché – sans transfert de masse – sur la période de révolution du système. Les étoiles formant un tel couple peuvent donc rester actives en X. C'est notamment le cas des RS CVn qui sont des binaires dont les deux étoiles sont de type tardif et dont la période orbitale varie de quelques jours à deux ou trois semaines. La luminosité X typique est de l'ordre de quatre fois celle du Soleil. Les RS CVn présentent des excès en UV, en IR, et sont variables (flares) en X et en Radio. L'activité chromosphérique est diagnostiquée en optique par la présence des raies d'émission H et K du calcium Ca II et de la raie d'émission de Balmer H_α (Strassmeier *et al.* 1988). Certaines sources dont les raies Ca II H et K sont non détectées en optique, mais dont le triplet du Ca II est détecté en infrarouge peuvent également être considérées comme chromosphériquement active (Strassmeier *et al.* 1993), si elles présentent la raie d'émission en H_α.

Trois autres types de systèmes binaires serrés mais séparés sont connus pour émettre en X, on mentionne les systèmes dits de Algol, les W UMa et les BY Dra (Schmitt & Stelzer 2008, §10.2.6) :

- RS CVn : deux étoiles de type tardif (pas forcément de classe de luminosité V) ; période orbitale allant de quelques jours à 2-3 semaines ; luminosité X 4 fois supérieure à la luminosité X du soleil.
- Algol : une étoile chaude (typiquement une étoile B) et une secondaire de type tardif tournant rapidement sur elle-même à l'origine du rayonnement X
- W UMa : deux étoiles de type tardif qui partagent une enveloppe commune ; période orbitale de quelques jours ; émission X importante
- BY Dra : deux naines de type tardif en orbite proche ; pas forcément couplées gravitation-

nellement, mais luminosités X importantes

1.3.6 Les étoiles post-séquence principale

a Les étoiles de type PG 1159

Les étoiles de type PG 1159 forment une étape transitoire dans l'évolution stellaire et se situent entre les étoiles post-AGB et les naines blanches (Hügelmeyer *et al.* 2007). On pense plus précisément que ces astres sont la suite évolutive des étoiles de type Wolf-Rayet qui se trouvent au centre des nébuleuses planétaires PNe (type spectral WC) et dont sont originaires les naines blanches de type spectral DO les plus chaudes (Motch *et al.* 1993). Les températures T_{eff} des étoiles de type PG 1159 se situent dans une gamme allant de 75 000 à 200 000 K (Hügelmeyer *et al.* 2007). Leurs spectres optiques, qui peuvent être ajustés par des modèles d'atmosphères qui ne sont pas localement en équilibre thermodynamique (dit NLTE pour *Non-Local Thermodynamic Equilibrium*, voir Werner *et al.* (2003, 1991), et pour une définition francophone de LTE voir Heyvaerts (2006), p. 37), sont dépourvus de raies d'hydrogènes mais montrent des raies d'ionisation en absorption He II , C IV et O VI (Motch *et al.* 1993, Werner *et al.* 1991, Werner 1992). Les mécanismes d'émission X sont semblables à ceux des naines blanches (voir ci-après).

b Les naines blanches

Les naines blanches font partie du stade final de l'évolution des étoiles de masse allant jusqu'à 8 masses solaires. Environ 97% des étoiles actuelles de notre galaxie finiront en naines blanches, puis en naines noires après une longue période de refroidissement. Le refroidissement dure plusieurs milliards d'années du fait de la faible surface et de la quantité importante d'énergie emmagasinée. Le Sloan Digital Sky Survey (SDSS) a permis la constitution de catalogues contenant actuellement plus de 9 000 de ces objets, sélectionnés sur des considérations optiques (Eisenstein *et al.* 2006).

Contrairement aux étoiles de type solaire et à la proposition de Strittmatter *et al.* (1972), les rayons X émis par les naines blanches ne proviennent pas d'une couronne chaude mais de l'émission thermique des couches profondes de la photosphère rendue peu opaque par l'absence ou la faible présence de métaux (Shipman 1976). Le rayonnement X est donc un rayonnement thermique de corps noir qui provient de zones où la température est supérieure à la température de surface donnée en optique. On classifie les naines blanches essentiellement en deux classes spectrales principales :

- les DA, qui ont une atmosphère composée quasi uniquement d'hydrogène et qui ont une vaste gamme de températures ;
- les non-DA, qui se décomposent en sous-familles :
  - les DO, qui ont une atmosphère riche en He et dont la température Teff varie entre 45 000 et 120 000 K ;
  - les DB, dont la température Teff varie entre 11 000 et 30 000 K ;
  - les DC, DQ, DZ et DH, qui ont pour point commun d'avoir une température Teff inférieure à 11 000 K et dont les premières présentent un continuum sans raie, les secondes des raies du carbone, les troisièmes des raies métalliques et les dernières des signes d'activité magnétique (Kleinman *et al.* 2004).

Si on considère les naines blanches dans l'ordre de présentation ci-dessus, leur atmosphère est de plus en plus riche en éléments lourds et donc l'opacité augmente. Il en résulte qu'on n'observe pas, en X, les types DB, DC, DQ, DZ et DH. Seules les naines blanches de type spectral DO les plus chaudes (Teff>100 000) sont détectées (O'Dwyer *et al.* 2003). Quant aux DA, elles sont détectées à partir de Teff=23 000 K si elles ne contiennent que de l'hydrogène et pour des températures supérieures sinon. Comme on peut s'y attendre, on détecte essentiellement en X les naines blanches de type DA (Werner 2008). Enfin, notons qu'en accord avec les températures mises en jeu, les naines blanches possèdent une forte émission en UV.

Même si ces températures sont élevées pour des photosphères, elles sont relativement faibles pour produire des rayons X et ce également dans des couches plus internes. De fait, l'émission X de ces étoiles ne dépasse pas les 0.5 keV (O'Dwyer *et al.* 2003). On la qualifie donc de molle, ou *soft*. A titre d'exemple, on donne Fig. I.7 le spectre XMM de la naine blanche GD 153 de type spectral DA

et ayant une température Teff de 38 500 K (voir Bohlin *et al.* 1995, Table 3). On montre également le spectre optique de la naine blanche WD 1657+343 de type spectral DA ayant une température de l'ordre de 61 000 K (voir Eisenstein *et al.* 2006, Fig. 19).

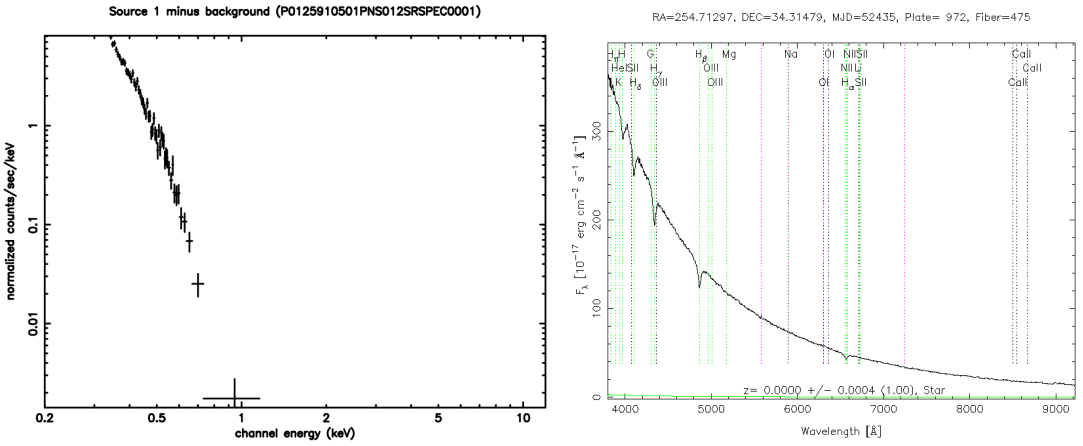


FIG. I.7 – Gauche : spectre EPIC PN de la naine blanche GD 153 de type spectral DA et de Teff=38 500 K. Ce spectre mou (pas de photon détecté au delà de 1 keV) et régulier correspond à une émission thermique de corps noir. Droite : spectre optique obtenue par le SDSS de la naine blanche WD 1657+343 de type spectral DA.

Certaines naines blanches ont été associées à des émissions X dures (>0.5keV). Dans la plupart des cas, on a prouvé que soit les naines blanches appartiennent à des systèmes binaires, l'émissions X dure provenant de la couronne du compagnon ou d'un phénomène d'accrétion (O'Dwyer *et al.* 2003), soit le rayonnement X dur provient d'un AGN d'arrière plan. Il reste cependant quelques candidates pour lesquelles le rayonnement X dur semble bien provenir de la naine blanche et dont l'origine n'est toujours pas certaine (O'Dwyer *et al.* 2003).

c Étoiles à neutrons isolées et pulsar

Les étoiles à neutrons se forment quand la masse d'un astre qui s'effondre sur lui-même dépasse la limite de Chandrasekhar qui se situe à environ 1.4 $M_\odot$. La force de gravité est telle que même la pression de dégénérescence des électrons confinés ne parvient pas à lui résister. Ces derniers finissent par être collés au noyau de leur atome où ils interagissent avec les protons pour former des neutrons[4]. Les étoiles à neutrons naissent après l'explosion en supernovae de l'étoile d'origine dans un milieu dont les températures sont de l'ordre de 10^{11} K (Becker *et al.* 2008, §14.2.2.2). A leur naissance, les étoiles à neutrons ont des températures internes comprises entre un et dix milliards de degrés K.

Elles font partie des objets dans la nature qui possèdent les plus forts champs magnétiques. L'origine de ces champs n'est pas bien connue mais doit venir d'un phénomène d'amplification lors de l'effondrement en étoile à neutrons du cœur de l'étoile juste avant son explosion en supernovae. Les magnétars, étoiles à neutrons les plus magnétisées, ont des champs de l'ordre de 10^{14} à 10^{15} Gauss (Lewin & van der Klis 2006, §1.2).

Les jeunes pulsars se trouvent souvent au milieu de la nébuleuse laissée par la supernovae qui les a engendrés. L'origine de l'émission X provient de plusieurs mécanismes thermiques et non-thermiques :

- pour les pulsars jeunes ayant un champ magnétique supérieur à 10^{12} Gauss, on constate une émission non thermique en loi de puissance qu'on attribue à l'émission cyclotron de particules accélérées dans la magnétosphère ;
- possible détection d'émission X synchroton venant de la nébuleuse ;
- émission thermique de corps noir venant de la surface chaude jusqu'en X mou ;

[4] Plus exactement, les protons se transforment en neutrons par réaction β inverse ($p + e^- \rightarrow n + \nu_e$, avec ν_e un neutrino électronique) alors que la réaction β ($n \rightarrow p + e^- + \nu'_e$) est inhibée par la trop forte concentration d'e^- dans le milieu.

– émission thermique X molle venant des pôles magnétiques chauffés par un flux de particules chargées venant de la magnétosphère.

En fonction de la puissance du champ magnétique, de la température de l'étoile à neutrons, etc., l'émission dominante change.

On connaît à l'heure actuelle un groupe de 7, et probablement 8 (Pires *et al.* 2009), étoiles à neutrons isolées pour lesquelles on ne détecte pas d'émission radio (*radio quiet*) et dont les spectres X, purement thermiques, s'apparentent à des spectres de corps noirs avec un peu d'absorption photoélectrique venant du milieu interstellaire. Ce groupe viendrait de l'association OB nommée Sco OB2 qui se situe dans la ceinture de Gould (Popov *et al.* 2003, Motch *et al.* 2003, 2009). Pour expliquer à la fois l'émission X et l'émission optique, on a besoin de faire intervenir, là aussi, des points chauds situés aux pôles de l'étoile à neutrons.

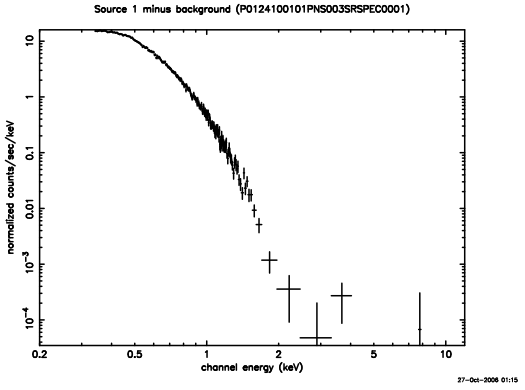


FIG. I.8 – Spectres EPIC PN de l'étoile à neutron isolée RX J0720.4-3125. Ce spectre, comme celui de la naine blanche Fig. I.7, est mou (très peu de photons détectés au delà de 1.5 keV) et régulier, il correspond également à une émission thermique de corps noir de température supérieure à celle de la naine blanche Fig. I.7.

Le lecteur intéressé par le sujet pourra lire l'ouvrage de Becker (2008).

1.4 LES BINAIRES X (XRB)

Les systèmes binaires X (XRB pour *X-Ray Binaries*) sont des systèmes binaires dans lesquels a lieu un transfert de masse entre une étoile et un astre compact : étoile à neutrons ou trou noir. Ces systèmes avec échange de matière entre les deux composantes sont dits en interaction.

Le puits de potentiel gravitationnel est beaucoup plus profond que pour les variables cataclysmiques (voir §1.5). L'énergie libérée par le transfert de masse sous forme de rayonnement est donc potentiellement plus important. Le nom de ces objets vient du fait que la majorité de leur émission est rayonnée dans les longueurs d'onde X. Cette émission varie typiquement entre 10^{34} et 10^{38} erg.s^{-1} dans la bande 0.5–10 keV.

Dans la suite, on utilisera plusieurs abréviations : les trous noirs seront notés BH pour *black hole*, les étoiles à neutrons seront notées NS pour *neutron star*. On appellera « primaire » l'astre compact accrétant et « secondaire », « donneur » ou « compagnon » l'astre dont la matière est accrétée. Il existe une large gamme de XRB dont les propriétés diffèrent essentiellement en fonction de la nature de la primaire, de son activité magnétique, de la période orbitale, de l'inclinaison de l'orbite par rapport à l'observateur, et enfin du type spectral et de la masse du compagnon. Le lecteur intéressé par le sujet pourra lire les ouvrages Lewin & van der Klis (2006), Staubert (2008), Tanaka (2008) auxquels nous ferons de multiples références dans cette section.

Les binaires X se divisent en plusieurs familles en fonction de la masse du compagnon, et en plusieurs sous-familles en fonction notamment de leurs caractéristiques observationnelles :

1. dans les binaires X de faible masse (LMXB pour *Low Mass X-ray Binaries*) le compagnon est une étoile de faible masse ($M \leq 2M_{\odot}$) de type spectral allant de A à M et qui a suffisam-

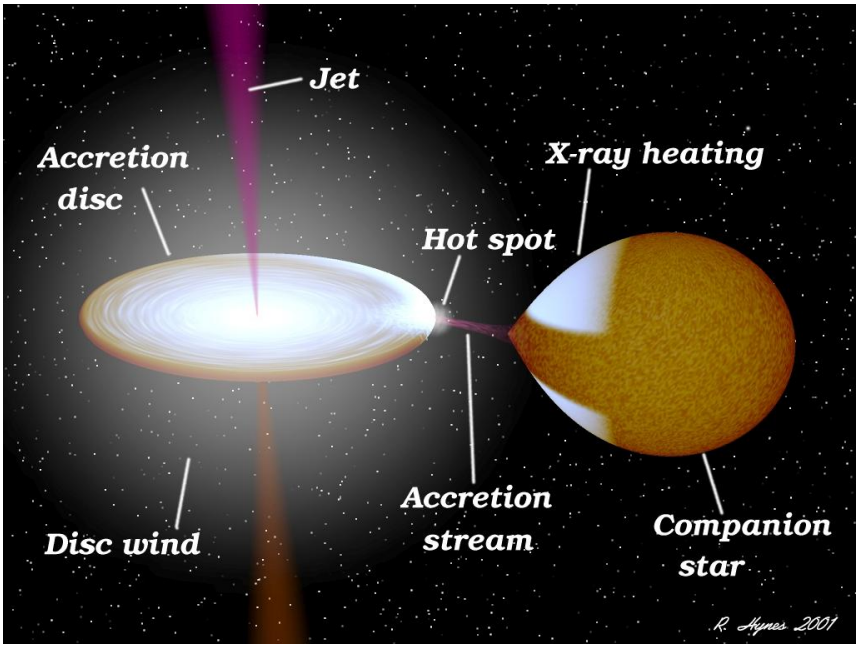


FIG. I.9 – Schéma synthétique d'un système binaire X semi-détaché non magnétisé (crédit Rob Hynes).

ment évolué, ou est suffisamment proche de la primaire, pour remplir son lobe de Roche. Le transfert de masse a lieu par le point de Lagrange interne L1 et l'émission X provient du disque d'accrétion et/ou de l'astre compact avant d'être éventuellement reprocessée et/ou comptonisée en partie. Le rapport Fx/Fopt, variable selon les sources, a une valeur comprise entre 20 et 1000 dans la bande 0.5–10 keV (Staubert 2008, §15.4). Ces objets possèdent de nombreux phénomènes et classes de variabilités parmi lesquels figurent :

- les transitoires X molles (SXT pour *Soft X-ray Transients*). Ce sont des sources qui alternent entre de longues périodes de quiescence ($\sim$1 à 10 ans) pendant lesquelles elles sont difficilement détectables, et des phases éruptives courtes ($\sim$ quelques semaines) pendant lesquelles leur luminosité approche la luminosité d'Eddington ;
- les sources dites ADC, pour *accretion-disk-corona*, pour lesquelles les rayons X venant de l'objet central sont diffusés dans la direction de l'observateur par les électrons de la couronne chaude qui, ayant une taille plus grande que le compagnon, a pour effet de lisser la courbe des éclipses X ;
- les *dippers*, qui représentent environ 10% des LMXB, sont des sources pour lesquelles on observe des éclipses X peu profondes qui peuvent venir du flux d'accrétion entre la secondaire et le disque, du point chaud (*hot spot*), ou toute autre structure en corotation avec le système ;
- les *bursters* X et les *Rapid Bursters* sont des sources qui présentent des sursauts, ou *bursts*, X dont la durée est de l'ordre de quelques minutes. Il s'agit de binaires X ayant pour astre compact une étoile à neutrons faiblement magnétisée. Elles présentent des sursauts de type I et, pour les *Rapid Burster*, des sursauts de type II :
  - les sursauts X de type I sont des réactions thermonucléaires ayant lieu à la surface d'étoiles à neutrons pas ou peu magnétisées (voir paragraphe 1.4.3) ;
  - les sursauts X de type II « *ont pour origine une instabilité de l'accrétion au voisinage de l'objet compact* [... ,] *un temps de récurrence de l'ordre de quelques minutes et* [...] *des comportements temporels assez complexes* » (Viallet 2008, §2.2) ;
- les QPO pour *Quasi Periodic Oscillations sources*, parmi lesquelles ont trouve les sources Z, les sources Atoll et les kHz QPO.

2. dans les binaires X de masse intermédiaire, IMXB pour *Intermediate Mass X-ray Binaries*, le compagnon possède une masse $\sim$ 2–4 $M_\odot$. On pense que de tels systèmes ne sont pas

stables dynamiquement. Leur observation est donc exceptionnelle. Il se peut que certains IMXB évoluent très rapidement et deviennent des LMXB. On suspecte notamment que ce soit le cas de Cyg X-2, (Podsiadlowski *et al.* 2002).

3. dans les binaires X de masse importante, HMXB pour *High Mass X-ray Binaries*, le compagnon est une étoile jeune massive de type Be ($M \geq 5M_{\odot}$) ou OB ($M > 15M_{\odot}$). L'astre compact accrète de la matière à partir du puissant vent stellaire émis par son compagnon. L'émission X provient à la fois de la primaire et de la secondaire et le rapport Fx/Fopt dans la bande 0.5–10 keV est de l'ordre de l'unité (Staubert 2008, §15.3). La matière accrétée ne forme pas nécessairement de disque. Parmi ces systèmes, on trouve :
   – les binaires X à éclipses, ou *eclipsing X-ray binaries* ;
   – les *Be/X-ray binaries* (BeXRB) ;
   – les *Supergiant X-ray binaries* (SGXB) ;
   – les *Supergiant Fast X-ray Transients* (SFXT).

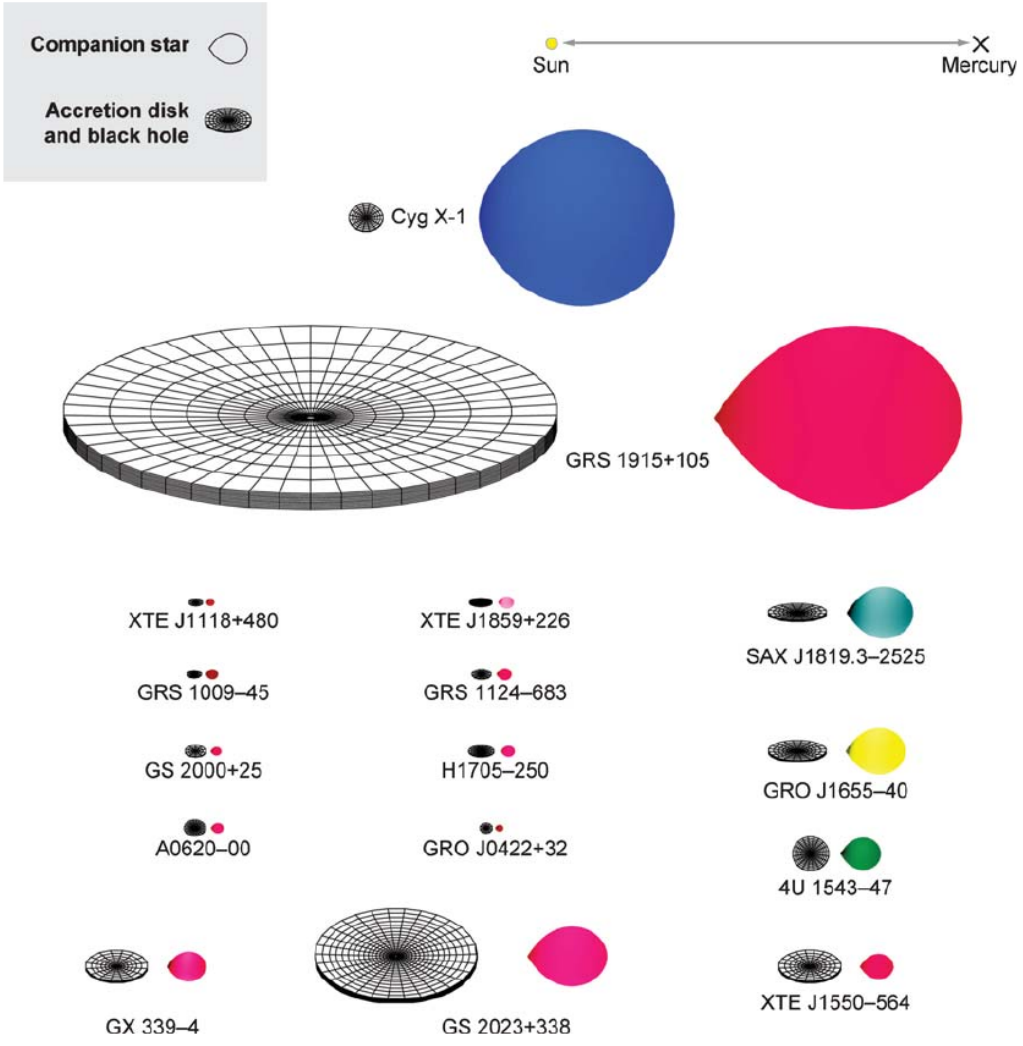


FIG. I.10 – Schéma synthétique montrant les échelles de différents systèmes binaires X à trou noir (extrait de Remillard & McClintock (2006)).

Certains objets sont similaires, qu'ils appartiennent aux LMXB ou aux HMXB. C'est notamment le cas des pulsars X accrétants et des *microquasars* qui présentent une forte émission radio et des jets. Enfin, les sources X ultra lumineuses, ULX pour *ultra luminous X-ray sources*, sont encore mal connues et sont classées uniquement sur la base de leur luminosité X exceptionnelle (voir section ci-après).

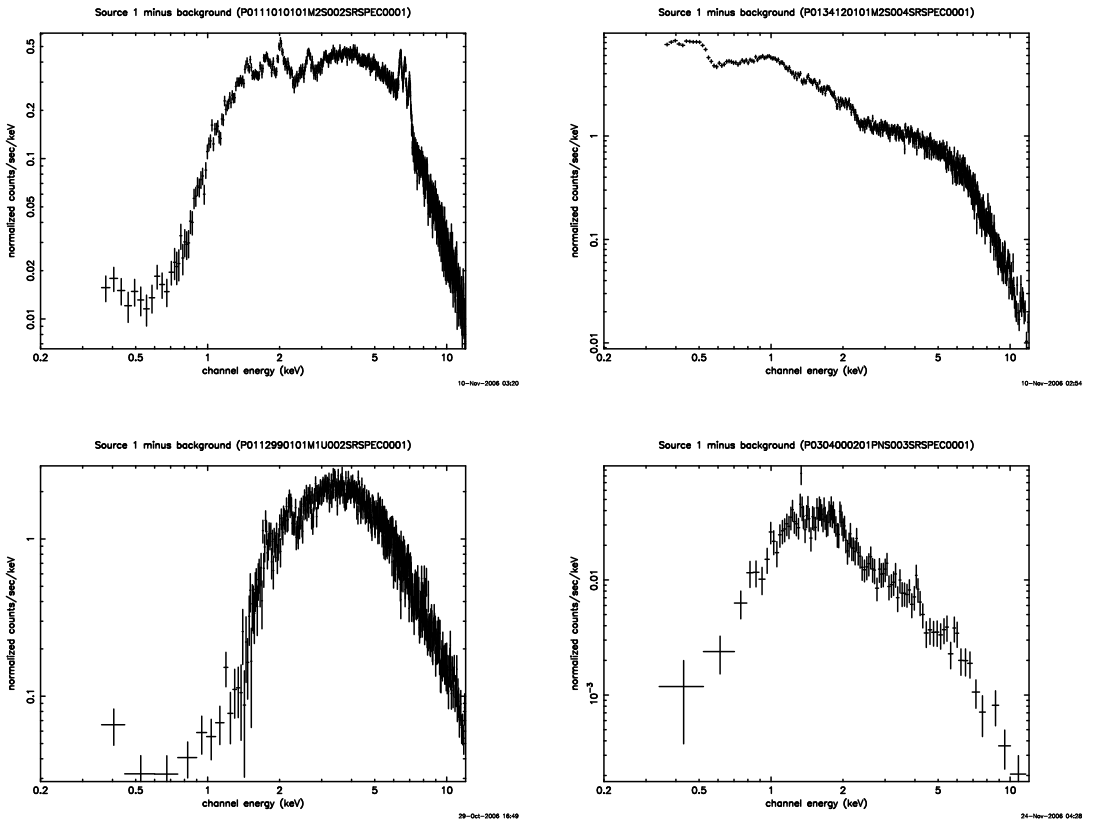


FIG. I.11 – Spectres EPIC PN ou MOS de quatre binaires X. En haut à gauche : HMXB Cen X-3 qui est un pulsar à eclipses. En haut à droite : IMXB Her X-1 dont l'astre compact est une étoile à neutrons. En bas à gauche : LMXB V1487 Aql qui est une source transitoire X (*X-ray transient*). En bas à gauche : LMXB V404 Cyg qui est une source transiente et dont la secondaire est de type K0III. À l'exception du deuxième, c'est spectres sont dur avec peu de photons reçus en dessous de 1.5 keV, la plupart des photons X mou étant reprocessés. Pour plus d'informations, voir par exemple Tugay & Vasylenko (2009).

La durée de vie des HMXB est contrainte par le temps que passe le compagnon massif sur la séquence principale. Cette valeur est typiquement de l'ordre de 10^5 à 10^7 ans (Lewin & van der Klis 2006, §1.1.1) et les HMXB sont localisées dans le plan galactique où se trouvent les étoiles jeunes. En revanche, du fait de leur durée de vie plus longue, de 10^7 à 10^9 ans, les LMXB sont associées à des populations plus âgées et se situent plutôt dans le centre galactique et dans les amas globulaires. On pense que notre Galaxie ne compte que quelques centaines de HMXB et de LMXB.

Les LMXB sont souvent significativement plus brillantes en X qu'en optique où l'on voit essentiellement le *reprocessing* d'une partie des rayons X mous émis. Même un faible taux de *reprocessing* en optique suffit à dominer l'émission du compagnon. Ce qui explique que le rapport Fx/Fopt est de l'ordre de 100 à 1000 sauf dans les ADC (*accretion-disk-corona*) où ce même rapport vaut ~ 20. On connaît aujourd'hui environ 150 LMXB (Liu *et al.* 2001) dont 20% contiennent des BH. Les NS sont souvent peu magnétisées ($B < 10^{10}$ Gauss) et sont donc plus des *burster* et des QPO que des pulsars (Staubert 2008, §15.4).

Les HMXB, elles, ont une valeur Fx/Fopt de l'ordre de l'unité. Seulement 4% des HMXB contiennent des BH, 50% des pulsars, et parmi les pulsars seulement 50% des spectres présentent des raies cyclotron (Staubert 2008, §15.3). On connaît aujourd'hui environ 130 HMXB (Liu *et al.* 2001).

1.4.1 Pulsars accrétants

Parmi les étoiles à neutrons que l'on trouve dans les XRB, certaines sont des pulsars. La période de pulsation de ces objets est de l'ordre de la seconde dans les HMXB et de la milliseconde dans les LMXB qui possèdent une période orbitale courte. Le flux d'accrétion sur l'étoile à neutron suit les lignes de champ et on interprète l'émission X comme étant l'émission modifiée par comptonisation d'un rayonnement thermique (Bremsstrahlung) venant des régions chauffées par le flux de matière accrétée au niveau des pôles magnétiques. La comptonisation par des électrons chauds aurait lieu dans les colonnes d'accrétion juste au-dessus des pôles magnétiques. S'ajoute éventuellement une petite contribution venant des autres mécanismes d'émission X décrits plus haut. Pour les étoiles à neutrons ayant un champ B $\sim 10^{12}$ Gauss, on s'attend à voir des raies cyclotron dans le spectre X.

1.4.2 Étoiles à neutrons faiblement magnétisées et trous noirs

Tous les objets compacts de plus de 3 $M_\odot$ sont théoriquement des trous noirs. Du fait de leur nature particulière impliquant la présence d'un horizon, ces objets ne disposent pas de champ magnétique. En dessous de $B \sim 10^8$ Gauss, une étoile à neutrons possède un champ magnétique trop faible pour être un pulsar (Lewin & van der Klis 2006, §1.3). Le disque d'accrétion n'est alors plus tronqué et apparaît une couche limite entre le disque et l'étoile. Cette couche est l'interface au niveau de laquelle la matière passe du disque sur la surface de l'étoiles à neutrons.

La plupart des modèles actuels qui tentent de décrire l'émission X des étoiles à neutrons et des trous noirs accrétants comportent à la fois une atmosphère d'électrons ténue chaude (10-100 keV) et un disque d'accrétion géométriquement mince. L'émission X molle vient des bords internes du disque d'accrétion. La plus grande partie de l'émission optique est associée au phénomène de *reprocessing* en optique des rayons X par les couches supérieures du disque irradié. La luminosité X observée dépend beaucoup de l'inclinaison du système et c'est logiquement que les sources X les plus faibles ont des éclipses ou des *dips*.

La plupart des binaires X avec BH et une grande partie de celles avec des NS ne sont pas constamment très lumineuses en X. Ce sont des systèmes transitoires dont l'émission X peut essentiellement être décrite par trois états : un état de luminosité X importante, un état de luminosité X plus faible et un état quiescent. S'ajoute, dans le cas des étoiles à neutrons d'autres phénomènes ponctuels comme les sursauts, ou *bursts*, de type I.

Les systèmes transitoires ont de longues périodes d'inactivité, allant de plusieurs mois à plusieurs dizaines d'années. On dit alors qu'ils sont dans un état quiescent. Entre deux états quiescents, ces systèmes connaissent de courtes éruptions (ou *outbursts*) pendant lesquels leur magnitude en X augmente de plusieurs ordres de grandeur. La fraction de transitoires par rapport aux persistants pour les BH XRB est supérieure à cette même fraction pour les NS XRB, mais leurs *outbursts* sont plus longs et plus rares (Lewin & van der Klis 2006, §1.3.1).

Deux théories ont été proposées pour expliquer les systèmes transitoires. Dans les deux, de la matière s'accumule sur les bords externes du disque d'accrétion pendant la phase de quiescence. Le transfert de masse est alors relativement faible entre le bord interne du disque et l'astre compact (AC). En *outburst*, la source devient brillante en X à cause d'une augmentation du transfert du disque à l'AC. Dans le premier modèle, la variation non linéaire du transfert de masse à travers le disque est due à des instabilités thermiques (Narayan *et al.* 1996). Dans le second, c'est le taux de transfert de masse entre l'étoile compagnon et le disque d'accrétion qui change en raison de l'irradiation X de l'étoile. Cette deuxième possibilité a été étudiée par Viallet (2008) lors d'un travail de thèse dont les résultats suggèrent un effet marginal de ce second mécanisme.

a Phase de luminosité X importante

Le disque multicouleur, MCD pour *MultiColor Disk*, est un modèle de disque d'accrétion basé sur le modèle standard de Shakura & Sunyaev (1973) (Mitsuda *et al.* 1984, Makishima *et al.* 2000). Ce modèle considère un disque constitué de plusieurs anneaux émettant un rayonnement de corps noir à différentes températures. La couleur de chaque anneau est donc différente. La température décroît avec l'éloignement au centre : $T(R) \propto R^{-3/4}$, $T_{in} \propto M^{-1/4}$, T_{in} se situant sur le bord interne (Makishima *et al.* 2000).

L'émission X molle (*soft*) détectée dans les binaires X ayant une luminosité X importante est compatible avec l'origine thermique associée à un disque MCD géométriquement mince, mais optiquement épais. On localise donc cette émission au niveau des bords internes du disque où les températures sont les plus importantes. La composante dure (*hard*), non thermique, suit une loi de puissance d'indice ~ 2.5 (Tanaka 2008, §16.4.3.1) et aurait pour origine un processus de *Comptonisation*, aussi appelée émission Compton inverse (voir section 1.1.4).

Dans le cas des étoiles à neutrons, la composante s'accorde également avec le modèle MCD, mais avec une température $kT_{in} \sim 1.5$ keV, supérieure à celle associée aux trous noirs. La composante thermique la plus dure est associée à une température ~ 2 keV qui proviendrait de la surface même de l'étoile à neutrons. La ressemblance entre le spectre de cette composante et celui des *burst* de type I abonde dans ce sens.

b Phase de luminosité X plus faible

Cette phase se distingue de la phase de luminosité X importante par une émission X plus dure mais d'intensité moindre dans la bande X classique. Elle a lieu à la fois pour les NS et les BH et on pense qu'elle correspond à un état différent du disque d'accrétion, dépendant du taux d'accrétion. L'émission X entre NS et BH est similaire dans la bande 0.1-10 keV, si bien qu'on ne peut discriminer, dans cette bande, entre les deux types d'astres compacts quand ils sont dans cet état. Cependant, à plus haute énergie il y a une cassure dans le spectre dur des étoiles à neutrons qui permet de les distinguer des BH.

Les propriétés X ressemblent globalement, à plus petite échelle, à celles des AGN (cf. §1.7.2). Les spectres sont durs et suivent une loi de puissance d'indice de photon typique $\sim 1.7 - 1.9$. Physiquement, on les interprète comme le résultat de *Comptonisation* thermique, c'est-à-dire de la *Comptonisation* de photons X mous avec des électrons d'origine thermique de température typique $kT \leq 100$ keV. On situe cette diffusion Compton inverse de photons mous dans une couronne entourant le disque d'accrétion.

c Phase de quiescence

Les transitoires X molles, *soft X-ray transients*, passent la majeure partie de leur temps en phase quiescente. Durant cette phase, la luminosité X des objets est beaucoup plus faible et est différente entre BH LMXB et NS LMXB : les NS LMXB sont plus lumineuses que les BH LMXB mais, alors que ces dernières ont un spectre en loi de puissance d'indice ~ 2, celui des NS LMXB ressemble en général à un corps noir de température $kT \leq 0.1$ keV en plus d'une partie dure. On pense que cette partie molle supplémentaire provient de l'émission thermique de l'étoile à neutrons. La luminosité optique, pendant cette phase quiescente, est dominée par l'émission de l'étoile compagnon.

1.4.3 Sursaut X de type I

Les *bursts* X de type I ont lieu à la surface d'étoiles à neutrons faiblement magnétisées. La matière accrétée est compressée à des densités et des températures auxquelles la combustion thermonucléaire de l'hélium devient instable. Se produit alors un flash thermonucléaire dont l'allumage est rapide (1 seconde) et provoque une augmentation de la luminosité X du système. Cette augmentation de luminosité soudaine est suivie par une décroissance exponentielle douce qui dure plusieurs dizaines de secondes (Staubert 2008, §15.6.3). Le spectre de l'émission est compatible avec un spectre de corps noir. Ces réactions thermonucléaires de surface ne peuvent pas avoir lieu sur un trou noir. On rencontre le même type de phénomène dans les novae où l'astre accrétant est une naine blanche.

On connaît 7 LMXB pour lesquelles ont lieu des *superbursts*. Presque 100 fois plus énergétique que les *bursts* classiques, la décroissance lumineuse dure de 1 à 6 heures. On pense que dans ces *superbursts* le combustible n'est pas l'hélium mais le carbone.

1.4.4 Sources X ultra-lumineuses (ULX)

Les sources X ultra-lumineuses, ou ULX pour *Ultra Luminous X-ray sources*, sont des sources dont la nature exacte est encore débattue (Fabbiano 2004). « *La définition communément admise pour*

les ULX est celle d'une source X qui ne se situe pas au centre de sa galaxie hôte [, il ne peut donc pas s'agir d'un AGN,] *et dont l'émission X dans la bande 0.5-10 keV est supérieure à* 10^{39} *erg.s*$^{-1}$ » (extrait de la thèse de Fabien Grisé, 2008). Il semblerait qu'il s'agisse de systèmes binaires formés d'un trou noir et d'une étoile. L'essentiel du débat porte sur la masse du trou noir. D'un coté l'émission X semble dépasser la limite d'Eddington de trous noirs accrétants de masse stellaire ($\sim$10 $M_\odot$). D'un autre coté, on explique mal la présence de trous noirs de masse intermédiaire ($M \sim 10^3 - 10^4\ M_\odot$) en dehors des centres galactiques et des amas globulaires. La température interne du disque ajusté sur les spectres X a semblée un temps trop élevée (Makishima *et al.* 2000) : avec une valeur kT_{in} de 1.1–1.8 keV, elle dépasse celle des trou noir de masse stellaire ($\sim$ 0.4–1.0 keV) alors que la température est invresement proportionelle à la masse du BH ($T_{in} \propto M^{-1/4}$). Mais l'ajustement de (Makishima *et al.* 2000) était uniquement basé sur un modéle MCD à partir de spectres ASCA. Depuis, d'autre ajustement ont été fait sur des spectres XMM avec des modéles qui mélange une loi de puissance avec une le MCD. Les résultats sont au contraire en faveur de trou noir de masse intermédiaire avec de température de l'ordre de 0.2 keV (Stobbart *et al.* 2006).
La figure I.12 extraite de Roberts (2007) montre le spectre d'un ULX et son ajustement à partir de différentes composantes. A l'heure actuelle, on dispose d'un échantillon d'environ 110 candidats ULX (Liu & Bregman 2005). Mais cet échantillon est sûrement biaisé par d'éventuels AGN d'arrière-plan.

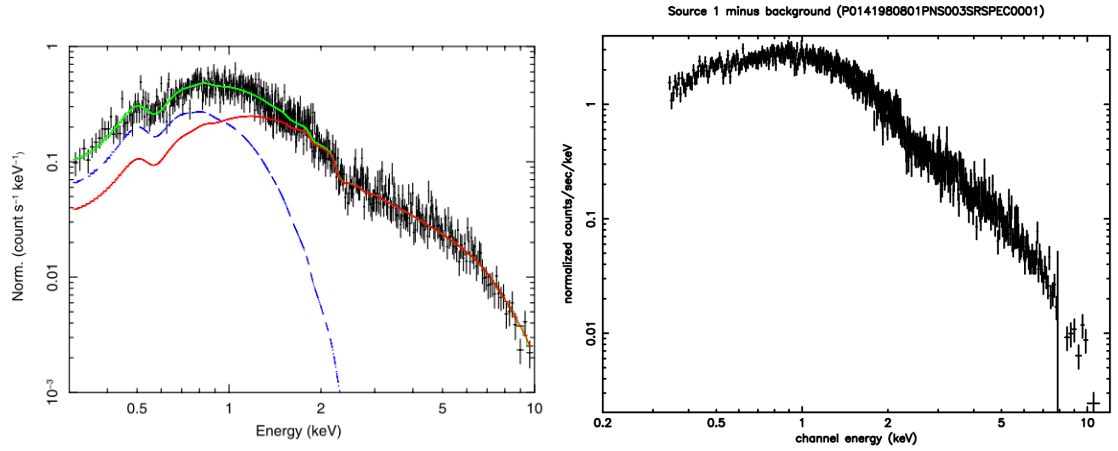


FIG. I.12 – Gauche : Spectre EPIC de NGC 1313 X-1 extrait de Roberts (2007). En vert, meilleur ajustement d'un modèle de IMBH. Il résulte de la superposition de spectres de corps noirs d'un disque multicouleurs de température $kT_{in} \sim 0.2keV$ (en bleu) et d'une composante en loi de puissance cassée (*broken power-law* venant d'une couronne optiquement épaisse) d'indice de photon $\alpha \sim 1.7$ qui domine ce spectre (en rouge). Droite : spectre EPIC PN de l'ULX M33 X-8.

Comme pour les LMXB, on évoque un modèle de disque multicouleurs (MCD) pour la composante thermique. Quant à la composante non thermique, à base de loi de puissance, on l'interprète comme la diffusion Compton inverse de photons mous dans une couronne ayant une certaine profondeur optique τ et température électronique $kT_{couronne}$.

1.5 VARIABLES CATACLYSMIQUES (CV)

Cette partie se base essentiellement sur Warner (1995), Lewin & van der Klis (2006), Beuermann (2008), Viallet (2008) dans lesquels le lecteur curieux trouvera de nombreuses informations.

1.5.1 Généralités sur les CV non magnétisées

Les variables cataclysmiques sont des systèmes binaires composés d'une naine blanche, WD pour *White Dwarf*, accrétant de la matière à partir d'une étoile compagnon de masse plus faible. Bien que la secondaire soit souvent une étoile de type M ou K, il peut également s'agir d'une étoile G, d'une naine brune (*brown dwarf*) ou d'une autre WD. L'énergie libérée par la matière tombant

dans le champ de gravité de la WD est à l'origine de la plupart des phénomènes X observés dans ces systèmes. Comme le puits de potentiel est plus faible que dans les binaires X, la conversion de l'énergie gravitationnelle de la matière accrétée en rayonnement est moins importante que pour ces dernières, et les processus thermonucléaires deviennent plus efficaces. De fait en dehors des phases éruptives, le disque d'accrétion rayonne principalement dans le visible et l'ultraviolet (UV) alors qu'il émet surtout en X dans les XRB. L'émission des CV dans le rouge et l'infrarouge provient avant tout de la secondaire, les UV et les rayons X de la surface ou de la proximité immédiate de la WD.
La plupart du temps, le donneur est une étoile de la séquence principale peu évoluée. Ce n'est en revanche pas le cas des binaires symbiotiques dans lesquelles le donneur appartient à la branche des géantes ou à la branche asymptotique des géantes (AGB).

La majorité des CV sont des analogues aux LMXB et ont des périodes orbitales allant de 75 minutes à 8 heures. Pour les périodes supérieures, allant de 8h à 3 jours, le donneur est une étoile sous-géante évoluée. Entre 3 et 200 jours, le système est instable si le donneur dépasse 67% de la masse de la WD. Enfin, au delà de 200 jours, on trouve les CV symbiotiques : il n'y a pas de transfert de masse par débordement du lobe de roche mais, à la manière des HMXB, la WD orbite dans le vent d'une super-géante et le capte. A l'autre extrémité, pour des périodes orbitales inférieures à 60 minutes, on trouve les systèmes de type AM CVn : ces systèmes sont composés de deux WD et la matière accrétée est essentiellement constituée d'hélium.

On distingue parmis les CV plusieurs sous-familles en fonction de l'évolution temporelle de leur courbe de lumière en optique :

1. les novae classiques sont des systèmes dont on a observé une éruption. Cette éruption, pendant laquelle la luminosité optique peut augmenter de 6 à 19 magnitudes, est anti-corrélée avec sa durée. Ces éruptions sont la conséquence du déclenchement de réactions thermonucléaires de surface pendant lesquelles l'hydrogène accumulé par accrétion sur la WD est consommé. On pense que les réactions thermonucléaires commencent quand la couche dépasse $10^{-4}\ M_{\odot}$. Les novae sont l'équivalent des *bursts* X ayant lieu dans les LMXB.
2. les novae naines , ou *dwarf novae*, sont des systèmes à éruptions récurrentes, dont l'amplitude varie de 2 à 5 magnitudes. La durée entre deux éruptions varie entre quelques jours et quelques semaines. L'origine de ces éruptions réside dans des instabilités du disque d'accrétion, à l'instar des transitoires X molles. Parmi les novae naines, on distingue les types :
   - U Gem pour lesquelles les éruptions, ou *outbursts*, durent quelques jours et se reproduisent sur des échelles de temps de l'ordre de la semaine. SS Cygni fait partie de cette classe.
   - Z Cam pour lesquelles les éruptions sont occasionnellement interrompues et stabilisées à des valeurs de luminosités intermédiaires entre les phases quiescentes et éruptives.
   - SU UMA dont la période orbitale est courte. Ces systèmes présentent des super éruptions qui sont plus brillantes et durent 5 fois plus longtemps que les éruptions normales. Ce phénomène est interprété comme étant dû à des forces de marées exercées par le donneur sur le disque externe, augmentant la viscosité de ce dernier. Pour plus de précisions, voir Viallet (2008) §5.1.3.
3. les novae récurrentes sont des novae classiques dont on a observé plusieurs éruptions.
4. la classe des *novae like* regroupe l'ensemble des variables cataclysmiques non-éruptives. Pour un certain nombre d'entre elles, il s'agit probablement de systèmes éruptifs dont les éruptions n'ont pas encore été observées (pré-nova, post-nova, novae naine au repos, ...).
5. les binaires X symbiotiques (SXB pour *Symbiotic X-ray binary*) qui ont également des réactions thermonucléaires de surface. Mais dans ce cas, la matière accumulée sur la WD provient du vent d'une géante rouge et les *outbursts* durent plusieurs dizaines d'années.

1.5.2 Les CV magnétisées : Polars et Polars intermédiaires

Les polars, ou étoiles de type AM Her, et les polars intermédiaires, IP pour *Intermediate Polar*, ou encore étoiles de type DQ Her, représentent environ 10% des CV (Downes *et al.* 2001). Dans ces systèmes, la WD possède un champ magnétique supérieur à respectivement 10^7 et 10^6 Gauss. Ce champ magnétique empêche la formation d'un disque, complètement dans le cas des polars

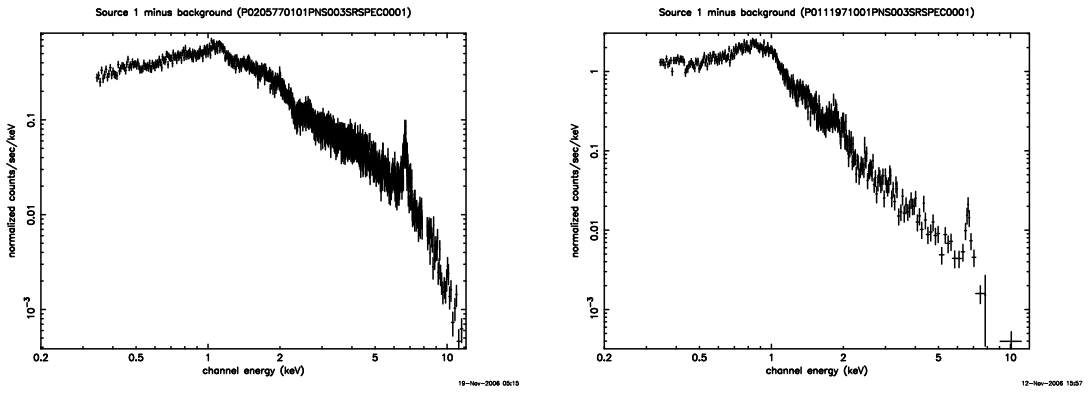


FIG. I.13 – Spectre EPIC PN de la *Dwarf Nova* Z CHA (gauche) et de la *Nova-like* IX Vel (droite). Pour une explication détaillée d'un spectre similaire (celui de OY Car) voir Ramsay *et al.* (2001).

et partiellement dans le cas des IP. Dans ce dernier cas, la partie interne du disque d'accrétion est tronquée et la matière est transférée du disque à la WD par des canaux qui suivent les lignes de champ. Pour les polars, la matière est directement accrétée depuis la primaire, et, contrairement aux IP, la période de rotation de la WD est contrainte (*locked*) par la période orbitale du système.

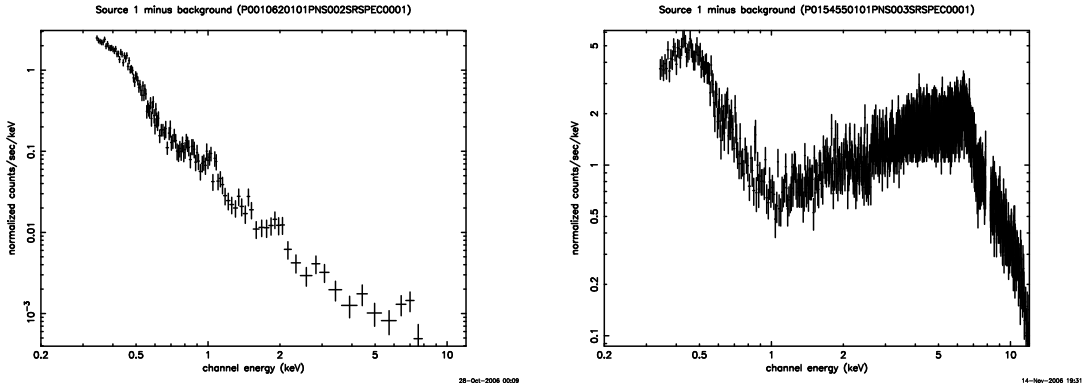


FIG. I.14 – Spectre EPIC PN du polar V1309 Ori (gauche) et du polar intermédiare GK Per (droite). Pour plus d'information sur ce dernier spectre, voir Vrielmann *et al.* (2005).

Les polars sont plutôt des émetteurs X mous du fait du *reprocessing* des X durs sur la surface de la WD (Lewin & van der Klis 2006, §10.3.2). Un modèle simple expliquant l'origine de ces rayons X durs reprocessés invoque une émission de type bremsstrahlung thermique qui proviendrait de plasmas stockés quand la matière accrétée rentre en collision avec l'atmosphère de la WD. Pour plus d'information sur l'émission X des polars voir par exemple le §6.4 de Warner (1995).

1.5.3 Sources X super-soft (SSS)

Les sources X super molles, SSS pour *luminous SuperSoft X-ray Sources* sont comme leur nom l'indique des sources intrinsèquement très lumineuses possédant des spectres X très mous. Une quarantaine de ces sources ont été découvertes par ROSAT. Leurs luminosités bolométriques, obtenues par ajustement d'un spectre d'atmosphère stellaire sur les données X, sont proches de la limite d'Eddington et vont de 10^{36} à environ 10^{37} erg.s^{-1}. Le maximum de leur spectre X se trouve dans une gamme d'énergie allant de 15 à 80 eV, ce qui équivaut à un rayonnement de corps noir ayant des températures de l'ordre de 300 000 à 500 000 K (Kahabka & van den Heuvel 1997). Ces températures sont deux ordres de grandeur plus faibles que celles des binaires X classiques.

L'estimation de ces luminosités absolues, élevées, ne semble pas être biaisée par une éventuelle focalisation de l'emission X. En effet, à l'instar des ULX, des regions ionisées faisant plusieurs dizaines de parsec de diamètre ont été détectées autour de certaines de ces sources, notamment CAL 83 dont le spectre EPIC est présenté Fig. I.15, ce qui tenderait à prouver que l'emission X est isotrope ou quasi isotrope.

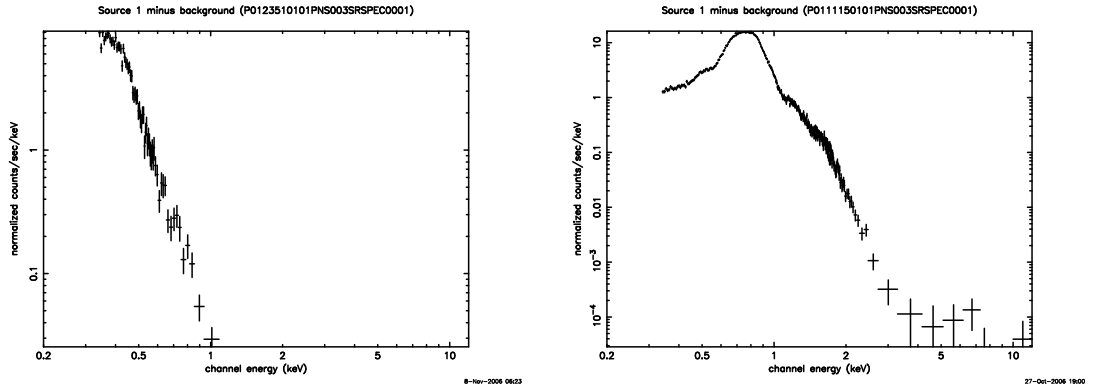


FIG. I.15 – Spectre EPIC PN des SSS CAL 83 (droite) et RX J0925.7-4758 (gauche). Pour plus d'informations sur CAL 83 et son spectre X, voir par exemple Lanz *et al.* (2005).

Les caractéristiques observées des ces sources correspondent à des naines blanches qui consument progressivement, ou de façon cyclique, de la matière riche en hydrogène dans leurs enveloppes via le cycle CNO. Cette matière est accrétée sur leurs surfaces à des taux de l'ordre de $10^{-7}\ M_\odot$ par an, contraignant les possibilités de CV formant les SSS à trois types :

1. les CV où le compagnon, proche, est une étoile de la séquence principale dont la masse, supérieure à celle de la WD, est comprise entre 1.3 et 2.5 $M_\odot$. Ces systèmes sont appelés CBSS pour *Close Binary Supersoft Sources*.
2. les binaires symbiotiques dont la secondaire est une géante rouge de masse solaire.
3. les binaires symbiotiques dont la secondaire est une géante rouge sur la branche asymptotique des géantes (AGB) et dont la perte de masse par un vent puissant avoisine les $10^{-5} M_\odot$ par an. La vitesse du vent est faible, de l'ordre de 30 km.s^{-1}.

Enfin, notons que les SSS sont suspectées d'être à l'origine de supernovae de type Ia. Le lecteur curieux trouvera des informations supplémentaires au sujet des SSS dans Kahabka & van den Heuvel (1997).

1.6 RESTES DE SUPERNOVAE

Les restes de supernovae, SNR pour *SuperNova Remnants*, émettent eux aussi des rayons X. Cette émission peut venir de l'astre compact central – étoile à neutrons ou pulsar – dans le cas où le progéniteur n'est pas complètement détruit par l'explosion (SNR de type Ib/c). Mais elle vient essentiellement du matériau chauffé par les chocs dans la coquille en expansion et à l'interface avec le milieu interstellaire. Le plasma peut atteindre des températures de l'ordre de 10^6 à 10^7 K (Vink 2004). L'émission X vient alors de raies d'émission par recombinaison, de Bremsstrahlung thermique, mais aussi de l'émission synchrotron des électrons les plus énergétiques accélérés au niveau de ces ondes de chocs. Pour en savoir plus sur les processus d'accélération, le lecteur pourra se référer à Reynolds (2008).

Les spectres X des supernovae présentent souvent des fortes raies du fer Fe K à 6.5 keV, de Si et, pour les Type Ia, de Fe L, mais aussi des raies O, Ne, Mg et Ca. La largeur des raies est directement liée à la dispersion de la distribution des vitesses des émetteurs sur la ligne de visée.

On trouvera par exemple une étude détaillée de la SNR Tycho dans Cassam-Chenaï *et al.* (2004).

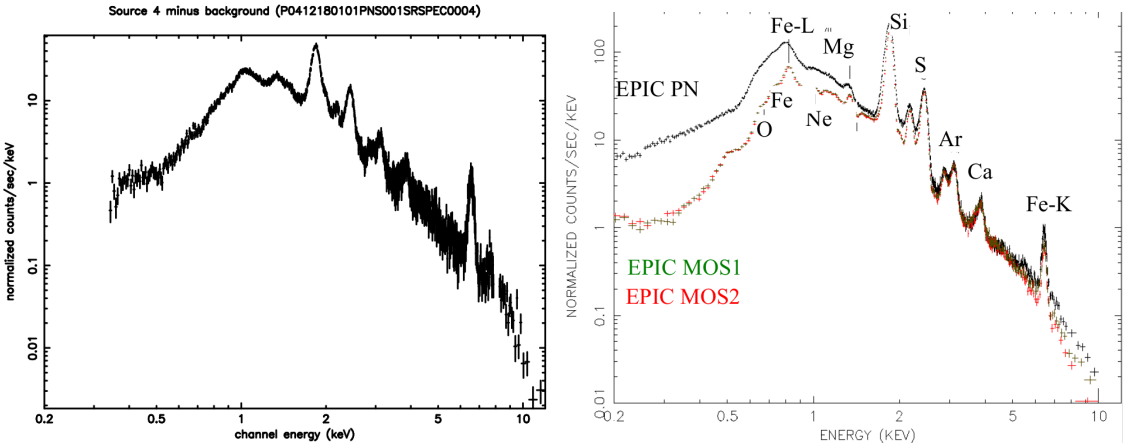


FIG. I.16 – Spectres EPIC de deux restes de supernovae observés par XMM-Newton. Gauche : Cassiopée A, SNR jeune de type Ib/c. Droite : Tycho, SNR de type Ia (figure extraite de Decourchelle *et al.* (2001)).

1.7 GALAXIES, AGN ET AMAS DE GALAXIES

1.7.1 Les galaxies

L'émission X que l'on reçoit de certaines galaxies a des origines multiples. On verra dans la section suivante le cas particulier des AGN dont l'émission provient en grande partie, voire en totalité, du noyau actif central qui comporte un trou noir supermassif de masse $M \sim 10^6 - 10^8 M_\odot$. On pense que pour la majorité des galaxies, surtout massives, celles-ci contiennent un tel trou noir central, mais que celui-ci peut ne pas être actif. Pour les galaxies dont on ne détecte pas d'activité du noyau, on voit essentiellement en X la contribution des binaires, des SNR jeunes et des zones de chocs dans les régions H II de formation d'étoiles photoionisées par des associations OB. Si la galaxie observée est suffisamment proche, ces différentes sources peuvent être résolues. Sinon, on observe une émission diffuse qui contient la somme de toutes ces contributions.

a Galaxies spirales non actives

On ne détecte en X que les galaxies spirales peu éloignées de nous. Pour les plus proches, ont distingue en X différentes régions comme les amas globulaires orbitant en dehors du plan, les éventuelles ULX ou XRB du plan et, toujours dans le plan, les zones de chocs dans le milieu interstellaire photoionisé par des associations OB, des SNR et des zones de formation d'étoiles. Certaines possèdent en leur centre des régions H II. On les appelle galaxies à noyau H II. Leur émission optique ressemble alors à celle des LINER Ho (2008) avec lesquelles elles peuvent être confondues.

b Galaxies elliptiques

Les galaxies elliptiques contiennent essentiellement des populations d'étoiles âgées groupées un peu à la façon d'immenses amas globulaires. Comme dans les amas globulaires, mais à plus grande échelle, on trouve de nombreuses XRB. Essentiellement des LMXB, les HMXB étant associées à des populations d'étoiles jeunes. La seule contribution des XRB permet parfois de détecter ces galaxies en X. Le spectre des galaxies elliptiques ne montre souvent aucune raie d'émission (voir Fig. I.17), si ce n'est parfois une légère raie H_α qui trahit une faible activité nucléaire (du noyau). Le rapport Fx/Fopt constaté avec XMM-Newton est inférieur à 0.01 (Georgakakis *et al.* 2006).

c Galaxies starburst *et* star forming

Les galaxies *starburst* sont des galaxies qui, pour différentes raisons comme par exemple une récente collision avec une autre galaxie, ont un taux de formation d'étoiles très élevé. Optique-

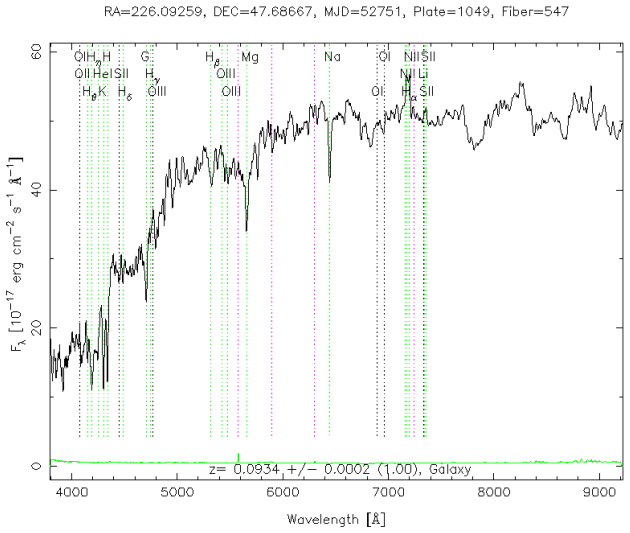


FIG. I.17 – Spectre SDSS de la galaxie elliptique 2MASX J15042218+4741116.

ment elle apparaissent très bleues et leur spectre et complètement dominé par quelques raies étroites caractéristiques de la photoionisation des régions H II par des étoiles jeunes de type OB (voir Fig. I.18 et I.19). Ces raies comportent notamment les raies de [O III]. Les raies de [N II], [S II] et [O I] sont plus faibles quand la photoionisation est d'origine stellaire que dans les LINER où elle provient des radiations plus dures émises par le noyau actif. On trouve de nombreuses HMXB et des étoiles de type Wolf-Rayet (He II) liées à la présence de populations jeunes d'étoiles et du gaz photoionisé par les étoiles jeunes. Pour séparer ces galaxies des LINER, on a recours au diagramme $\log_{10}([\text{O III}]\lambda 5007/\text{H}_\beta)$ versus $\log_{10}([\text{N II}]\lambda 6583/\text{H}_\alpha)$, voir Ho (2008), §2.3 et Stasińska *et al.* (2006), §6.2. La valeur de $\log_{10}([\text{N II}]\lambda 6583/\text{H}_\alpha)$ est typiquement inférieure à 0.4 pour les galaxies et supérieure pour les LINER. Enfin, le rapport Fx/Fopt constaté avec XMM-Newton est inférieur à 0.05.

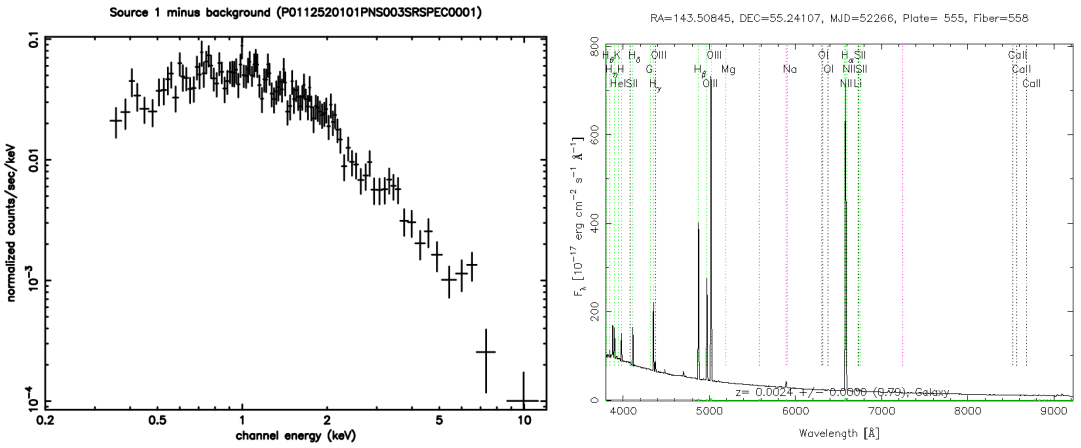


FIG. I.18 – Spectres EPIC et SDSS de la galaxie *starburst* Mrk 116A. Le lecteur trouvera des informations sur des spectres similaires dans Jenkins *et al.* (2004).

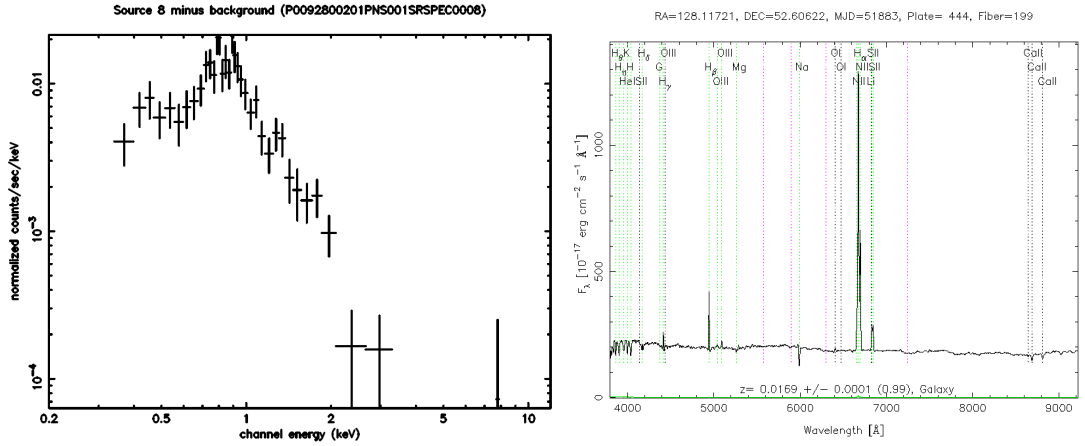


FIG. I.19 – Spectres EPIC et SDSS de la galaxie elliptique Mrk 91 ayant subi une formation stellaire récente.

1.7.2 Noyaux actifs de galaxies (AGN)

a Les QSO

Les QSO, pour *Quasi Stellar Objects*, sont des sources extragalactiques éloignées et ponctuelles (non-résolues). Il s'agit de galaxies actives dont la luminosité est complètement dominée par l'activité du noyau. La composante galactique n'est pas visible dans le spectre optique. La période de variabilité optique de ces objets, relativement courte (maximum 1 mois), a permis de poser des contraintes sur les dimensions de la région à l'origine de l'émission. Le processus le plus efficace pour transformer de la matière en énergie est l'accrétion. On en a déduit que l'émission des QSO provient d'un gigantesque disque d'accrétion autour des trous noirs supermassifs au centre de galaxies.
Le spectre optique de ces objets présente des raies très larges en H_α, Mg II, etc et s'apparente souvent, en X, à une loi de puissance d'indice de photon ayant une valeur proche de 1.9. La présence d'un *Big Blue Bump*, excès du continuum atteignant un maximum dans l'UV et souvent interprétée comme une émission thermique venant du disque d'accrétion, fait des QSO d'importants émetteurs UV (Elvis *et al.* 1994).

Les QSO dits *radio loud*, par opposition aux *radio quiet*, sont pourvus de jets focalisés. Si l'observateur se trouve dans le cône d'émission de ces jets, on parle de BL Lac, ou blazar.

L'émission X est interprétée comme ayant une composante thermique (corps noir) venant des régions les plus internes du disque avec Comptonisation et éventuellement une composante molle de réflexion sur le disque dûe au *reprocessing* des photons X durs.

Le rapport Fx/Fopt constaté avec XMM-Newton est typiquement compris entre et 0.3 et 10.

La majorité des quasars sont bleus, mais il existe également une population dont l'émission est rougie par la présence de poussières (Young *et al.* 2008). Ces objets émettent alors en infrarouge et ont un spectre X plus dur que ceux de la population bleue. D'autres objets sont rougis par l'effet du *redshift*. Enfin, il semblerait qu'il existe une population, plus faible, de quasars intrinsèquement rouges. Une des hypothèses, confortée par la présence de raies larges de Mg II dans certains spectres, serait liée à un taux d'accrétion plus faible que celui de la population bleue (Young *et al.* 2008).

b Les galaxies de Seyfert

Les galaxies de Seyfert sont des galaxies possédant également un noyau actif. A la différence des QSO, l'émission optique n'est pas complètement dominée par la région centrale mais le continu de la galaxie est visible. La classification de ces galaxies se fait sur les caractéristiques du spectre optique. Les objets présentant des raies d'émissions larges sont classés comme Seyfert de type I alors que les Seyfert de type II ne présentent que des raies étroites. Cette division est dûe à l'existence de deux régions distinctes dans les AGN : la BLR pour *Broad emission Line Region* et

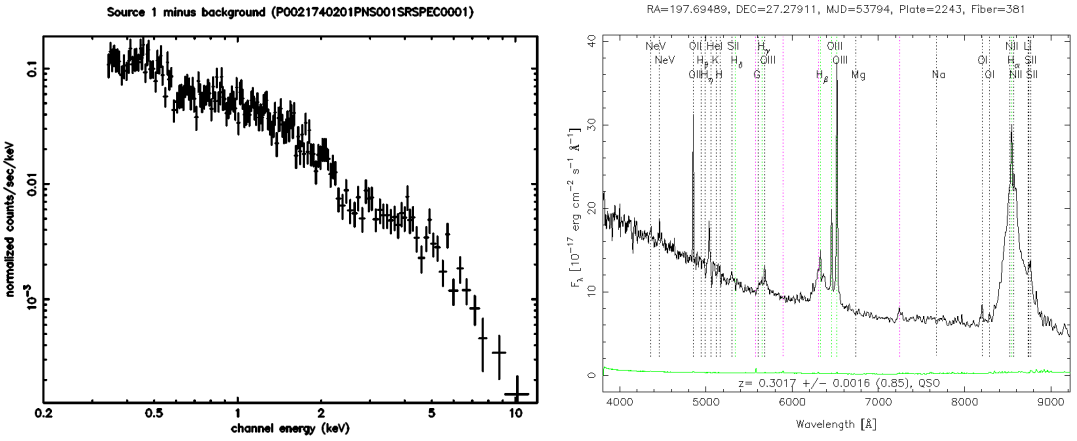


FIG. I.20 – Spectres EPIC de SDSS du QSO de coordonnées equatoriales J2000 13h11m04.57s +27°29″47.109′. Le spectre à la forme d'une *broken power law*.

la NLR pour *Narrow emission Line Region*. La première se situe près du disque d'accrétion et du trou noir où le mouvement rapide de la matière est à l'origine de l'élargissement par effet Dopler des raies d'excitations visibles en optique. La seconde se trouve plus éloignée dans la direction perpendiculaire au plan. Les nuages moléculaires sont photoionisés par l'émission UV provenant du disque. La matière, plus ténue, n'ayant pas une grande dispersion de vitesse dans cette région permet l'émission de raies optiques interdites et étroites.

Les Seyfert de type I sont divisées en plusieurs sous-classes en fonction de la valeur du rapport de raie H_β/O III (Véron-Cetty & Véron 2006). Mais cette séparation est quelque peu arbitraire. Tout aussi arbitraire la séparation entre QSO et Seyfert fixée par certains auteurs (Véron) en fonction de la magnitude optique absolue : QSO si $M_B > 23$, Seyfert sinon. Le plus probable est que les mécanismes en jeu dans les QSO et les galaxies de Seyfert sont les mêmes mais le taux d'accrétion du trou noir central est plus faible pour ces dernières.

c LLAGN : les LINERS et les TO

Les LINER, pour *Low Ionization Nuclear Emission-line Regions*, ont été définies à l'origine comme une sous-classe d'AGN de faible luminosité (LLAGN pour *Low-Luminosity AGNs*). Ils sont caractérisés à partir de leur spectre optique qui est dominé par des raies d'émission de largeur similaire à celles des *narrow-line regions* des galaxies de Seyfert mais avec une intensité lumineuse généralement plus faible. Pour les objets comportant les noyaux les plus actifs, les raies sont un peu plus larges (Heckman 1980). Elles correspondent à du gaz plus faiblement ionisé que pour les AGN classiques (González-Martín *et al.* 2006). Plus précisément, comme mentionné dans González-Martín *et al.* (2006), les LINER sont définies arbitrairement dans Heckman (1980) comme étant des galaxies ayant un rapport [O II]λ3727/[O III]λ5007 ≥ 1 et [O I]λ6300/[O III]λ5007 $\geq 1/3$.

Il semblerait qu'environ un quart des LINER émettent en UV et soient variables en UV. Cependant on suspecte que dans un grand nombre de cas, le noyau central est tellement obscurci/assombri par la matière qu'on ne puisse l'observer en UV, infrarouge ou optique. On recherche alors la trace de l'activité du trou noir central par la présence d'une contrepartie radio ou X.

Les spectres X se décomposent en deux grandes familles représentant respectivement 60% et 40% des cas environ :

- Les spectres X marqués par la présence d'un noyau actif : le spectre X comporte une loi de puissance entre 1 à 2 et 10 keV d'indice de photon valant 1.8 à 1.9 ; on voit parfois la raie K du fer (Fe K) à environ 6.4 KeV qui est typique, dans ce cas, de gaz relativiste près de l'horizon d'un trou noir accrétant ; éventuellement on voit une composante thermique entre 0.5 et 1 à 2 keV provenant d'un plasma optiquement mince (starburst).
- Les spectres X de type galaxies starburst : comportent l'emission thermique d'un plasma optiquement mince à une température $kT \sim 0.6$ keV et éventuellement une composante en loi de puissance provenant d'un noyau faiblement actif ou absorbé.

Les luminosités X sur la plage 2-10 keV varient entre 1.5e+38 et 1.5e+42 erg.s^{-1}

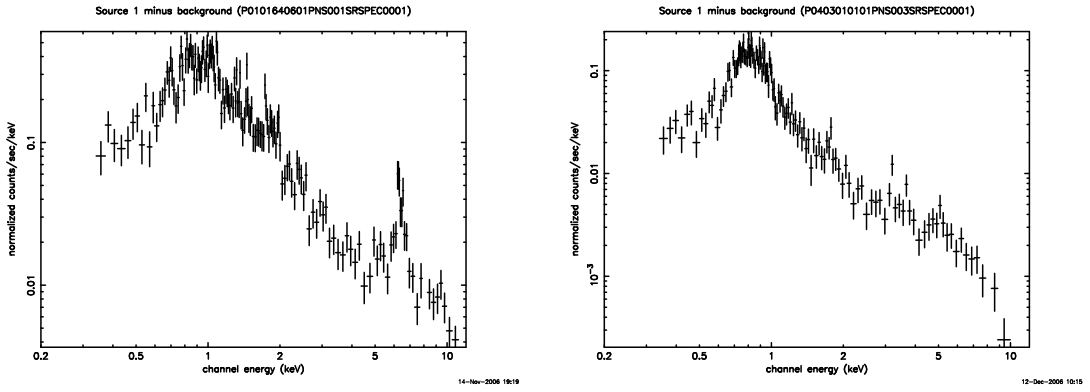


FIG. I.21 – Spectres EPIC de deux LINER observées par XMM-Newton. Gauche : NGC 6240. Droite : 3C 88. Le lecteur trouvera plus d'information sur le spectre de 3C 88 dans Gliozzi *et al.* (2008).

Il semblerait que les LINER ne soit pas des analogues des Seyfert avec un taux d'accrétion plus faible. L'environnement du trou noir semble différent. Pour plus d'informations sur le sujet, voir (Ho 2008, §8.3).

Les objets de transition, TO pour *Transition-type Objects*, sont des objets qui contiennent à la fois un noyau faiblement actif et un noyau H II.

1.7.3 Les Amas de galaxies

Les amas de galaxies sont des regroupements de plusieurs centaines de galaxies liées entre elles gravitationnellement. Se trouve parfois au centre de l'amas une galaxie elliptique géante (Dubinski 1998). On détecte en X l'émission diffuse du rayonnement par Bremstrahlung thermique du gaz chaud en équilibre hydrostatique dans le puits de potentiel de l'amas. Il se peut donc que le centre de cette émission diffuse n'ait pas de contrepartie optique.

1.8 LE FOND DIFFUS X

1.8.1 Le fond diffus X galactique

Un fond diffus X non résolu en provenance du disque galactique a été découvert il y a environ 25 ans (Worrall *et al.* 1982, Warwick *et al.* 1985). On l'appelle le GRXE pour *Galactic Ridge X-ray Emission*. Les caractéristiques spectrales de cette émission correspondent à un plasma optiquement mince de température de l'ordre de 10^8 K auquel s'ajoute une raie du fer bien visible à 6.7 keV. Il semble aujourd'hui qu'une grande partie de cette émission ($\sim 80\%$ entre 6 et 7 keV) provienne de sources ponctuelles de faible luminosité X comme les RS CVn et plus généralement les couronnes actives (Revnivtsev *et al.* 2009). Pour en savoir plus sur les GRXE, voir Revnivtsev *et al.* (2006, 2009).

1.8.2 Le fond diffus X extragalactique

Le fond diffus X, XRB pour *X-ray Background*, à été découvert il y a plus de 40 ans à partir des données enregistrées par trois compteurs Geiger installés dans une fusée Aerobee (Giacconi *et al.* 1962). Il a depuis été confirmé et on sait maintenant qu'il provient de la somme des contributions de sources ponctuelles extragalactiques de faible intensité. En effet, dans les relevés X actuels les plus profonds, la somme des flux X de sources non résolues détectées permet d'expliquer la totalité du fond diffus X pour des énergies de quelques keV et environ la moitié pour des énergies allant de 7 à 10 keV (Gilli *et al.* 2007). Les programmes de suivi optique des observations réalisées

en X montrent que l'immense majorité de ces sources ponctuelles sont des AGN qui sont plus ou moins obscurcis. Les modèles actuels cherchant à reproduire le XRB nécessitent l'ajout d'une population d'AGN extrêmement obscurcis, notamment pour expliquer les valeurs de ce fond à des énergies supérieures à 30 keV (Gilli *et al.* 2007). Le maximum du XRB se trouve à des énergies dépassant 10 keV et se trouve donc en grande partie au delà de la limite supérieure de détection d'XMM-Newton.

L'étude du XRB est intéressante à plusieurs titres. Elle nécessite la recherche de la population d'AGN extrêmement obscurcis qui est introduite dans les modèles mais n'a pas encore été observée. On s'interroge également sur un lien éventuel entre *redshift*, luminosité X et obscurcissement. Ce lien pourrait aider notre compréhension de la formation et de l'évolution des galaxies et servir en cosmologie. La population de quasars à haut redshift (z>3) et la fonction de masse des trous noirs au centre des galaxies sont également mal connus et pourrait profiter des études sur le XRB.

2 De XMM-Newton au catalogue 2XMMi

2.1 L'OBSERVATOIRE SPATIAL XMM-NEWTON

Le satellite XMM-Newton (X-ray Multi Mirror) est un satellite européen d'observation en rayons X appartenant à l'ESA (European Space Agency). Il a été mis en service après avoir été placé en orbite le 10 décembre 1999 par le premier lancement commercial d'une fusée Ariane-V. Avec ses 10 m de longueur, 16 m d'envergure, 4 m de diamètre et ses 3.8 tonnes, XMM-Newton est le plus gros satellite d'observation de rayons X jamais mis en orbite. Les financements permettent actuellement son utilisation jusqu'à au moins fin 2012.

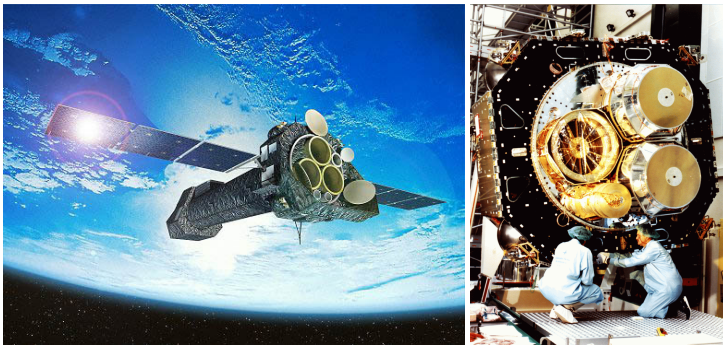


FIG. I.22 – Gauche : une vue d'artiste du satellite XMM-Newton (crédits ESA/Ducros). Droite : photo réalisé lors du montage alors que deux des trois « miroirs » ont été posés (crédits Dornier Satellitensysteme GmbH).

L'observatoire spatial XMM-Newton (Jansen *et al.* 2001), cf. Fig. I.22, est principalement constitué de trois télescopes de type Wolter (Aschenbach 1985, Trümper & Hasinger 2008) et observe en rayons X dans une gamme d'énergie allant de ∼0.2 à 12 keV. La résolution angulaire de chaque instrument est de l'ordre de 2″ à 5″ pour un champ de vue d'environ 30′ de diamètre. A chacun de ces trois télescopes est associé une caméra, cf. Fig. I.23. Deux d'entre elles, M1 et M2, sont basées sur la technologie MOS (*Metal-Oxyde-Semiconductor*), la troisième, la caméra PN, est basée sur une technologie CCD (*Charge Coupled Device*) classique à base de jonctions p-n. L'ensemble de ces trois caméras constitue l'instrument EPIC (*European Photon Imaging Camera*). La largeur à mi-hauteur[5] (FWHM pour *Full Width at Half Maximum*) de la PSF (*Point Spread Function*) de ces trois caméras varie avec la gamme d'énergie, mais vaut environ 6″ (Jansen *et al.* 2001).

Les deux télescopes équipés de caméra MOS sont également équipés de spectromètres à réseau scindant les faisceaux incidents en deux. Pour chacun de ces deux télescopes, la moitié du faisceau

[5] Pour une gaussienne de variance σ^2, FWHM $= 2\sqrt{2\log 2}\sigma$

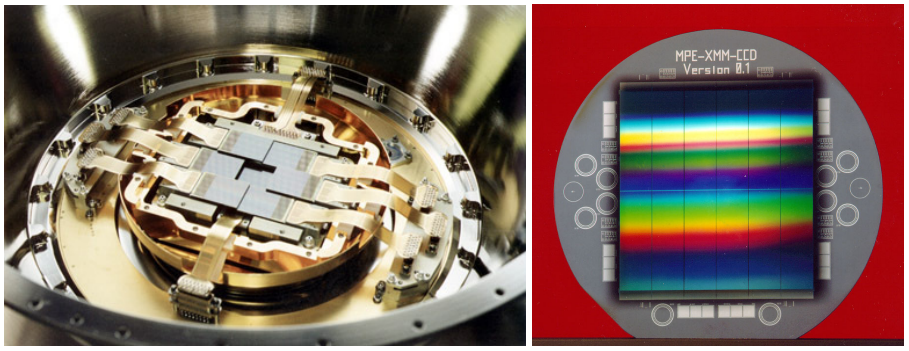


FIG. I.23 – Gauche : une des deux caméras EPIC MOS constituée de 7 capteurs. Droite : la caméra EPIC PN qui contient 12 capteurs CCD. (crédits ESA/MPE).

incident est ainsi redirigé vers une caméra MOS et l'autre moitié vers une caméra permettant de visualiser les spectres. L'ensemble des deux spectromètres, R1 et R2 associés respectivement à M1 et M2, constitue l'instrument RGS pour *Reflection Grating Spectrometer* (den Herder *et al.* 2001).
XMM-Newton dispose également d'un moniteur optique OM (pour *Optical Monitor*). Cet instrument (Mason *et al.* 2001) est constitué d'un télescope de type Ritchney Chrétien de 30 cm de diamètre observant dans le visible et l'UV (Ultra Violet). Son champ de vue est un carré de 17′ de coté. En raison de difficultés techniques dues à des effets thermiques et de réflexion sur les parois de l'OM, les données EPIC et OM sont indépendantes. On ne dispose donc pas à l'heure actuelle d'associations source X EPIC – source optique/UV OM.

Parmi les évènements marquants de la vie d'XMM-Newton, on peut noter que lors de l'assemblage du satellite, une des 182 grilles de diffraction d'un des RGA (*Reflection Grating Array*) n'a pu être installée. Depuis son lancement, au moins 3 micrométéorites (Meidinger *et al.* 2003) sont entrées dans les mirroirs du satellite et ont heurté les caméras MOS1, PN et MOS2 respectivement pendant les révolutions #107, #156 et #325. La micrométéorite ayant percuté l'instrument MOS1 est à l'origine de 27 pixels défectueux et plus de 30 pour l'évenement lié à PN (Strüder *et al.* 2001). Ces incidents n'ont pas de répercution notable sur l'exploitation scientifique. En revanche, une autre micrométéorite qui a heurté la caméra MOS1 le 9 mars 2005 pendant la révolutions #961 est à l'origine de la perte du capteur CCD 6[6]. Enfin, la communication avec le satellite a été perdue le 18 octobre 2008[7] suite à la défaillance d'un capteur de fin de course de l'antenne principale. Elle a heureusement pu être rétablie quatre jours plus tard[8].

2.2 DU SATELLITE AU CATALOGUE 2XMMI

L'ESA délègue au SSC (Science Survey Center) le traitement systématique des données issues de XMM ainsi que la construction d'un catalogue contenant les sources X détectées. Le SSC est un consortium impliquant des scientifiques de différents laboratoires européens. L'observatoire de Strasbourg, et plus particulièrement une partie de l'équipe Hautes Énergies, fait partie de ce consortium. Une des tâches incombant au SSC est la création d'un catalogue de sources issues des observations réalisées par XMM-Newton. A ce jour plusieurs catalogues ont été rendus public, chacun incluant le précédent. Dans l'ordre chronologique : le 1XMM, le 2XMMp, le 2XMM et le 2XMMi. Voyons succinctement les étapes de la création de ces catalogues.

Le centre de contrôle de XMM-Newton se trouve à l'ESOC (European Spatial Operation Center) à Darmstadt en Allemagne. Il communique avec le satellite via plusieurs centres de relais situés en Guyane Française (Kourou), au Chili (Santiago) et en Australie (Perth). Le centre de Darmstadt transfère les données d'observations brutes à VILSPA, compression anglosaxone de Villafranca Spain, du nom de la ville espagnole proche de Madrid ou se trouve le VSTS (Villa-

[6] http://xmm.esac.esa.int/external/xmm_news/items/MOS1-CCD6/

[7] http://www.esa.int/esaSC/SEM7A3RTKMF_index_0.html

[8] http://www.esa.int/esaSC/SEM268RTKMF_index_0.html

franca Satellite Tracking Station). Les données sont alors archivées dans une base sous forme de fichiers ODF (Observational Data File).

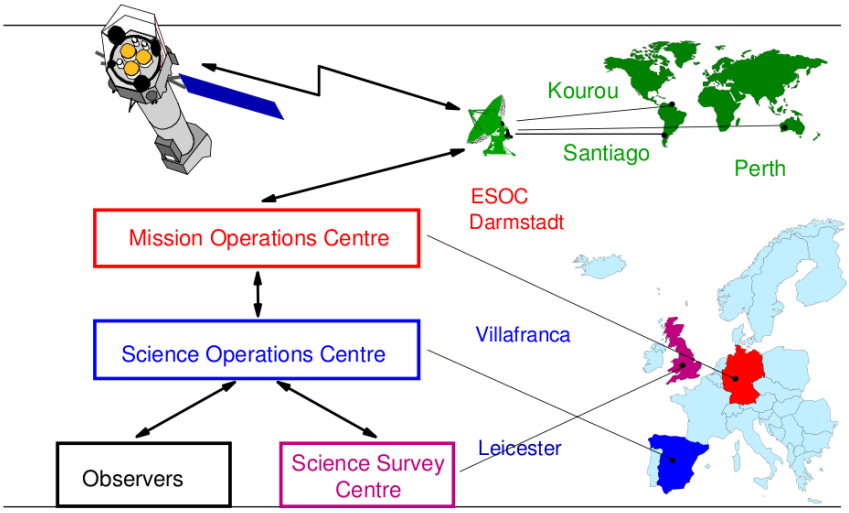


FIG. I.24 – Vue synthétique du segment sol ("Ground Segment") de XMM-Newton (crédits Dietmar Heger).

Les cibles observées par XMM-Newton sont sélectionnées par un système de demande de temps ouvert à la communauté internationale. La pression est de l'ordre d'une demande acceptée pour trois soumises, et la pression en temps est de l'ordre 1 pour 9. Les observateurs invités disposent des produits PPS[9], mais peuvent aussi appliquer eux même les traitements sur les données brutes. Ils sont pour cela aidés par le SAS (*Science Analysis Software*), logiciel de traitement de données développé spécifiquement par le consortium pour XMM-Newton.

Outre la source centrale, objet de la demande, une observation contient d'autres informations intéressantes. Aussi les fichiers ODF sont transmis au SSC (Survey Science Center) à l'université de Leicester au Royaume-Uni pour une analyse systématique de l'environnement de l'objet central. Là, les données brutes vont être réduites par un pipeline de traitement de données qui va permettre d'extraire des listes de sources, des images et des spectres corrigés des effets instrumentaux. Ces données sont ensuite envoyées à l'observatoire de Strasbourg ou les listes de sources X vont être corrélées avec des sources issues des catalogues d'archives stockés au CDS (Centre de Données de Strasbourg) et dans NED (Nasa Extragalactic Database). Des produits – tel que des cartes de pointages, des produits graphiques, des cartes couleurs des zones couvertes par le SDSS,... – sont également créés par l'ACDS (*Astronomical Catalogue Data Subsystem*). La cross-corrélation effectuée par l'ACDS recense, pour chaque source X, les sources d'archives compatibles avec une complétude de 99.7%. L'ensemble des produits générés par l'ACDS sont retournés à VILSPA pour être stocké dans la XSA (*XMM Science Archive*).

La création du catalogue de sources X comporte une étape de *data screening* : la sortie du pipeline est systématiquement vérifiée à l'œil, à l'aide d'un logiciel, pour être validée.

L'observatoire de Strasbourg met à disposition sur Internet une interface avec la XCATDB[10] (Motch *et al.* 2007), base de données hébergeant les sources XMM et les produits d'archives. Cette base présente la particularité, et l'avantage, de pouvoir exploiter facilement toute la richesse des produits issus de l'ACDS, notamment les corrélations avec les catalogues d'archives. Ce travail de thèse a permis, dans une certaine mesure, d'enrichir la XCATDB. Il a en effet apporté des informations supplémentaires aux relations entre sources X et sources d'archives issues des grands relevés (*survey*). Parmi ces informations figurent les probabilités d'identification et les densités locales. La XCATB est basée sur SAADA (Michel *et al.* 2005).

[9] Les produits PPS sont les produits issus du *Pipeline Processing System*

[10] `http://amwdb.u-strasbg.fr/2xmmi`

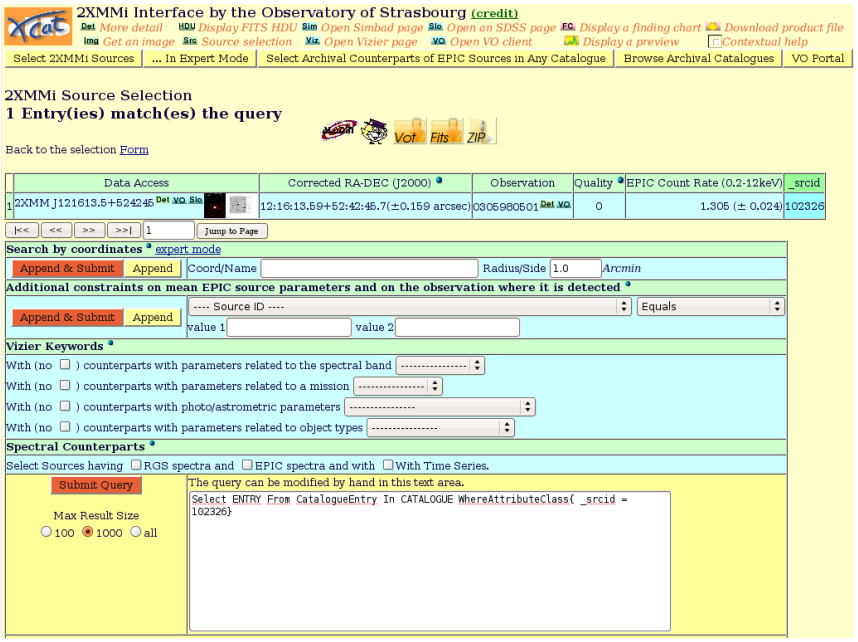


FIG. I.25 – Interface simplifiée de la XCAT-DB pour l'exploitation du 2XMMi.

2.3 LE 2XMMI EN CHIFFRES

2.3.1 Statistiques globales

Le 2XMMi (*incremental Second XMM-Newton Serendipitous Source Catalogue*) est une version étendue du 2XMM (Watson *et al.* 2008). Le 2XMMi a été construit à partir de 4117 observations et contient 289 083 détections hétérogènes pour un total de 221 012 sources X uniques. Parmi les détections, 24 533 sont des sources étendues (ep_ext_ml>4) pour un total de 21 970 sources uniques étendues.
Le catalogue couvre une surface représentant $\sim 1\%$ du ciel sur une large gamme de latitudes et longitudes galactiques.
Les caméras de l'instrument EPIC couvrent un champ de vue d'environ $\sim 30'$ de diamètre et sont sensibles à des énergies allant de 0.2 à 12 keV. La précision typique des positions est de l'ordre de $\sim 2''$, voir Fig. I.26. Le flux médian des détections dans la bande d'énergie totale (0.2 - 12 keV) vaut $\sim 2.5\text{x}10^{-14}$ erg.cm^{-2}.s^{-1}. Elle vaut respectivement $\sim 5.8\text{x}10^{-15}$ et $\sim 1.4\text{x}10^{-14}$ erg.cm^{-2}.s^{-1} (voir Fig. I.27) dans les bandes « soft » (0.2 - 2 keV) et « hard » (2 - 12 keV). Environ 20% des sources ont un flux total en dessous de $1.0\text{x}10^{-14}$ erg.cm^{-2}.s^{-1}.

Pour plus d'information concernant le 2XMMi, nous renvoyons le lecteur au guide de l'utilisateur en ligne à l'adresse suivante :
`http://xmmssc-www.star.le.ac.uk/Catalogue/2XMM/UserGuide_xmmcat.html`

2.3.2 Les grandeurs physiques

Plus de 260 paramètres – grandeurs physiques, marques, noms – sont associés à une source X. Dans le 2XMM, la bande totale d'énergie 0.2 - 12 keV est divisée en cinq sous bandes, dont les valeurs sont données dans le tableau I.1.

Le 2XMMi contient 4 rapports de dureté, ou HR pour *Hardness Ratio* en anglais. Un rapport de dureté caractérise la dureté d'un spectre X. En effet, comme dit précédemment (§1.1.5) plus une source émet de photons « durs », de haute énergie, par rapport aux photons « mous », d'énergie moindre, plus son spectre est dit « dur ». Le HR est donc naturellement décrit par la formule Eq. I.9 comme étant l'excès (ou le déficit) de photons durs reçus par unité de temps par rapport au

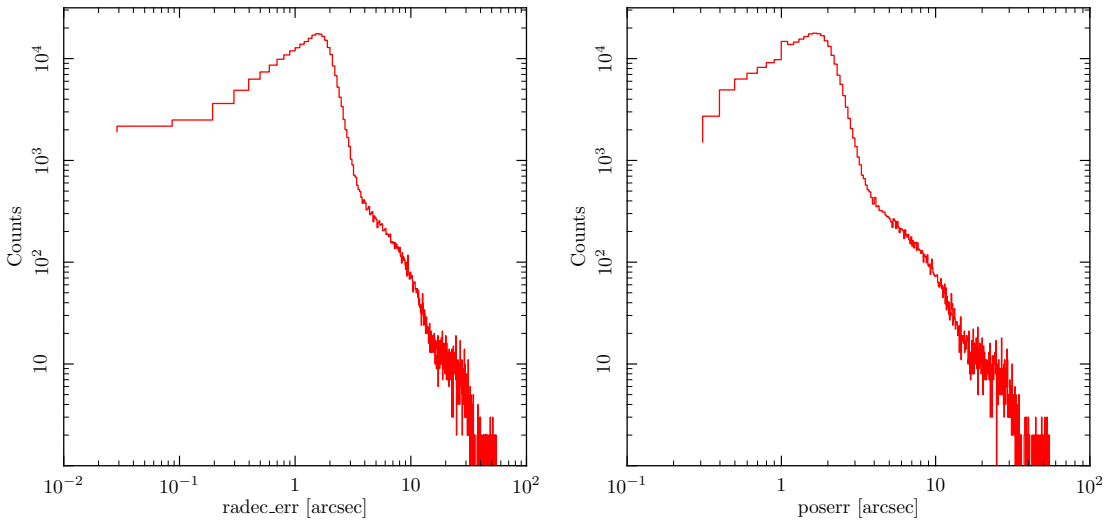


FIG. I.26 – Erreurs de position des sources ponctuelles (sc_ext_ml<4) du 2XMMi sans (gauche) et avec erreur systématique (droite).

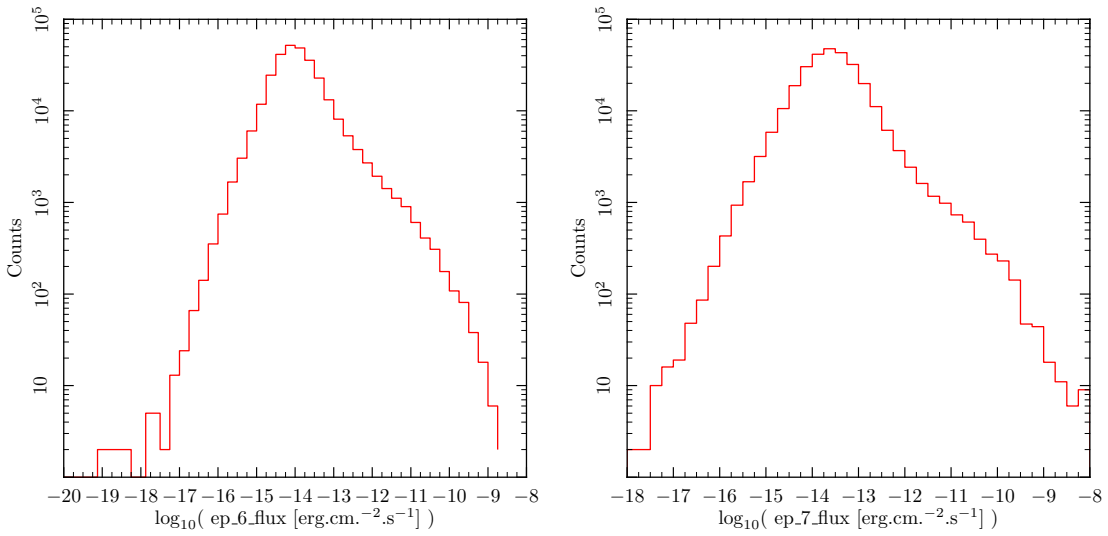


FIG. I.27 – Flux dans les bandes « soft » (gauche) et « hard » (droite) des sources du 2XMMi.

nombre total de photons « durs » et « mous » reçus par unité de temps.

$$HR = \frac{nbr_photon_dur.s^{-1} - nbr_photon_mou.s^{-1}}{nbr_photon_dur.s^{-1} + nbr_photon_mou.s^{-1}} \tag{I.9}$$

Le nombre de photons reçus par seconde dans une bande d'énergie donnée est également appelé *count rate*. Le tableau I.2 présente les bandes d'énergie du 2XMMi, voir tableau Tab. I.1, mises en jeu dans les différents rapports de dureté. Les valeurs de HR sont comprises entre -1 et 1. Le rapport de dureté HR1 est sensible à l'absorption galactique. Il en est de même, dans une moindre mesure, pour HR2.

La caméra PN possède deux fois plus de surface effective que M1 et M2. Comme mentionné dans la section 2.1, ces deux instruments ne bénéficient que de la moitié du faisceau incident, l'autre moitié servant à des analyses spectrales à haute résolution. La caméra PN possède également une meilleure résolution temporelle comparativement à M1 et M2. De fait, sauf mention

TAB. I.1 – Tableau présentant les différentes bandes d'énergies du 2XMMi

Bande	Gamme d'énergie	Equivalents en λ	Désignation
1	0.2 - 0.5 keV	6.2 - 2.48 nm	
2	0.5 - 1 keV	2.48 - 1.24 nm	
3	1 - 2 keV	1.24 - 0.62 nm	
4	2 - 4.5 keV	0.62 - 0.275 nm	
5	4.5 - 12 keV	0.275 - 0.103 nm	
6	0.2 - 2 keV	6.2 - 0.62 nm	soft band
7	2 - 12 keV	0.62 - 0.103 nm	hard band
8	0.2 - 12 keV	6.2 - 0.103 nm	total band
9	0.5 - 4.5 keV	2.48 - 0.275 nm	XID band

TAB. I.2 – Tableau présentant les différents rapports de dureté du 2XMM

Rapport de dureté	Bande d'énergie dur-molle
HR1	2 - 1
HR2	3 - 2
HR3	4 - 3
HR4	5 - 4

contraire, les flux X et les rapports de dureté utilisés sont ceux associés à la caméra PN.

Parmi les grandeurs souvent utilisées dans ce travail, on peut également mentionner EP_EXT_ML (pour *EPIC EXTent Maximum Likehood*). Ce paramètre caractérise la probabilité pour qu'une source détectée par la caméra EPIC soit étendue. Le critère habituel pour considérer qu'une source est ponctuelle consiste à prendre EP_EXT_ML < 4.

Comme nous verrons dans la suite, on a pris en compte les erreurs associées à ces grandeurs.

2.3.3 Les filtres

Pour chacune des trois caméras, il existe 4 filtres : *Medium* (médium), *Thick* (épais), *Thin1* (mince 1), *Thin2* (mince 2). Ces filtres ont pour but de bloquer une partie de l'émission optique, IR et UV. En effet les détecteurs sont sensibles à ces longueurs d'onde qui peuvent donc contaminer les valeurs mesurées en X. Cependant, l'utilisation de ces filtres à pour effet le blocage d'une partie de l'émission X, notamment molle. Il faut donc adapter le filtre utilisé en fonction de l'objet observé pour avoir le meilleur compromis entre suppression du bruit et diminution de la surface effective en X mou. Plus le filtre est épais, moins la surface effective en X mou est importante, mais plus la contamination par les autres longueurs d'onde est faible. Plus la source émet en optique, plus le filtre choisi devra être épais. Pour plus d'information, voir la documentation (*Users' Handbook* §3.3.6 « *EPIC filters and effective area* ») sur le web[11]. Le tableau I.3 fournit le nombre de détections associées à chaque filtre.

2.3.4 Correction des positions

Nous verrons dans le chapitre suivant que les positions des sources XMM peuvent souffrir d'un biais composé d'une translation et d'une rotation. Ce biais est dû aux erreurs de positionnement du centre de l'observation et des axes Nord–Sud et Est–Ouest. Afin de recalibrer les positions des sources X, le pipeline de traitement de données d'XMM-Newton contient une routine, nommée *eposcorr*, qui les corrèle avec le catalogue USNOB1. Un critère statistique basé sur une fonction de vraisemblance décide de valider ou de rejeter la meilleure solution trouvée par un système de grilles adaptatives faisant subir des rotations et des translations aux sources X. Quand

[11] http://xmm2.esac.esa.int/external/xmm_user_support/documentation/uhb/node32.html

TAB. I.3 – Nombre de détections associées à chaque filtre

Filtre\Instrument	PN	MOS1	MOS 2
Thin1	112 478	111 171	110 666
Thin2	200	225	91
Medium	109 794	145 429	145 356
Thick	21 353	28 602	29 130
UNDEF	45 258	3 656	3 840

l'algorithme valide la solution, une erreur systématique de 0.35 seconde d'arc doit être ajoutée quadratiquement à l'erreur dûe à la PSF (*radec_err*) pour obtenir l'erreur sur la position d'une source. Quand l'algorithme rejette la solution, cette erreur systématique est fixée à 1.0 seconde d'arc.
Le catalogue 2XMMi contient 210 069 détections dont les positions ont pu être recalibrées par *eposcorr* et 79 014 pour lesquelles *eposcorr* a échoué.



Références

ASCHENBACH, B., 1985 ; *X-ray telescopes*. Reports on Progress in Physics, **48** (5), 579. URL `http://stacks.iop.org/0034-4885/48/579`

BECKER, W., 2008 ; *Neutron Stars and Pulsars*. Springer-Verlag Berlin and Heidelberg GmbH & Co. K, Berlin ; New York

BECKER, W., HABERL, F. & TRÜMPER, J., 2008 ; *The Universe in X-Rays*. Springer, Berlin ; New York

BEUERMANN, K., 2008 ; *The Universe in X-Rays*. Springer, Berlin ; New York

BOHLIN, R. C., COLINA, L. & FINLEY, D. S., 1995 ; *White Dwarf Standard Stars : G191-B2B, GD 71, GD 153, HZ 43.* AJ, **110**, 1316

BOUVIER, J., FORESTINI, M. & ALLAIN, S., 1997 ; *The angular momentum evolution of low-mass stars.* A&A, **326**, 1023

CASSAM-CHENAÏ, G., DECOURCHELLE, A., BALLET, J., HWANG, U., HUGHES, J. P., PETRE, R. & ET AL., 2004 ; *XMM-Newton observation of Kepler's supernova remnant*. A&A, **414**, 545

DECOURCHELLE, A., SAUVAGEOT, J. L., AUDARD, M., ASCHENBACH, B., SEMBAY, S., ROTHENFLUG, R., BALLET, J., STADLBAUER, T. & WEST, R. G., 2001 ; *XMM-Newton observation of the Tycho supernova remnant*. A&A, **365**, L218

DEN HERDER, J. W., BRINKMAN, A. C., KAHN, S. M., BRANDUARDI-RAYMONT, G., THOMSEN, K., AARTS, H., AUDARD, M., BIXLER, J. V., DEN BOGGENDE, A. J., COTTAM, J., DECKER, T., DUBBELDAM, L., ERD, C., GOULOOZE, H., GÜDEL, M., GUTTRIDGE, P., HAILEY, C. J., JANABI, K. A., KAASTRA, J. S., DE KORTE, P. A. J., VAN LEEUWEN, B. J., MAUCHE, C., MCCALDEN, A. J., MEWE, R., NABER, A., PAERELS, F. B., PETERSON, J. R., RASMUSSEN, A. P., REES, K., SAKELLIOU, I., SAKO, M., SPODEK, J., STERN, M., TAMURA, T., TANDY, J., DE VRIES, C. P., WELCH, S. & ZEHNDER, A., 2001 ; *The Reflection Grating Spectrometer on board XMM-Newton*. A&A, **365**, L7

DENNERL, K., 2008 ; *The Universe in X-Rays*. Springer, Berlin ; New York

DONATI, J. & LANDSTREET, J., 2009 ; *Magnetic fields of non-degenerate stars*. ArXiv e-prints

DOWNES, R. A., WEBBINK, R. F., SHARA, M. M., RITTER, H., KOLB, U. & DUERBECK, H. W., 2001 ; *A Catalog and Atlas of Cataclysmic Variables : The Living Edition*. PASP, **113**, 764

DUBINSKI, J., 1998 ; *The Origin of the Brightest Cluster Galaxies*. ApJ, **502**, 141

EISENSTEIN, D. J., LIEBERT, J., HARRIS, H. C., KLEINMAN, S. J., NITTA, A., SILVESTRI, N., ANDERSON, S. A., BARENTINE, J. C., BREWINGTON, H. J., BRINKMANN, J., HARVANEK, M., KRZESIŃSKI, J., NEILSEN, JR., E. H., LONG, D., SCHNEIDER, D. P. & SNEDDEN, S. A., 2006 ; *A Catalog of Spectroscopically Confirmed White Dwarfs from the Sloan Digital Sky Survey Data Release 4*. ApJS, **167**, 40

ELVIS, M., WILKES, B. J., MCDOWELL, J. C., GREEN, R. F., BECHTOLD, J., WILLNER, S. P., OEY, M. S., POLOMSKI, E. & CUTRI, R., 1994 ; *Atlas of quasar energy distributions*. ApJS, **95**, 1

FABBIANO, G., 2004 ; *Ultraluminous X-ray Sources : an Observational Review*. Dans *Revista Mexicana de Astronomia y Astrofisica Conference Series*, G. Tovmassian & E. Sion, réds., tm. 20 de *Revista Mexicana de Astronomia y Astrofisica Conference Series*

FEIGELSON, E. D. & MONTMERLE, T., 1999 ; *High-Energy Processes in Young Stellar Objects*. ARA&A, **37**, 363

GEORGAKAKIS, A. E., CHAVUSHYAN, V., PLIONIS, M., GEORGANTOPOULOS, I., KOULOURIDIS, E., LEONIDAKI, I. & MERCADO, A., 2006 ; *Mining for normal galaxies in the first XMM-Newton Serendipitous Source Catalog*. MNRAS, **367**, 1017

GIACCONI, R., GURSKY, H., PAOLINI, F. R. & ROSSI, B. B., 1962 ; *Evidence for x Rays From Sources Outside the Solar System*. Physical Review Letters, **9**, 439

GILLI, R., COMASTRI, A. & HASINGER, G., 2007 ; *The synthesis of the cosmic X-ray background in the Chandra and XMM-Newton era*. A&A, **463**, 79

GLIOZZI, M., FOSCHINI, L., SAMBRUNA, R. M. & TAVECCHIO, F., 2008 ; *The polyhedral nature of LINERs : an XMM-Newton view of LINERs in radio galaxies*. A&A, **478**, 723

GONZÁLEZ-MARTÍN, O., MASEGOSA, J., MÁRQUEZ, I., GUERRERO, M. A. & DULTZIN-HACYAN, D., 2006 ; *X-ray nature of the LINER nuclear sources*. A&A, **460**, 45

GÜDEL, M., 2004 ; *X-ray astronomy of stellar coronae*. A&A Rev., **12**, 71

GÜDEL, M. & TELLESCHI, A., 2007 ; *The X-ray soft excess in classical T Tauri stars*. A&A, **474**, L25

HECKMAN, T. M., 1980 ; *An optical and radio survey of the nuclei of bright galaxies - Activity in normal galactic nuclei*. A&A, **87**, 152

HEYVAERTS, J., 2006 ; *Astrophysique : étoiles, univers et relativité*. Dunod, Paris

HO, L. C., 2008 ; *Nuclear Activity in Nearby Galaxies*. ARA&A, **46**, 475

HÜGELMEYER, S. D., DREIZLER, S., WERNER, K., KRZESINSKI, J., NITTA, A. & KLEINMAN, S. J., 2007 ; *Observational Constraints on the Evolutionary Connection between PG 1159 Stars and DO White Dwarfs*. Dans *15th European Workshop on White Dwarfs*, R. Napiwotzki & M. R. Burleigh, réds., tm. 372 de *Astronomical Society of the Pacific Conference Series*

JANSEN, F., LUMB, D., ALTIERI, B., CLAVEL, J., EHLE, M., ERD, C., GABRIEL, C., GUAINAZZI, M., GONDOIN, P., MUCH, R., MUNOZ, R., SANTOS, M., SCHARTEL, N., TEXIER, D. & VACANTI, G., 2001 ; *XMM-Newton observatory. I. The spacecraft and operations*. A&A, **365**, L1

JENKINS, L. P., ROBERTS, T. P., WARD, M. J. & ZEZAS, A., 2004 ; *XMM-Newton observations of the starburst merger galaxies NGC 3256 and NGC 3310*. MNRAS, **352**, 1335

KAHABKA, P. & VAN DEN HEUVEL, E. P. J., 1997 ; *Luminous Supersoft X-Ray Sources*. ARA&A, **35**, 69

KLEINMAN, S. J., HARRIS, H. C., EISENSTEIN, D. J., LIEBERT, J., NITTA, A., KRZESIŃSKI, J., MUNN, J. A., DAHN, C. C., HAWLEY, S. L., PIER, J. R., SCHMIDT, G., SILVESTRI, N. M., SMITH, J. A., SZKODY, P., STRAUSS, M. A., KNAPP, G. R., COLLINGE, M. J., MUKADAM, A. S., KOESTER, D., UOMOTO, A., SCHLEGEL, D. J., ANDERSON, S. F., BRINKMANN, J., LAMB, D. Q., SCHNEIDER, D. P. & YORK, D. G., 2004 ; *A Catalog of Spectroscopically Identified White Dwarf Stars in the First Data Release of the Sloan Digital Sky Survey*. ApJ, **607**, 426

LANZ, T., TELIS, G. A., AUDARD, M., PAERELS, F., RASMUSSEN, A. P. & HUBENY, I., 2005 ; *Non-LTE Model Atmosphere Analysis of the Large Magellanic Cloud Supersoft X-Ray Source CAL 83*. ApJ, **619**, 517

LEWIN, W. H. G. & VAN DER KLIS, M., 2006 ; *Compact stellar X-ray sources*

LIU, J.-F. & BREGMAN, J. N., 2005 ; *Ultraluminous X-Ray Sources in Nearby Galaxies from ROSAT High Resolution Imager Observations I. Data Analysis*. ApJS, **157**, 59

LIU, Q. Z., VAN PARADIJS, J. & VAN DEN HEUVEL, E. P. J., 2001 ; *A catalogue of low-mass X-ray binaries*. A&A, **368**, 1021

LUCY, L. B. & WHITE, R. L., 1980 ; *X-ray emission from the winds of hot stars*. ApJ, **241**, 300

MAKISHIMA, K., KUBOTA, A., MIZUNO, T., OHNISHI, T., TASHIRO, M., ARUGA, Y., ASAI, K., DOTANI, T., MITSUDA, K., UEDA, Y., UNO, S., YAMAOKA, K., EBISAWA, K., KOHMURA, Y. & OKADA, K., 2000 ; *The Nature of Ultraluminous Compact X-Ray Sources in Nearby Spiral Galaxies*. ApJ, **535**, 632

MASON, K. O., BREEVELD, A., MUCH, R., CARTER, M., CORDOVA, F. A., CROPPER, M. S., FORDHAM, J., HUCKLE, H., HO, C., KAWAKAMI, H., KENNEA, J., KENNEDY, T., MITTAZ, J., PANDEL, D., PRIEDHORSKY, W. C., SASSEEN, T., SHIREY, R., SMITH, P. & VREUX, J.-M., 2001 ; *The XMM-Newton optical/UV monitor telescope*. A&A, **365**, L36

MEIDINGER, N., ASCHENBACH, B., BRAEUNINGER, H. W., DROLSHAGEN, G., ENGLHAUSER, J., HARTMANN, R., HARTNER, G. D., SRAMA, R., STRUEDER, L., STUEBIG, M. & TRUEMPER, J. E., 2003 ; *Experimental verification of a micrometeoroid damage in the pn-CCD camera system aboard XMM-Newton*. Dans *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, J. E. Truemper & H. D. Tananbaum, réds., tm. 4851 de *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*

MICHEL, L., NGUYEN, H. N. & MOTCH, C., 2005 ; *SAADA : Astronomical Databases Made Easier*. Dans *Astronomical Data Analysis Software and Systems XIV*, P. Shopbell, M. Britton & R. Ebert, réds., tm. 347 de *Astronomical Society of the Pacific Conference Series*

MITSUDA, K., INOUE, H., KOYAMA, K., MAKISHIMA, K., MATSUOKA, M., OGAWARA, Y., SUZUKI, K., TANAKA, Y., SHIBAZAKI, N. & HIRANO, T., 1984 ; *Energy spectra of low-mass binary X-ray sources observed from TENMA*. PASJ, **36**, 741

MOTCH, C., MICHEL, L. & PINEAU, F. X., 2007 ; *The XCATDB : a Complex Database Based on Saada*. Dans *Astronomical Data Analysis Software and Systems XVI*, R. A. Shaw, F. Hill & D. J. Bell, réds., tm. 376 de *Astronomical Society of the Pacific Conference Series*

MOTCH, C., PIRES, A. M., HABERL, F., SCHWOPE, A. & ZAVLIN, V. E., 2009 ; *Proper motions of thermally emitting isolated neutron stars measured with Chandra*. A&A, **497**, 423

MOTCH, C., WERNER, K. & PAKULL, M. W., 1993 ; *A new PG 1159 star discovered in the ROSAT XRT all sky survey - NLTE analysis of X-ray and optical spectra*. A&A, **268**, 561

MOTCH, C., ZAVLIN, V. E. & HABERL, F., 2003 ; *The proper motion and energy distribution of the isolated neutron star <ASTROBJ>RX J0720.4-3125</ASTROBJ>*. A&A, **408**, 323

NARAYAN, R., MCCLINTOCK, J. E. & YI, I., 1996 ; *A New Model for Black Hole Soft X-Ray Transients in Quiescence*. ApJ, **457**, 821

O'DWYER, I. J., CHU, Y.-H., GRUENDL, R. A., GUERRERO, M. A. & WEBBINK, R. F., 2003 ; *Hard X-Ray Emission Associated with White Dwarfs*. AJ, **125**, 2239

PAERELS, F. B. S. & KAHN, S. M., 2003 ; *High-Resolution X-Ray Spectroscopy with CHANDRA and XMM-NEWTON*. ARA&A, **41**, 291

PAGANO, I., AYRES, T. R., LANZAFAME, A. C., LINSKY, J. L., MONTESINOS, B. & MARCELLO-RODONÒ, 2006 ; *Key problems in cool-star astrophysics*. Ap&SS, **303**, 17

PATERNO, L., 1998 ; *Do we understand the 22-year solar activity cycle ?* Academie des Science Paris Comptes Rendus Serie B Sciences Physiques, **326**, 393

PIRES, A. M., MOTCH, C., TUROLLA, R., TREVES, A. & POPOV, S. B., 2009 ; *The isolated neutron star candidate 2XMM J104608.7-594306*. A&A, **498**, 233

PODSIADLOWSKI, P., RAPPAPORT, S. & PFAHL, E. D., 2002 ; *Evolutionary Sequences for Low- and Intermediate-Mass X-Ray Binaries*. ApJ, **565**, 1107

POPOV, S. B., COLPI, M., PROKHOROV, M. E., TREVES, A. & TUROLLA, R., 2003 ; *Young isolated neutron stars from the Gould Belt*. A&A, **406**, 111

PRAVDO, S. H., TSUBOI, Y., SUZUKI, Y., THOMPSON, T. J. & REBULL, L., 2009 ; *GGD 27 : X-rays from a Massive Protostar with an Outflow*. ApJ, **690**, 850

RAMSAY, G., CÓRDOVA, F., COTTAM, J., MASON, K., MUCH, R., OSBORNE, J., PANDEL, D., POOLE, T. & WHEATLEY, P., 2001 ; *First XMM-Newton observations of a Cataclysmic Variable II : The X-ray spectrum of OY Car*. A&A, **365**, L294

REMILLARD, R. A. & MCCLINTOCK, J. E., 2006 ; *X-Ray Properties of Black-Hole Binaries*. ARA&A, **44**, 49

REVNIVTSEV, M., SAZONOV, S., CHURAZOV, E., FORMAN, W., VIKHLININ, A. & SUNYAEV, R., 2009 ; *Discrete sources as the origin of the Galactic X-ray ridge emission*. Nature, **458**, 1142

REVNIVTSEV, M., SAZONOV, S., GILFANOV, M., CHURAZOV, E. & SUNYAEV, R., 2006 ; *Origin of the Galactic ridge X-ray emission*. A&A, **452**, 169

REYNOLDS, S. P., 2008 ; *Supernova Remnants at High Energy*. ARA&A, **46**, 89

ROBERTS, T. P., 2007 ; *X-ray observations of ultraluminous X-ray sources*. Ap&SS, **311**, 203

RYBICKI, G. B. & LIGHTMAN, A. P., 1979 ; *Radiative processes in astrophysics*. New York, Wiley-Interscience, 1979. 393 p.

SCHMITT, J. & STELZER, B., 2008 ; *The Universe in X-Rays*. Springer, Berlin ; New York

SCHMITT, J. H. M. M., 1997 ; *Coronae on solar-like stars.* A&A, **318**, 215

SCHMITT, J. H. M. M. & LIEFKE, C., 2004 ; *NEXXUS : A comprehensive ROSAT survey of coronal X-ray emission among nearby solar-like stars*. A&A, **417**, 651

SHAKURA, N. I. & SUNYAEV, R. A., 1973 ; *Black holes in binary systems. Observational appearance.* A&A, **24**, 337

SHIPMAN, H. L., 1976 ; *Sirius B - A thermal soft X-ray source*. ApJ, **206**, L67

STASIŃSKA, G., CID FERNANDES, R., MATEUS, A., SODRÉ, L. & ASARI, N. V., 2006 ; *Semi-empirical analysis of Sloan Digital Sky Survey galaxies - III. How to distinguish AGN hosts*. MNRAS, **371**, 972

STAUBERT, R., 2008 ; *The Universe in X-Rays*. Springer, Berlin ; New York

STOBBART, A.-M., ROBERTS, T. P. & WILMS, J., 2006 ; *XMM-Newton observations of the brightest ultraluminous X-ray sources*. MNRAS, **368**, 397

STRASSMEIER, K. G., HALL, D. S., FEKEL, F. C. & SCHECK, M., 1993 ; *A catalog of chromospherically active binary stars (second edition)*. A&AS, **100**, 173

STRASSMEIER, K. G., HALL, D. S., ZEILIK, M., NELSON, E., EKER, Z. & FEKEL, F. C., 1988 ; *A catalog of chromospherically active binary stars*. A&AS, **72**, 291

STRITTMATTER, P. A., BRECHER, K. & BURBIDGE, G. R., 1972 ; *X-Ray Emission from White Dwarfs*. ApJ, **174**, 91

STRÜDER, L., ASCHENBACH, B., BRÄUNINGER, H., DROLSHAGEN, G., ENGLHAUSER, J., HARTMANN, R., HARTNER, G., HOLL, P., KEMMER, J., MEIDINGER, N., STÜBIG, M. & TRÜMPER, J., 2001 ; *Evidence for micrometeoroid damage in the pn-CCD camera system aboard XMM-Newton*. A&A, **375**, L5

TANAKA, Y., 2008 ; *The Universe in X-Rays*. Springer, Berlin ; New York

TOMCZYK, S., MCINTOSH, S. W., KEIL, S. L., JUDGE, P. G., SCHAD, T., SEELEY, D. H. & EDMONDSON, J., 2007 ; *Alfvén Waves in the Solar Corona*. Science, **317**, 1192

TRÜMPER, J. E. & HASINGER, G., 2008 ; *The Universe in X-Rays*. Springer, Berlin ; New York

TUGAY, A. V. & VASYLENKO, A. A., 2009 ; *XMM-Newton Observations of X-ray Pulsar Cen X-3*. Dans *YSC'16 Proceedings of Contributed Papers (eds. Choliy V. Ya., Ivashchenko G.), pp. 58-61*, V. Y. Editors : Choliy & G. Ivashchenko, réd.

VAN BELLE, G. T. & VON BRAUN, K., 2009 ; *Directly Determined Linear Radii and Effective Temperatures of Exoplanet Host Stars*. ApJ, **694**, 1085

VÉRON-CETTY, M.-P. & VÉRON, P., 2006 ; *A catalogue of quasars and active nuclei : 12th edition*. A&A, **455**, 773

VIALLET, M., 2008 ; *Irradiation de lâĂŹenvironnement et instabilités du disque dâĂŹaccrétion dans les systèmes binaires et les noyaux actifs de galaxies*. Thèse de doctorat, Strasbourg, France

VINK, J., 2004 ; *A Review of X-ray Observations of Supernova Remnants*. Nuclear Physics B Proceedings Supplements, **132**, 21

VRIELMANN, S., NESS, J. & SCHMITT, J. H. M. M., 2005 ; *On the nature of the X-ray source in GK Persei*. A&A, **439**, 287

WARNER, B., 1995 ; *Cataclysmic variable stars*

WARWICK, R. S., TURNER, M. J. L., WATSON, M. G. & WILLINGALE, R., 1985 ; *The galactic ridge observed by EXOSAT*. Nature, **317**, 218

WATSON, M. G., SCHRÖDER, A. C., FYFE, D., PAGE, C. G., LAMER, G., MATEOS, S., PYE, J., SAKANO, M., ROSEN, S., BALLET, J., BARCONS, X., BARRET, D., BOLLER, T., BRUNNER, H., BRUSA, M., CACCIANIGA, A., CARRERA, F. J., CEBALLOS, M., DELLA CECA, R., DENBY, M., DENKINSON, G., DUPUY, S., FARRELL, S., FRASCHETTI, F., FREYBERG, M. J., GUILLOUT, P., HAMBARYAN, V., MACCACARO, T., MATHIESEN, B., MCMAHON, R., MICHEL, L., MOTCH, C., OSBORNE, J. P., PAGE, M., PAKULL, M. W., PIETSCH, W., SAXTON, R., SCHWOPE, A., SEVERGNINI, P., SIMPSON, M., SIRONI, G., STEWART, G., STEWART, I. M., STOBBART, A., TEDDS, J., WARWICK, R., WEBB, N., WEST, R., WORRALL, D. & YUAN, W., 2008 ; *The XMM-Newton Serendipitous Survey. VI. The Second XMM-Newton Serendipitous Source Catalogue*. ArXiv e-prints, **807**

WERNER, K., 1992 ; *Analysis of PG1159 stars (review)*. Dans *The Atmospheres of Early-Type Stars*, U. Heber & C. S. Jeffery, réds., tm. 401 de *Lecture Notes in Physics, Berlin Springer Verlag*

—, 2008 ; *The Universe in X-Rays*. Springer, Berlin ; New York

WERNER, K., DEETJEN, J. L., DREIZLER, S., NAGEL, T., RAUCH, T. & SCHUH, S. L., 2003 ; *Model Photospheres with Accelerated Lambda Iteration*. Dans *Stellar Atmosphere Modeling*, I. Hubeny, D. Mihalas & K. Werner, réds., tm. 288 de *Astronomical Society of the Pacific Conference Series*

WERNER, K., HEBER, U. & HUNGER, K., 1991 ; *Non-LTE analysis of four PG1159 stars*. A&A, **244**, 437

WORRALL, D. M., MARSHALL, F. E., BOLDT, E. A. & SWANK, J. H., 1982 ; *HEAO 1 measurements of the galactic ridge*. ApJ, **255**, 111

YOUNG, M., ELVIS, M. & RISALITI, G., 2008 ; *The Reddest DR3 SDSS/XMM Quasars*. ArXiv e-prints, **809**

Chapitre II

Corrélation croisée des sources XMM

La corrélation croisée (ou *cross-correlation* en anglais) des sources du catalogue 2XMMi avec des catalogues d'archives est une étape préliminaire très importante. C'est cette étape qui permet de récupérer les informations multi-longueurs d'onde indispensables pour la classification des sources X. De nombreuses stratégies sont possibles pour effectuer une corrélation croisée de deux catalogues. Les plus simples consistent par exemple à :

- choisir dans le catalogue secondaire la source la plus proche de celle du catalogue primaire, en fixant éventuellement un rayon de recherche maximum
- choisir, dans un rayon donné autour de la source primaire, la source la plus brillante du catalogue secondaire

Aucune de ces deux méthodes ne permet de quantifier la probabilité pour que la source du catalogue primaire et sa source associée dans le catalogue secondaire soient effectivement une seule et même source. Pour calculer une telle probabilité d'identification, nous avons recours à une méthode basée sur les probabilités Bayesiennes. Ce chapitre est dédié à la présentation, en détails, de la solution que nous avons mise en place et utilisée.

Après avoir posé le cadre général du problème de corrélation croisée, ainsi que les hypothèses simplificatrices auxquelles nous avons recours pour le résoudre (1), nous détaillerons la méthode mise en place (2). Celle-ci est basée sur les rapports de vraisemblance (LR pour *likelihood ratio*), et sur une estimation originale du taux de fausses identifications par tranche de LR. La partie suivante (3) aborde certains détails de calcul, dont un peu de géométrie sphérique, laissés de coté dans la partie (2) pour ne pas alourdir le texte. Le calcul du LR, nous le verrons, nécessite l'estimation de densités locales de sources. La dernière partie de ce chapitre (5) sera consacrée aux méthodes disponibles pour estimer ces densités et mettra plus particulièrement l'accent sur le kernel smoothing que nous avons largement utilisé. Deux méthodes de recalibration des positions des sources X, complémentaires, seront également abordées (4). Ces méthodes permettent de minimiser le biais dû à l'éventuelle existence d'un écart systématique (ou *offset*) entre les sources X et les source d'archives.

1 Cadre général du problème et simplifications

Comme nous avons à travailler sur les données XMM, nous allons nous placer dans le champ de vue (FOV pour *Field Of View*) de l'instrument EPIC. Ce qui suit peut être généralisé à un tout autre découpage du ciel, voir même à des catalogues *all-sky*, qui couvrent la totalité du ciel.
On suppose que dans notre champ de vue XMM de 15 arcmin de rayon, N_x sources X ont été détectées. Pour ce même champ de vue, on dispose de N_o sources issues d'un catalogue d'archive qui peut couvrir les longueurs d'onde optiques, infrarouges (IR), radio, X ou gamma.

1.1 Problèmes de positionnement

Les premières difficultés auxquelles nous sommes confrontés concernent le référentiel des positions des sources. Il faut que les sources soient placées dans le même référentiel. On dispose en général de coordonnées ICRS (Seidelmann & Kovalevsky 2002) ou équatoriales J2000. Mais il n'est pas rare que des biais existent dans la ou les positions de référence. Par exemple, dans le cas d'XMM, on constate qu'il peut y avoir une légère rotation, de l'ordre de quelques dixièmes de degrés, entre l'axe nord-sud du référentiel du FOV XMM et celui des coordonnées équatoriales. De même, il peut y avoir un *offset* entre les coordonnées du centre du FOV et sa vraie position. Ces effets peuvent conduire, en bord de champs, à des différences de positions de l'ordre de quelques secondes d'arc. Nous verrons section (4) et plus particulièrement paragraphe (4.2) comment on peut corriger, dans certains cas, ces effets.
Pour les anciens relevés réalisés à partir de plaques photographiques, tel que l'USNOB1 (Monet *et al.* 2003), l'optique et la nature des plaques sont à l'origine de distortions dont les effets augmentent au fur et à mesure que l'on s'éloigne du centre optique. Des modèles de ces déformations ont été établis afin de corriger au mieux ces déformations. Il existe cependant toujours des effets résiduels qui ajoutent, en plus de la PSF, une incertitude sur le positionnement des sources. Ces effets sont souvent négligeables ou introduits dans l'erreur de positionnement des sources via une erreur systématique.

L'autre difficulté, plus gênante, concerne le référentiel temporel. Les observations utilisées pour la construction d'un catalogue et les observations XMM ne sont pas réalisées en même temps. Or le ciel n'est pas statique. Chaque source dispose de mouvements propres qui sont à l'origine du changement de leurs positions dans le temps. Ce phénomène touche principalement les objets proches, donc essentiellement les étoiles brillantes, et n'affecte pas les quasars. Ces derniers sont trop lointains pour que leurs déplacements sur le ciel soient mesurables. Les coordonnées doivent donc être ramenées à une certaine date. Le repère spatial et temporel doit être le même. Le choix classique pour les coordonnées équatoriales J2000 est de prendre l'époque J2000.0. Ce n'est possible que si on dispose des vitesses tangentielles des différentes sources. Si ces mesures existent elles ne sont pas toujours fiables et sont elles même entachées d'erreurs elliptiques qu'il faut prendre en compte et rajouter aux erreurs de positionnement. Nous avons tenu compte des mouvements propres pour les sources du catalogue Tycho 2. Pour les autres catalogues, et dans ce qui suit, nous négligerons ces effets.

On fait maintenant l'hypothèse que les coordonnées de chaque source et les ellipses d'erreurs associées sont exprimées en coordonnées équatoriales J2000 à l'époque J2000.0. On fait également l'hypothèse simplificatrice que dans ce système les positions des sources tiennent compte des vitesses tangentielles, ou que ces vitesses sont nulles. On fait également l'hypothèse qu'il n'y a pas de biais dans la détermination des positions des deux catalogues.

1.2 Des origines de la différence des sources observées

Toutes les sources que nous observons font partie d'une même distribution, non connue, de sources en une région du ciel donnée. Faisons l'hypothèse que cette distribution spatiale projetée est localement poissonienne. Cette supposition implique qu'on ne prend pas en considération les effets de regroupement dûs : aux systèmes binaires ou ternaires résolus, aux amas de galaxies, ...

Pour un certain nombre de raisons physiques et techniques, seul un nombre N_x de ces sources apparaissent dans notre champ de vue et seul un certain nombre N_o de ces sources apparaissent dans le catalogue d'archive. Parmis ces raisons, on peut citer par ordre d'importance :

- l'émission intrinsèque des objets qui n'est pas la même selon la longueur d'onde ;
- les sensibilités différentes des instruments ;
- les temps de pose plus ou moins longs pour des observations réalisées par un même instrument ;
- la colonne densité entre les objets et l'instrument dont l'absorption varie en fonction de la longueur d'onde ;

- les décalages en longueur d'onde, différents en raison des *redshifts* et, plus généralement, des vitesses d'éloignement (vitesses radiales), qui donnent accès à diverses régions des spectres des objets ;
- la résolution spatiale différente pour chaque instrument (confusion, *blending*) ;
- l'état d'un objet variable observé à deux moments différents ;
- les phénomènes transitoires ;
- les conditions d'observation (par exemple, le *seeing* est variable dans le temps pour les télescopes terrestres) ;
- les divers algorithmes d'extraction de sources ;
- ...

On note N_b (b pour *both*) le nombre de sources détectées à la fois par XMM et par l'instrument du catalogue des contreparties. Ce chiffre est a priori inconnu.

1.3 Superposition X de sources résolues en optique

En réalité, il faut séparer N_b en deux nombres pour tenir compte du pouvoir séparateur de chaque instrument :

- N_{bx} : nombre de sources X qui sont réellement associées à des sources du catalogue des contreparties
- N_{bo} : nombre de sources du catalogue des contreparties qui sont réellement associées à des sources X

Selon la résolution des instruments, on peut avoir $N_{bx} \leq N_{bo}$ ou $N_{bx} \geq N_{bo}$. En effet, une source X peut, par exemple, contenir la somme des émissions X de deux sources proches résolues en optique mais pas en X.

Dans la suite on suppose qu'on a pas de tels effets, ce qui, en pratique, est faux. On suppose donc que N_b est le nombre de sources communes en X et dans le catalogue de contreparties. On suppose également que pour une source X donnée ne peut correspondre qu'une contrepartie dans le catalogue d'archive.

2 Algorithme de corrélation croisée mis en place

On vient de voir, dans la section précédente, certaines difficultés intrinsèques au problème de corrélation de sources en astronomie. On ignore maintenant certaines de ces complications et on prend en compte les simplifications mentionnées. On les rappellera quand elles interviendront dans ce qui suit.

Cette section donne une vision d'ensemble du système de corrélation croisée mis en place dont certains points techniques seront détaillés dans la section suivante (3).

On commence cette description en abordant (2.1) le critère de sélection des sources d'archives candidates. Cette sélection prend en compte des erreurs elliptiques sur les coordonnées sphériques des sources. On calcule ensuite un rapport de vraisemblance pour chaque paire source X - source d'archive (2.2). Ce rapport de vraisemblance tient compte de mesures de densités locales obtenues par un kernel smoothing (2.3). L'estimation de la distribution du taux de fausses associations par tranches de LR (2.4) permet alors de dériver les probabilités d'identification, basées uniquement sur la coïncidence en position, des associations source X - source d'archive (2.5).

Dans la suite, on utilise par commodité le terme "source optique", mais on peut généraliser aux sources de différentes longueurs d'onde.

2.1 Critère de sélection des candidats optiques

On considère une source X et un candidat optique avec :

- α_X, δ_X les coordonnées équatoriales de la source X
- σ_{α_X}, σ_{δ_X} et $\rho_{X_{\alpha,\delta}}$ respectivement les erreurs sur $\alpha_X \cos\delta_X$ et sur δ_X et la corrélation entre σ_{α_X} et σ_{δ_X}

– α_o, δ_o les coordonnées équatoriales d'une source optique
– σ_{α_o}, σ_{δ_o} et $\rho_{o_{\alpha,\delta}}$ respectivement les erreurs sur $\alpha_o \cos\delta_o$ et sur δ_o et la corrélation entre σ_{α_o} et σ_{δ_o}

Comme il est généralement fait implicitement, sauf dans Budavari & Szalay (2007), on convertit le problème sphérique en problème plan et les erreurs sur les positions sont directement interprétées comme des Gaussiennes 2D classiques. On traite cet aspect dans le paragraphe (3.1.2) où on justifie ce choix en quantifiant l'approximation ainsi réalisée. On a choisi un plan de projection 2D centré sur la position de la source X et ayant pour axe des x la direction du candidat optique (Fig. II.1). Dans ce référentiel, les erreurs sur les positions deviennent les Gaussiennes : $\mathcal{N}_X(x,y;\sigma^2_{x_X},\sigma^2_{y_X},\rho_X\sigma_{x_X}\sigma_{y_X})$ et $\mathcal{N}_o(x-d,y;\sigma^2_{x_o},\sigma^2_{y_o},\rho_o\sigma_{x_o}\sigma_{y_o})$ avec d la distance angulaire entre les sources X et optique. Comme suggéré dans Sinnott (1984), d est calculé par la fonction d'Haversine (voir Eq. II.20).

La densité de probabilité pour que les deux sources soient à la même position, et donc qu'elles soient issues d'un même objet, est donnée par le produit de convolution de ces deux distributions. Ce qui nous donne une nouvelle Gaussienne :

$$P(x,y) = \mathcal{N}_c(x,y;\sigma^2_{x_c},\sigma^2_{y_c},\rho_c\sigma_{x_c}\sigma_{y_c}) \tag{II.1}$$

avec $\sigma^2_{x_c} = \sigma^2_{x_X} + \sigma^2_{x_o}$, $\sigma^2_{y_c} = \sigma^2_{y_X} + \sigma^2_{y_o}$ et $\rho_c\sigma_{x_c}\sigma_{y_c} = \rho_X\sigma_{x_X}\sigma_{y_X} + \rho_o\sigma_{x_o}\sigma_{y_o}$.

Si la source optique est la contrepartie de la source X, elle se trouve, avec une probabilité γ, dans l'ellipse définie par l'équation :

$$\begin{pmatrix} x \\ y \end{pmatrix}^t \begin{pmatrix} \sigma^2_{x_c} & \rho_c\sigma_{x_c}\sigma_{y_c} \\ \rho_c\sigma_{x_c}\sigma_{y_c} & \sigma^2_{y_c} \end{pmatrix}^{-1} \begin{pmatrix} x \\ y \end{pmatrix} = k^2_\gamma \tag{II.2}$$

La complétude que nous avons choisie correspond au critère des 3σ, qui sert souvent de compromis entre le nombre total d'associations et le nombre de contreparties manquées (0.3%). Pour obtenir cette complétude, $\gamma = 99.7\%$, en 2D, il faut prendre la valeur $k_\gamma = 3.43935$. Dans le repère que nous avons choisi, les coordonnées de la source optique sont $x = d$ et $y = 0$. La critère de sélection que nous adoptons est de garder toutes les sources candidates qui satisfont :

$$\boxed{\frac{d}{\sigma_{x_c}\sqrt{1-\rho_c^2}} \leq k_\gamma} \tag{II.3}$$

Comme mentionné en introduction, on fait les hypothèses suivantes : on néglige les décalages systématiques qui peuvent exister entre les deux catalogues ; on suppose que les positions et leurs erreurs associées ont été calculées dans un même référentiel, à une même époque et qu'elles tiennent compte des mouvements propres.

2.2 LE RAPPORT DE VRAISEMBLANCE, OU *likelihood ratio*

Pour chaque association source X - source optique vérifiant le critère Eq. II.3, on calcule un rapport de vraisemblance (LR pour *likelihood ratio*). Ce LR est défini comme étant le rapport entre la probabilité pour avoir la contrepartie à la distance normalisée r, et la probabilité d'avoir une source aléatoire, due à une distribution poissonienne de source, à cette même distance.

La densité de probabilité pour que les deux sources soient à la même position sachant x et y correspond à la densité de probabilité pour avoir une contrepartie en x et y, en faisant l'hypothèse que la source optique correspond au même objet astrophysique que la source X. La Gaussienne $\mathcal{N}_c$ (eq. II.1) peut être écrite, sous sa forme canonique, $\frac{1}{2\pi\sigma_M\sigma_m}\exp -\frac{1}{2}(\frac{x_1^2}{\sigma_M^2} + \frac{y_1^2}{\sigma_m^2})$. Où σ_M et σ_m sont les demi-grand axe et demi-petit axe, dans le repère défini par les vecteurs propres (x_1, y_1) issus de la diagonalisation de la matrice de variance-covariance de $\mathcal{N}_c$:

$$\sigma_{M,m} = \frac{1}{2}(\sigma^2_{x_c} + \sigma^2_{y_c} \pm \sqrt{(\sigma^2_{x_c} - \sigma^2_{y_c})^2 + 4(\rho\sigma_{x_c}\sigma_{y_c})^2}) \tag{II.4}$$

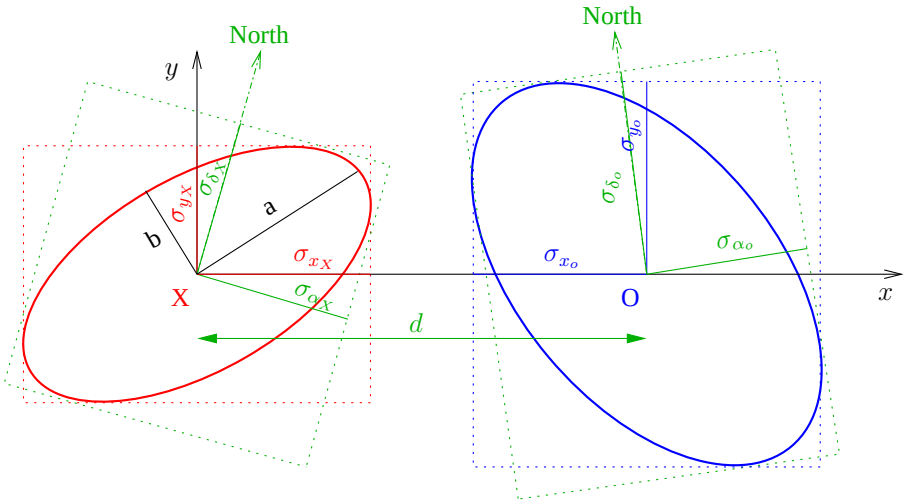


FIG. II.1 – Plan de projection choisi : le repère xy est centré sur la source X (point X) ; l'axe des x est en direction du candidat optique (point O). d est la distance angulaire entre les deux sources. Ce repère est indispensable à hautes latitudes où on ne peut plus considérer que les méridiennes sont parallèles. On peut donc traiter proprement le cas des sources se trouvant près des pôles.

On change l'échelle et on passe en coordonnées polaires, ce qui permet d'obtenir la distance angulaire sans dimension :

$$r = \sqrt{\frac{x_1^2}{\sigma_M^2} + \frac{y_1^2}{\sigma_m^2}} \tag{II.5}$$

Ainsi, la nouvelle surface élémentaire devient $\pi\sigma_M\sigma_m$, la surface de l'ellipse à 1σ (obtenue en fixant $r = 1$).

Le LR que nous utilisons est inspiré de celui décrit dans De Ruiter *et al.* (1977). Comme Wolstencroft *et al.* (1986), on ne considère pas seulement le premier candidat mais toutes les sources qui satisfont l'équation Eq. II.3. On remplace donc la probabilité de trouver le premier objet aléatoire à une distance comprise entre r et $r + \mathrm{d}r$ par celle de trouver *un* objet aléatoire à une distance comprise entre r et $r + \mathrm{d}r$.

La probabilité de trouver la contrepartie optique (cp) à une distance comprise entre r et $r + \mathrm{d}r$ vaut :

$$\mathrm{d}p(r|cp) = re^{-\frac{1}{2}r^2}\mathrm{d}r \tag{II.6}$$

Et celle de trouver un objet aléatoire (spur pour "spurious") à une distance comprise entre r et $r + \mathrm{d}r$ est donnée par la loi de Poisson :

$$\mathrm{d}p(r|spur) = 2\lambda r\mathrm{d}r \tag{II.7}$$

Nous avons choisi de prendre comme densité la densité locale de sources au moins aussi brillantes que m_o, la magnitude de la source candidate. Dans l'ellipse à 1σ (Eq. II.5), la nouvelle surface élémentaire, la densité Poissonienne de source vaut $\lambda = \pi\sigma_M\sigma_m f(\alpha_x, \delta_x, m < m_o)$.

Puisque, avec cette définition, le nombre de sources disponibles dans une même surface est plus important que si on prenait en compte une seule tranche de magnitudes, les densités calculées sont plus précises (ou plus locales) que les méthodes basées sur des tranches de magnitudes. Avec cette définition, c'est comme si on calculait les densités locales de sources avec un instrument de plus en plus sensible.

Le rapport de vraisemblance (LR) est le rapport de ces deux densités de probabilités (II.6 et II.7) :

$$\boxed{LR(r) = \frac{\mathrm{d}p(r|cp)}{\mathrm{d}p(r|spur)} = \frac{1}{2\lambda}e^{-\frac{1}{2}r^2}} \tag{II.8}$$

Le formalisme que l'on applique ici a pour but de calculer des probabilités d'identification basées uniquement sur la coïncidence en position. On n'utilise pas d'information supplémentaire sur les sources telle que le spectre de distribution en énergie (SED pour *Spectral Energy Distribution*). De fait, on n'ajoute pas de terme supplémentaire à notre LR comme c'est la cas, par exemple dans Wolstencroft *et al.* (1986), Sutherland & Saunders (1992) et Brusa *et al.* (2007). La quantité $q(m)$ qu'ils définissent correspond à la probabilité d'avoir, parmis les contreparties réelles, une source de magnitude m, ou dans une tranche Δm autour de m (voir la formule Eq. II.41). Dans ce cas, $q(m)$ devrait être local mais devient alors difficile à estimer. En général, l'estimation de $q(m)$ est entachée de larges erreurs qui, en plus des erreurs sur l'estimation de la densité locale, affectent l'erreur sur le LR.

Nous verrons dans la section 2.4 que ce facteur $q(m)$ est, d'une certaine manière, pris en compte dans notre fonction de fiabilité.

2.3 Estimation de la densité locale

La précision du *likelihood ratio* (Eq. II.8) dépend de la précision de l'estimation de la densité locale de sources $f(\alpha_x, \delta_x, m < m_o)$. Une façon simple d'estimer une densité locale est d'avoir recours au k-NN, détaillé section 5.2. Dans ce cas, l'estimation de la densité locale a une erreur bien connue valant $\sqrt{k}/\pi d_k^2$, avec d_k la distance au $k^{\text{ème}}$ plus proche voisin de la source X. Cependant, le k-NN n'est pas lisse : il présente des maximums locaux en lesquels il n'est pas dérivable ; il peut fluctuer de façon importante entre deux points peu distants ayant ou non les k mêmes voisins. De fait, deux candidats ayant une magnitude proche – ou deux sources proches ayant une même magnitude – peuvent, avec cette méthode, avoir des LRs assez différents. Nous présentons, section 5 d'autres méthodes pour calculer des densité locales de sources.

Nous avons choisi d'utiliser le *sample-point density estimator*, qui consiste en un *kernel smoothing* à largeur de bande variable. Nous avons choisi comme fonction noyau le profil d'Epanechnikov et, comme largeur de bande, la distance au $100^{\text{ème}}$ plus proche voisin.

Nous n'avons pas utilisé de *kernel smoothing* à largeur de bande fixe déterminée, par exemple, par la méthode de Markov Chain Monte Carlo (MCMC) présentée section 5.5.2. En effet, dans notre cas on considère pour chaque candidat la densité de sources au moins aussi brillantes que le candidat. Dans ces conditions, il faudrait créer des cartes de densités avec à peu près autant de cartes que de candidats. L'algorithme étant assez lent, on ne peut pas se permettre de l'appliquer N fois.

Plus k est petit, plus la densité estimée est locale, mais plus l'erreur faite sur son estimation est grande. L'utilisation de $k = 100$ mène à une erreur relative sur le LR de l'ordre de 15%, ce qui correspond à une erreur relative d'environ 6.5% sur $\log_{10}(LR)$. Notre algorithme prend en compte les effets de bords.

Nous avons testé le k-NN (knn) et le *kernel smoothing* (ks) sur des données simulées. Une carte de densité de taille 400x400, soit 2.25 secondes d'arc de côté pour chaque pixel, a été calculée sur un FOV synthétique contenant 20,000 sources uniformément réparties auxquelles on a ajouté 500 sources issues d'une distribution Gaussienne 2D de moyennes 3 et 2 respectivement en x et y et ayant la valeur 1.5 comme déviation standard. Le rayon du FOV vaut 15 arcminutes. L'utilisation de $k = 64$ pour le knn et de $k = 100$ pour le ks mène approximativement à la même erreur relative – respectivement 13% et 15% – sur l'estimation de la densité locale (voir Fig. II.2). L'ajustement d'une Gaussienne 2D centrée en (3,2) sur le profil de densité obtenu dans un rayon de 3σ ($r = 4.5$) mène à une déviation standard de $\sigma_f = 1.476$ et $\sigma_f = 1.492$ respectivement pour le knn et le ks, pleinement compatible avec la valeur d'entrée de 1.5.

2.4 Calcul des fiabilités

2.4.1 Définition de la fiabilité

Bien que nous utilisions une définition différente du LR, une estimation différente du taux de fausses associations et une fonction d'ajustement de l'histogramme de fiabilité différente, nous avons plus ou moins suivi le travail présenté dans la partie 3 de Oyabu *et al.* (2005). La méthode

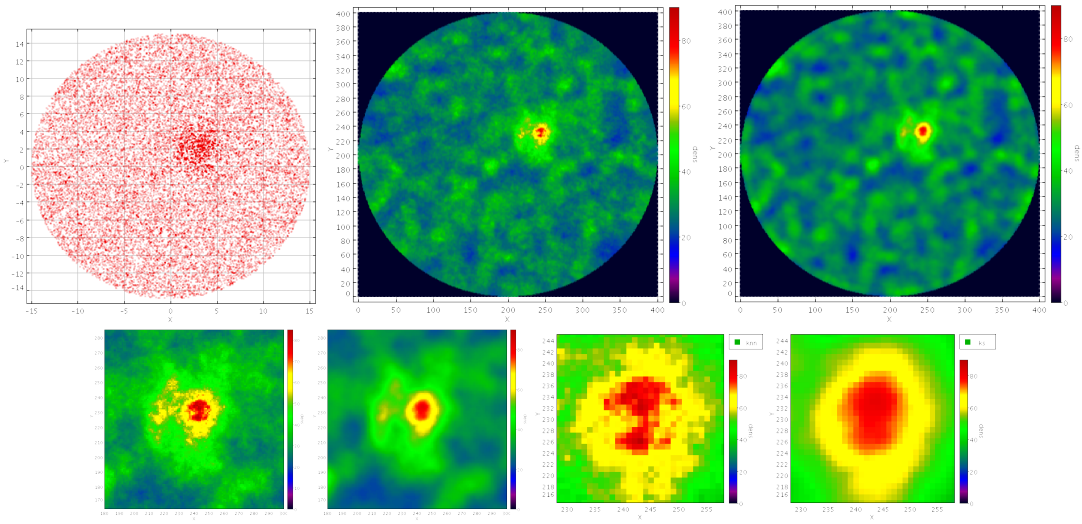


FIG. II.2 – Application d'un moyennage par k-NN et *kernel smoothing* sur des données simulées (voir section 2.3). En haut à gauche : distribution de sources générée (distribution uniforme + Gaussienne) ; les unités sont en minutes d'arc. En haut au centre : moyennage par k-NN avec k=64. En haut à droite : kernel smoothing avec k=100. En bas : zooms sur la partie Gaussienne de la distribution.

semble originellement venir de l'article Rutledge *et al.* (2000).

On définit la fiabilité R d'une association ayant un LR se trouvant dans une tranche de LR donnée par :

$$R(LR) = \frac{N_{real}(LR)}{N_{real}(LR) + N_{spur}(LR)} = \frac{N_{cand}(LR) - N_{spur}(LR)}{N_{cand}(LR)} \tag{II.9}$$

où N_{real} et N_{spur} sont respectivement les nombres, non connus, de candidats qui sont respectivement réels et faux pour une tranche de LR donnée ; N_{cand} est le nombre total de candidats dans une tranche de LR donnée. Si on se réfère à la section 1, $N_{real} = N_b$.

On doit donc estimer N_{spur}.

2.4.2 Estimation du taux de fausses associations

Une méthode souvent utilisée pour estimer N_{spur} consiste à corréler les sources X avec des échantillons de sources optiques synthétiques. Les échantillons générés ont les mêmes caractéristiques que les sources réelles : même densité, même distribution des erreurs sur les positions, etc. Les positions sont générées aléatoirement. La somme des résultats de ces échantillons issue de simulations de Monte-Carlo fournit une estimation du nombre de fausses associations en fonction des distances, des LRs, etc. Cette approche est utilisée, par exemple, par Oyabu *et al.* (2005). Dans Stephen *et al.* (2005), l'échantillon aléatoire consiste en une liste d'*"anti [...] sources"* qui sont *"le reflet en longitude et en latitude galactique"* des sources de départ.

Nous avons proposé une nouvelle méthode pour estimer le nombre de fausses associations. Cette dernière n'est pas basée sur des simulations de Monte-Carlo, mais sur le calcul direct des résultats qu'elles doivent fournir. Cette méthode offre des résultats plus précis et plus rapide qui ne dépendent pas du nombre de simulations réalisées. L'idée de départ, qui consiste à estimer la surface qu'occupe une association par rapport à la surface totale disponible, peut être trouvé dans Boller *et al.* (1998).

Considérons le cas d'une source X et d'une source optique de magnitude m dans un FOV. La probabilité d'observer la source X en un point (α_X, δ_X) du FOV est donnée par la distribution $g_X(\alpha_X, \delta_X)$ de sources X dans le FOV. La probabilité pour que la source optique de magnitude m soit aléatoirement associée avec la source X est donnée par l'intégrale de la distribution de sources optiques au moins aussi brillantes que m sur la surface de l'ellipse de convolution $S_{x,o} = k_\gamma^2 \pi \sigma_M \sigma_m$. La probabilité que l'association ait un LR dans une tranche ΔLR est donnée

par l'intégrale de la distribution de sources optiques au moins aussi brillantes que m sur la surface que cette tranche occupe dans l'ellipse de convolution $S_{x,o}$: $S_{x,o}(\Delta LR)$. Enfin, la somme de ces probabilités pour toutes les sources X et optiques fournit une estimation du nombre de fausses associations par tranche de LR. En considérant maintenant le cas de N sources X et M sources optiques dans un FOV, on trouve la formule :

$$N_{spur}(\Delta LR) = \sum_{x=1}^{N} \iint_{S_{FOV}} g_X(\alpha_X, \delta_X) \left(\sum_{o=1}^{M} \iint_{S_{x,o}(\Delta LR)} g_o(\alpha_o, \delta_o, m) \mathrm{d}\alpha_o \mathrm{d}\delta_o \right) \mathrm{d}\alpha_X \mathrm{d}\delta_X \tag{II.10}$$

où $S_{x,o}(\Delta LR)$ dépend de la source X, de la source optique, et de la position de la source X, et peut donc être écrit $S_{x,o}(\Delta LR, X, o, \alpha_X, \delta_X)$.

Une façon de simplifier le problème consiste à supposer que $g_o(\alpha_o, \delta_o, m)$ est à peu près constant dans la surface $S_{x,o}(\Delta LR)$. La formule (II.10) peut alors être écrite :

$$N_{spur}(\Delta LR) = \iint_{S_{FOV}} g_X(\alpha_X, \delta_X) \left(\sum_{x=1}^{N} \sum_{o=1}^{M} S_{x,o}(\Delta LR) g_o(\alpha_X, \delta_X, m) \right) \mathrm{d}\alpha_X \mathrm{d}\delta_X \tag{II.11}$$

Afin d'estimer $N_{spur}(\Delta LR)$ on doit calculer la carte de distribution complète de sources X et autant de cartes de distribution de sources optiques que de magnitudes différentes. Comme de tels calculs sont potentiellement très longs, on suppose que les sources optiques sont uniformément distribuées dans le FOV, et on choisit pour densité la valeur moyenne de leurs densités locales. Moins rigoureuse, cette version permet d'obtenir la simple formule :

$$\boxed{N_{spur}(\Delta LR) = \sum_{x=1}^{N} \sum_{o=1}^{M} S_{x,o}(\Delta LR) \,/ S_{FOV}} \tag{II.12}$$

La quantité $S_{x,o}(\Delta LR)$ peut être calculée comme suit : le rayon correspondant à un LR donné est facilement dérivé de l'équation (II.8) :

$$r^2(LR) = \begin{cases} 0 & \text{if} \quad LR \geq LR_{max} = \frac{1}{2\lambda} \\ \ln \frac{1}{(2\lambda LR)^2} & \text{if} \quad LR_{min} < LR < LR_{max} \\ k_\gamma^2 & \text{if} \quad LR \leq LR_{min} = \frac{1}{2\lambda} e^{\frac{1}{2}k_\gamma^2} \end{cases} \tag{II.13}$$

Donc, pour une tranche $\Delta LR = LR_2 - LR_1$, $LR_2 > LR_1$, si on a $LR_1 \leq LR_{max}$ et $LR_2 \geq LR_{min}$, la surface de la tranche de LR dans l'ellipse est exprimée par :

$$S_{x,o}(\Delta LR) = 2\pi\sigma_M\sigma_m \ln \frac{\min(\max(LR_2, LR_{min}), LR_{max})}{\max(\min(LR_{max}, LR_1), LR_{min})} \tag{II.14}$$

En principe, on devrait retirer de ce calcul les sources optiques qui sont des contreparties réelles puisqu'elle ne sont pas distribuées aléatoirement. Cependant, leur fraction peut être négligée pour les faibles valeurs de LR ; pour les hautes valeurs de LR, l'influence sur la probabilité d'identification est infime. En réalité la distribution de sources n'est pas complètement Poissonienne. En effet, par exemple, deux sources ne peuvent pas être infiniment proches à cause de la PSF, de la résolution de l'instrument et de l'algorithme d'extraction de sources. Il y a donc une petite surface autour de chaque source, notament les sources brillantes, qui ne peut pas en contenir une autre. Si on néglige cet effet, ce que l'on fait, on a tendance à surestimer légèrement le taux de fausses associations.

2.4.3 Calculs en pratique

Afin d'éviter de calculer trop de densités locales pour estimer le taux de fausses associations, on divise l'échelle des magnitudes en tranches et on associe à toutes les sources se trouvant dans une même tranche la moyenne de leurs densités locales. La largeur de la tranche dépend de la précision du catalogue sur la magnitude. On calcule ensuite pour toutes les sources optiques et toutes

les sources X le facteur $\sigma_M \sigma_m$, LR_{min} et LR_{max}. Il est ainsi possible de construire l'histogramme du nombre de fausses associations attendues en fonction des valeurs de LR. Pour augmenter la vitesse de calcul, on peut faire un histogramme des valeurs $\sigma_M \sigma_m$. Cependant, cette approche implique une nouvelle perte de précision pour une réduction parfois modeste des temps de calcul.

Comme montré sur la figure II.3, les histogrammes utilisés dans le calcul de la fiabilité sont le nombre de candidats et le nombre de fausses associations par tranche de $\log_{10}(LR)$.

2.4.4 Ajustement de la fonction de fiabilité

Afin d'estimer le nombre de fausses associations, on prend le cas assez souvent réaliste dans lequel une source X aurait au plus un candidat. La fiabilité d'une association (à ne pas confondre avec la fiabilité intégrée de toutes les associations ayant un $LR \geq l$), est directement donnée par (pour les détails, voir section 3.3, Eq. II.40) :

$$R(LR) = p(id|r) = \frac{1}{1 + \frac{p(spur)}{p(cp)} \frac{1}{LR}} \tag{II.15}$$

Le terme $p(cp)/p(spur)$ – probabilité que la source optique soit la contrepartie divisée par la probabilité qu'elle ne le soit pas – doit être indépendant de la distance normalisée r (Eq. II.5). Cette probabilité correspond au terme $(1-\theta)/\theta$ utilisé dans De Ruiter *et al.* (1977). Cependant, $p(cp)/p(spur)$ doit dépendre de la nature de la population de sources X (e.g. étoiles, noyaux actifs de galaxies) et doit donc varier avec les propriétés telles que la magnitude, la couleur ou les rapports de flux. Pour obtenir un LR similaire à celui utilisé dans Brusa *et al.* (2007) par exemple, il faudrait prendre en compte un paramètre supplémentaire $q(m)$ décrivant la variation de $p(cp)/p(spur)$ avec la magnitude (ou n'importe quelle autre propriété pertinente) du candidat. Autrement, $q(m)$ peut être remplacé par un autre terme jouant le rôle du B_{phot} dans Budavari & Szalay (2007). Un histogramme $R(LR)$ peut être construit à partir des histogrammes $N_{cand}(LR)$ et $N_{spur}(LR)$ obtenus à partir de la méthode présentée dans le paragraphe précédent. Si le rapport $p(cp)/p(spur)$ était indépendant des propriétés des sources, $R(LR)$ devrait pouvoir être ajusté à l'aide de la fonction II.15 en n'utilisant qu'un seul paramètre : $a = p(cp)/p(spur)$. Inclure un terme $q(m)$ dans LR avec $N_{\Delta m}$ tranches de magnitudes, requière la construction de $N_{\Delta m}$ histogrammes $R(LR, \Delta m)$ et de les ajuster par des fonctions II.15 ayant différents paramètres a. Cependant, en général, le manque de statistiques ne nous permet pas de le faire.

L'histogramme $R(LR)$ peut alors être vu comme la somme de $N_{\Delta m}$ histogrammes $R(LR, \Delta m)$, contenant chacun m_i sources, et peut donc être modélisé par la fonction suivante :

$$R_f(LR) = \sum_{i=1}^{N_{\Delta m}} \frac{b_i}{1 + a_i \frac{1}{LR}} \text{ , avec } b_i = \frac{m_i}{\sum_{j=1}^{N_{\Delta m}} m_j} \tag{II.16}$$

En pratique, les histogrammes ne sont pas fonction de LR mais de $\log_{10}(LR)$. Les meilleurs ajustements obtenus l'ont été avec la fonction :

$$R_f(x = \log_{10}(LR)) = \frac{\sum_{i=1}^{N_{\Delta m}} b'_i 10^{(1-i)x}}{1 + \sum_{i=1}^{N_{\Delta m}} a'_i 10^{-ix}} \tag{II.17}$$

avec $N_{\Delta m} = 3$, c'est à dire 6 paramètres.

L'ajustement est réalisé par un algorithme de Levenberg-Marquard. En pratique, on calcule autant de fonctions de fiabilité, et on crée et ajuste autant d' histogrammes de LR que le nombre de bandes de magnitude dans le catalogue d'archive.

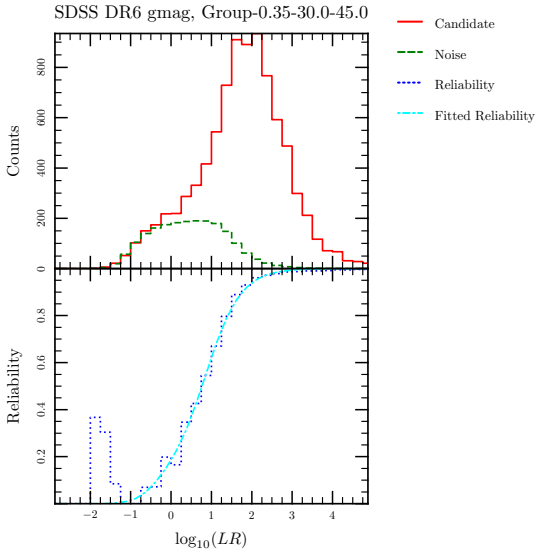


FIG. II.3 – Haut : histogrammes du nombre d'associations N_{cand} et du nombre de fausses associations N_{spur} estimés par tranche de $\log_{10}(LR)$. Bas : histogramme de fiabilité $R(\log_{10}(LR))$ par tranche de $\log_{10}(LR)$ et sa courbe d'ajustement. Dans l'exemple, les LRs ont été calculés à partir de magnitudes g issues du SDSS DR7 pour les sources du 2XMMi ayant une erreur systématique de 0.35 arcsec et une latitude galactique $30° < |b| < 45°$.

2.5 CALCUL DES PROBABILITÉS D'IDENTIFICATION DANS LE CAS GÉNÉRAL

On étend ici l'approche Bayesienne à des sources X ayant N_{cand} candidats. On suppose qu'au plus une association est réelle. Bien que cette hypothèse puisse être fausse, notamment dans le cas d'objets étendus ou de sources proches en optique et non résolues en X, elle permet de simplifier le problème et de le relier directement au cas où on a un seul candidat. On formule $N_{cand}+1$ hypothèses :

- H_{cp_i} : la i$^{\text{ème}}$ source optique est la contrepartie ;
- $H_{spur_{all}}$: il n'y a pas de contrepartie.

La probabilité Bayesienne que la i$^{\text{ème}}$ source soit la contrepartie sachant r_j, $j \in [1, N_{cand}]$ vaut (pour les détails, voir section 3.3.2, Eq. II.42) :

$$P_{id,i} = R'_i = p(H_{cp_i}|r_1 \cap \cdots \cap r_{N_{cand}}) = \frac{LR_i \frac{p(H_{cp_i})}{p(H_{spur_{all}})}}{1 + \sum_{j=1}^{N_{cand}} LR_j \frac{p(H_{cp_j})}{p(H_{spur_{all}})}} \tag{II.18}$$

Si $\frac{p(H_{cp_i})}{p(H_{spur_{all}})} = \frac{p(cp_i)}{p(spur_i)}$, l'équation (II.18) mène à la formule ci-dessous, obtenue en suivant les prescriptions de Rutledge *et al.* (2000) :

$$P_{id,i} = R'_i = \frac{\frac{R_i}{1-R_i}}{1 + \sum_{i=1}^{M} \frac{R_i}{1-R_i}} \tag{II.19}$$

A partir de l'équation II.15, on montre facilement que $R/(1-R) = LRp(cp)/p(spur)$. Si on calcule $p(H_{cp_i})$ et $p(H_{spur_{all}})$ en normalisant les termes $p(cp_i)$, $p(spur_i)$ comme Rutledge *et al.* (2000) le font pour construire $P_{id,i}$ à partir de R_i, on obtient l'égalité $\frac{p(H_{cp_i})}{p(H_{spur_{all}})} = \frac{p(cp_i)}{p(spur_i)}$. On applique donc la formule II.19 pour calculer la probabilité d'identification finale attachée à chaque association.

Chaque candidat possède autant de fiabilités que le nombre de bandes de magnitudes dans le catalogue d'archive. La valeur R_i prise en compte pour le calcul de la probabilité finale est, pour

chaque source, la meilleure parmi les différentes bandes photométriques.

Nous rappelons que ces probabilités d'identifications sont uniquement basées sur la coïncidence en position.

2.6 TESTS ET IMPLÉMENTATION

2.6.1 Groupement d'observations

Les sources XMM sont corrélées FOV par FOV. Or, afin de diminuer le bruit de comptage dans les tranches de LR de l'histogramme permettant le calcul de la fiabilité, sans perdre en résolution, il faut augmenter la statistique de comptage. On a donc additionné les données venant de FOV similaires :

- on a séparé les FOV XMM ayant des erreurs systématiques sur les positions différentes, à savoir 0.35″ et 1.0″ (voir section traitant des données d'XMM-Newton) ;
- on a mis à part les observations des petit et grand nuages de Magellan (SMC et LMC) ;
- puisqu'on suppose qu'elles contiennent des objets de même type et des lois logN-logS semblables, les observations ont été groupées en fonction de leur latitude galactique.

2.6.2 Tests sur des données simulées

Nous avons appliqué notre algorithme sur des échantillons de données simulées assez simples. On a généré 90 FOV contenant chacun 100 sources X distribuées aléatoirement dans des FOV de 15 minutes d'arc de rayon, ainsi que 90 FOV de 18 minutes d'arc de rayon contenant 2000 sources optiques. Les 3 minutes d'arc supplémentaires sont nécessaires pour éviter les effets de bord lors du calcul des densités locales. Les positions des sources X et optiques sont respectivement bruitées par des gaussiennes de 2.5 et 1 secondes d'arc. On a pour cela utilisé la méthode polaire de Marsaglia (voir Saporta 2006, p. 375-376). Parmis les 100 et 2000 sources, 30 sont communes, c'est-à-dire ont exactement la même position avant qu'elle ne soit bruitée. Toutes les sources optiques ont la même magnitude. On n'a donc pas généré de loi logN-logS. Une comparaison entre le taux de fausses associations et la fiabilité obtenus par notre méthode avec ceux entrés dans la simulation est donnée en exemple Fig. II.4. Les fiabilités estimées concordent bien avec celles des simulations.

2.6.3 Implémentation

Cet algorithme de corrélation croisée, de calcul et d'ajustement de la fiabilité a entièrement été écrit en JAVA. JAVA est un langage de programmation orientée objet qui est performant, portable, très répandu et qui possède des outils de développement de grande qualité (Eclipse ou encore NetBeans). Il est utilisé au seins de la communauté dans des logiciels tels que Topcat, Aladin ou encore SAADA. Le choix de ce langage a été motivé par la possibilité de réutiliser des codes et des librairies existants, par la simplicité du développement (outils puissant, programmation orienté objet, ...) ainsi que par ses performances parfois supérieures à celle du C.
Les sources du 2XMMi sont directement extraites de la XCATDB à l'aide de requêtes SQL et d'une connexion JDBC. Les sources d'archive sont récupérées du VO à l'aide du protocole ConeSearch. C'est une des raisons pour lesquelles la corrélation se fait FOV par FOV. Les résultats des corrélations sont ajoutés, via SQL et Saada, à la XCATDB et sont accessibles sur internet.

Les estimations de densités locales par *kernel smoothing* nécessitent, pour être performantes, de nombreuses recherches de type k-NN. De plus, pour chaque candidat, on est seulement intéressé par les sources de son voisinage ayant une magnitude inférieure ou égale à celle du candidat. Pour accélérer les recherches de type k-NN, les sources d'archive sont stockées dans des quadtrees spécialement conçus pour cette tâche : chaque feuille stocke la magnitude la plus grande parmi toutes les sources du sous-arbre ; les requêtes de type k-NN contiennent comme paramètre supplémentaire la magnitude du candidat ; en plus du critère de distance, on peut donc facilement stopper la recherche dans un sous-arbre en fonction d'un critère sur la magnitude.

On a préféré, pour ce cas particulier, les quadtree au kd-tree. En effet, les quadtree sont plus faciles à mettre à jour. On peut donc les créer en mode streaming, au fur et à mesure qu'on parse

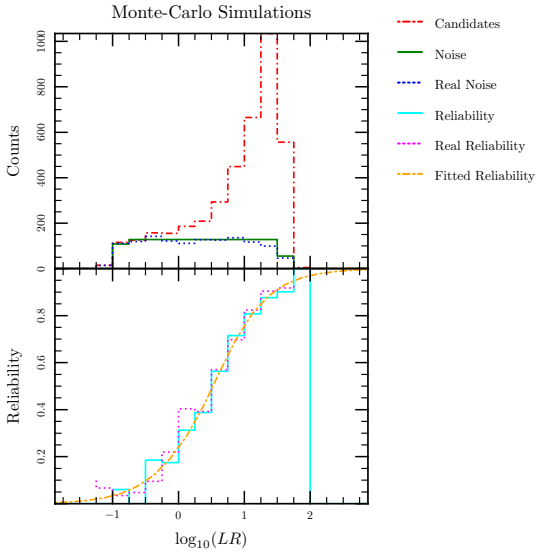


FIG. II.4 – Comparaison entre les fiabilités réelles et estimées à partir de données simulées (voir la section 2.6.2 pour les explications). Haut : la courbe en pointillés de taille variable représente le nombre total d'associations ; la ligne solide et celle en pointillés sont respectivement le nombre de fausses associations estimé par notre algorithme et le nombre réel de fausses associations. Bas : les courbes solide et en pointillés représentent respectivement la fiabilité estimée et la fiabilité réelle ; la courbe en semi-pointillé est la courbe d'ajustement de l'histogramme de fiabilité.

la VOTable retournée par une requête ConeSearch.

3 Approfondissement de quelques aspects techniques

Cette section est un peu technique et nous conseillons au lecteur d'y revenir plus tard s'il a besoin de démonstrations . Le lecteur peu soucieux de détails techniques pourra donc passer directement à la section 4.

Pour traiter proprement les convolutions d'ellipses d'erreur, notamment près des pôles, il faut choisir un référentiel approprié (3.1.1). Après avoir vérifié que le passage des coordonnées sphériques à un repère plan a, dans notre cas, un effet négligeable sur les distances et les surfaces, on pourra calculer proprement la distance en sigma entre deux sources et la probabilité qu'elles ont de se trouver à la même position (3.2).

Nous verrons ensuite les détails de l'origine du LR, la manière dont nous le calculons ainsi que le calcul de probabilité d'identification finale (3.3). Enfin, nous proposons la démonstration statistique basée sur le maximum de vraisemblance pour justifier la valeur de l'erreur sur l'estimation de densités locales par k-NN (3.4).

3.1 RÉFÉRENTIEL POUR LA CONVOLUTION DES ELLIPSES D'ERREUR

Pour calculer le produit de convolution des ellipses d'erreur sur les positions des sources X et des sources optiques, nous avons besoin de projeter ces ellipses en coordonnées sphériques sur un plan euclidien. Cette partie décrit et justifie le choix du référentiel 2D que nous avons adopté. Après un peu de géometrie sphérique et la définition de notre repère (3.1.1), nous quantifierons

les approximations résultant de cette projection (3.1.2).

3.1.1 Passage d'un repère local sphérique à un repère local plan

a Triangle sphérique défini par le pôle nord et deux sources

Sur la sphère unité de centre C, Fig. II.5, on considère le triangle sphérique NXO où N désigne le pôle Nord, X et O les positions respectives (α_x, δ_x) d'une source X et (α_o, δ_o) d'une source optique. On note :

- x, n, o les distances angulaires entre respectivement N et O, O et X et N et X
- $\hat{N}$, $\hat{X}$ et $\hat{O}$ les angles dièdres entre les plans respectivement NCX et NCO, XCN et XCO, OCN et OCX. Ces angles ont des valeurs comprises entre $-180°$ exclu et $180°$ inclus et sont définis positifs dans la direction Est.

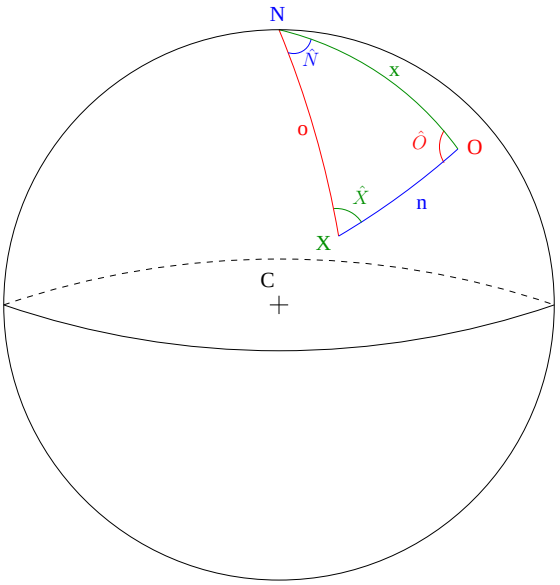


FIG. II.5 – Triangle quelconque sur une surface sphérique.

La valeur de l'angle $\hat{N}$ est immédiate :

$$\hat{N} = \Delta\alpha = \alpha_O - \alpha_X$$

On détermine n, la distance angulaire entre X et O à l'aide de la formule d'Haversine (Sinnott 1984) :

$$\boxed{n = 2 \cdot \arcsin\sqrt{\sin^2\frac{\delta_O - \delta_X}{2} + \sin^2\frac{\hat{N}}{2}\cos\delta_X\cos\delta_O}} \qquad \text{(II.20)}$$

On a également immédiatement les valeurs de x et o :

$$\begin{aligned} x &= 90° - \delta_O \\ o &= 90° - \delta_X \end{aligned}$$

On connaît la relation :

$$\frac{\sin\hat{N}}{n} = \frac{\sin\hat{X}}{x} = \frac{\sin\hat{O}}{o}$$

On remarquera que comme n, x et o ont des valeurs strictement positives, $\hat{X}$ et $\hat{O}$ sont du même signe que $\hat{N}$, qui dépend de $\alpha_O - \alpha_X$.
Comme la fonction $arcsin$ renvoie des valeurs $\in [-90, +90]$, on déduit deux valeurs possibles pour chaque angle $\hat{X}$ et $\hat{O}$. Le calcul du cosinus des angles moitiés $\hat{X}/2$ et $\hat{O}/2$ nous permet de

savoir respectivement si $|\hat{X}| > 90°$ et $|\hat{O}| > 90°$.
A partir des formules de trigonométrie sphérique, on écrit :

$$s = \frac{1}{2}(n + x + o)$$

$$\begin{aligned} \cos\frac{\hat{X}}{2} &= \sqrt{\frac{\sin s \sin(s-x)}{\sin n \sin o}} \\ \cos\frac{\hat{O}}{2} &= \sqrt{\frac{\sin s \sin(s-o)}{\sin n \sin x}} \end{aligned}$$

On peut donc exprimer la valeur des angles $\hat{X}$ et $\hat{O}$:

$$\hat{X} = \begin{cases} 180 - \arcsin\frac{x\cdot\sin\hat{N}}{n} & \text{si} \quad \cos\frac{\hat{X}}{2} > \frac{\sqrt{2}}{2} \quad \text{et} \quad \arcsin\frac{x\cdot\sin\hat{N}}{n} \geq 0 \\ -180 - \arcsin\frac{x\cdot\sin\hat{N}}{n} & \text{si} \quad \cos\frac{\hat{X}}{2} > \frac{\sqrt{2}}{2} \quad \text{et} \quad \arcsin\frac{x\cdot\sin\hat{N}}{n} < 0 \\ \arcsin\frac{x\cdot\sin\hat{N}}{n} & \text{sinon} \end{cases} \tag{II.21}$$

$$\hat{O} = \begin{cases} 180 - \arcsin\frac{o\cdot\sin\hat{N}}{n} & \text{si} \quad \cos\frac{\hat{O}}{2} > \frac{\sqrt{2}}{2} \quad \text{et} \quad \arcsin\frac{o\cdot\sin\hat{N}}{n} \geq 0 \\ -180 - \arcsin\frac{o\cdot\sin\hat{N}}{n} & \text{si} \quad \cos\frac{\hat{O}}{2} > \frac{\sqrt{2}}{2} \quad \text{et} \quad \arcsin\frac{o\cdot\sin\hat{N}}{n} < 0 \\ \arcsin\frac{o\cdot\sin\hat{N}}{n} & \text{sinon} \end{cases} \tag{II.22}$$

b *Passage à un repère plan*

On définit comme centre de notre repère cartésien le point X défini par la source X. Le plan choisi est le plan tangent, en X, à la sphère unité (Fig. II.6). Si on s'intéresse à des objets proches, disons dans un cercle de rayon $1°$ autour de ce centre, on peut considérer la distance angulaire n et l'angle $\hat{X}$ comme des coordonnées cartésiennes polaires. Soit N' la projection du pôle nord, N, sur le plan que nous venons de définir. On assimile alors l'axe XN' à l'axe des x. L'axe des y est l'axe passant par X, perpendiculaire à XN', et dirigé vers l'est. Les axes de ce repère sont gradués en unités d'angle.

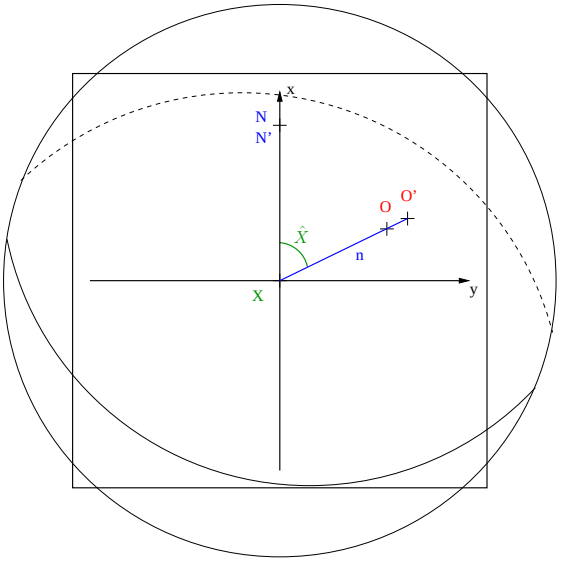


FIG. II.6 – Plan tangent en X à la sphère unité.

Une gaussienne en dimension 2 a une expression différente sur une sphère et sur une surface plane (Budavari & Szalay 2007). Cependant, lorsque les demi-axes sont de dimensions inférieures à quelques minutes d'arc, on peut considérer , avec une très bonne approximation, des gaussiennes 2D classiques dans le repère qu'on vient de définir.

c Repère défini pour le calcul du produit de convolution

Lors du calcul du produit de convolution de deux gaussiennes quelconques en dimension 2, il est important de connaître les orientations des deux ellipses d'erreur dans le même repère. L'angle d'orientation d'une ellipse, θ, est défini comme l'angle entre le pôle nord N et l'axe majeur de cette ellipse. Si on se trouve près d'un des deux pôles, la différence d'ascension droite $\hat{N} = \Delta\alpha$ peut être importante entre deux sources. L'orientation par rapport à laquelle est défini l'angle de leur ellipse (axe source – pôle nord dirigé vers le nord) peut donc être très différent (voir Fig. II.1).

Première solution On transforme le repère plan précédemment défini par un repère de même centre, X, mais où l'axe des x est maintenant défini par le rayon polaire n (Fig. II.7). Orienté positif vers l'Est. L'axe des y est défini par la droite perpendiculaire en X à l'axe des x et orientée vers le Nord. Si OX est sur un méridien, l'axe x est orienté plein Nord et l'axe des y plein Ouest. On note θ_x l'angle θ de l'axe majeur de l'ellipse d'erreur associée à la source X avec le pôle nord et θ_o l'équivalent pour la source candidate.

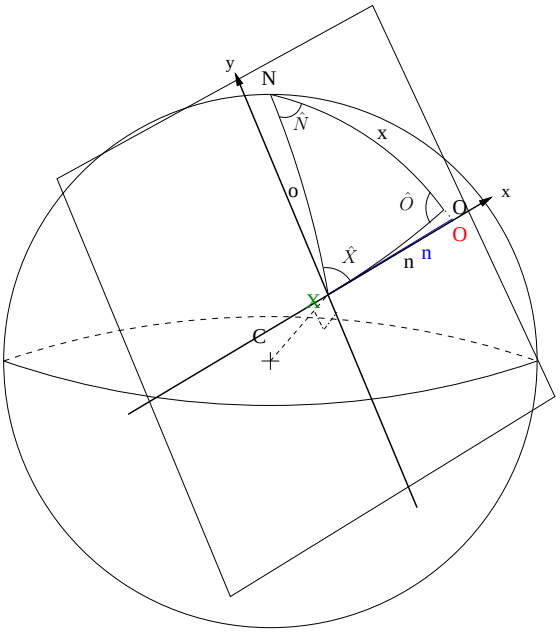


FIG. II.7 – Plan tangent en X à la surface de la sphère et ayant pour axe des x la direction de la source O.

Dans ce nouveau repère, les angles entre l'axe x et l'axe majeur des ellipses s'expriment à l'aide des formules :

$$\theta_X = \begin{cases} \hat{X} - \theta_x & \text{si} \quad \hat{X} \geq 0 \\ 180 + \hat{X} - \theta_x & \text{sinon} \end{cases}$$

$$\theta_O = \begin{cases} 180 - \hat{O} - \theta_o & \text{si} \quad \hat{O} \geq 0 \\ -\hat{O} - \theta_o & \text{sinon} \end{cases}$$

Deuxième solution L'autre solution consiste à conserver le repère plan ayant pour centre X et pour axe x l'axe XN'. Il faut alors calculer dans ce repère l'orientation de l'ellipse d'erreur sur la position de la source optique. D'après le schéma Fig. II.8, on obtient :

$$\begin{aligned} \theta_X &= \theta_x \\ \theta_O &= \theta_o + (\hat{X} - \hat{O} - 180^\circ) \end{aligned}$$

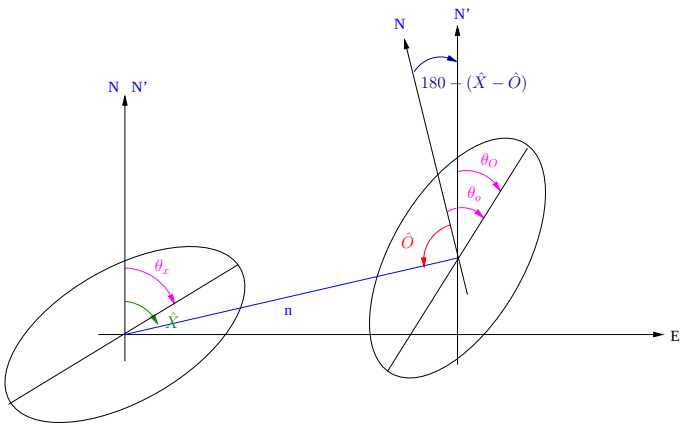


FIG. II.8 – Repère $(X, N\vec{o}rd, E\vec{s}t)$.

3.1.2 Erreurs associées au passage à un repère plan

a Erreurs sur les distances

Prenons le cas $n \leq 1°$. La mesure de distance cartésienne entre deux points qui subira la plus grande erreur par rapport à la distance en géometrie sphérique sera celle séparant 2 points de distance au centre $n = 1°$ et dont la différence d'angle $\Delta\hat{X} = 90°$. La distance cartésienne entre deux de ces points vaut :

$$\sqrt{2} = 1.414213562°$$

alors qu'en géométrie sphérique elle vaut (cas du centre au pôle nord) :

$$2 \cdot \arcsin\sqrt{\sin^2\frac{89° - 89°}{2} + \sin^2\frac{90°}{2}\cos 89°\cos 89°} = 1.414177661°$$

On obtient donc une différence de 0.13 arcsec et une erreur sur la distance angulaire calculée en cartésien de 0.0025%
Dans le cas d'un champ de vue XMM de rayon $r = 15'$, la différence est de $2 \cdot 10^{-3}$ arcsec $= 2$ mas et l'erreur est de $1.6 \cdot 10^{-3}\%$. En pratique les boîtes de recherche de candidats sont au maximum de quelques minutes d'arc et les distances toujours calculées de façon exacte par la fonction d'Haversine.
Les ellipses à 1σ des gausiennes 2D d'erreurs sur les positions ont toujours un paramètre inférieur à la minute d'arc. On peut donc négliger les effets de la géométrie sphérique sur le produit de convolution des gausiennes d'erreur.

b Erreurs sur les surfaces

En cartésien, la surface d'un disque de rayon $r = 1°$ vaut :

$$\pi r^2 = 9.5698 \cdot 10^{-4} \text{ sr} = 11309.73 \text{ arcmin}^2$$

En sphérique on s'intéresse à une surface sur une sphère unité. La surface élémentaire $r^2 \sin\theta d\theta d\phi$ vaut donc ici $\sin\theta d\theta d\phi$. Si on place le centre de notre disque plan au pôle nord, on obtient la surface sphérique en intégrant :

$$\int_{\theta=0°}^{\theta=1°}\int_{\phi=0}^{\phi=2\pi}\sin\theta d\theta d\phi$$

Ce qui donne :

$$2\pi(1 - cos(1°)) = 9.5696 \cdot 10^{-4} \text{ sr} = 11309.45 \text{ arcmin}^2$$

L'erreur faite sur le calcul de la surface sphérique en cartesien est faible, moins de 0.0025%.
En pratique on s'intéresse à des surfaces encore plus faibles : maximum 1000 arcmin^2. On négligera donc la géometrie sphérique pour les calculs de densités locales, souvent limités à une centaine de sources.

3.2 CONVOLUTION DES ELLIPSES D'ERREUR ET DISTANCE EN σ

La probabilité que deux sources ont de se trouver exactement à la même position est donnée par le produit de convolution des gaussiennes d'erreur 2D sur leurs positions. En effet, on somme sur tous les points de l'espace la probabilité que la première source se trouve en ce point par la probabilité que la deuxième source s'y trouve également.

3.2.1 Calcul sur les ellipses

On note ell_1 l'ellipse d'erreur associée à la source X et ell_2 celle associée à la source candidate. σ_M et σ_m correspondent respectivement aux axes majeurs et mineurs. Les angles précedents θ_x et θ_O seront respectivement noté θ_1 et θ_2.

La probabilité que la source X ait une position (x, y) dans notre nouveau repère est donnée par la Gaussienne 2D suivante :

$$P_1(x,y) = \frac{1}{2\pi\sigma_{M_1}\sigma_{m_1}} e^{-\frac{1}{2}\left(\frac{x_1^2}{\sigma_{M_1}^2}+\frac{y_1^2}{\sigma_{m_1}^2}\right)}$$

Où x_1 et y_1 sont obtenus par le changement de repère :

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} +\cos\theta_1 & -\sin\theta_1 \\ +\sin\theta_1 & +\cos\theta_1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

De même, la probabilité que la source 2 ait une position (x,y) dans notre nouveau repère est donnée par la Gaussienne 2D suivante :

$$P_2(x,y) = \frac{1}{2\pi\sigma_{M_2}\sigma_{m_2}} e^{-\frac{1}{2}\left(\frac{x_2^2}{\sigma_{M_2}^2}+\frac{y_2^2}{\sigma_{m_2}^2}\right)}$$

Avec le changement de repère :

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} +\cos\theta_2 & -\sin\theta_2 \\ +\sin\theta_2 & +\cos\theta_2 \end{pmatrix} \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix}$$

Où (x_0, y_0) est la position du centre de l'ellipse 2 dans le nouveau repère, soit $(n, 0)$.

La probabilité que les deux sources dont les positions sont représentées par les deux Gaussiennes 2D soient exactement au même endroit est donnée par le produit de convolution des deux Gaussiennes 2D en $(x_0 = a, y_0 = 0)$ (dans ce cas pour passer des coordonnées (x, y) à (x_2, y_2), ne pas faire intervenir la soustraction avec (x_0, y_0)).

$$P_{SP}(x_0,y_0) = (P_1(x,y) * P_2(x,y))(x_0,y_0) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} P_1(x,y)\cdot P_2(x_0-x, y_0-y)dxdy$$

Dans le cas simplifié où les deux ellipses sont symétriques ($\sigma_{M_1} = \sigma_{m_1}$ et $\sigma_{M_2} = \sigma_{m_2}$) , les angles θ_x et θ_o n'ont pas d'importance et on peut les choisir tel que $\theta_1 = \theta_2 = 0°$, le calcul direct dans ce cas donne :

$$P_{SP}(x_0,y_0) = \frac{1}{2\pi\sigma_{12}^2} e^{-\frac{1}{2}\left(\frac{x_0^2}{\sigma_{12}^2}+\frac{y_0^2}{\sigma_{12}^2}\right)}$$

Avec :

$$\sigma_{12} = \sqrt{\sigma_1^2 + \sigma_2^2}$$

La probabilité finale est donnée par :

$$P_{SP}(a,0) = \frac{1}{2\pi(\sigma_1^2+\sigma_2^2)} e^{-\frac{1}{2}\left(\frac{a^2}{\sigma_1^2+\sigma_2^2}\right)}$$

On trouvera une formule plus générale dans l'article de Condon *et al.* (1995) :

$$\boxed{P_{SP}(x,y) = \frac{1}{2\pi\sigma_M\sigma_m} e^{-\frac{1}{2}\left(\frac{x_3^2}{\sigma_M^2}+\frac{y_3^2}{\sigma_m^2}\right)}} \tag{II.23}$$

Avec :

$$\begin{aligned}
\xi &= \frac{1}{2}\arctan\left[\frac{(\sigma_{M_2}^2-\sigma_{m_2}^2)\sin(2(\theta_2-\theta_1))}{(\sigma_{M_1}^2-\sigma_{m_1}^2)+(\sigma_{M_2}^2-\sigma_{m_2}^2)\cos(2(\theta_2-\theta_1))}\right] \\
\sigma_M^2 &= \frac{1}{2}\left[\frac{(\sigma_{M_2}^2-\sigma_{m_2}^2)\cos(\theta_2-\theta_1)\sin(\theta_2-\theta_1)}{\cos\xi\sin\xi}+\sigma_{M_1}^2+\sigma_{m_1}^2+\sigma_{M_2}^2+\sigma_{m_2}^2\right] \\
\sigma_m^2 &= \sigma_{M_1}^2+\sigma_{m_1}^2+\sigma_{M_2}^2+\sigma_{m_2}^2-\sigma_M^2
\end{aligned}$$

Comme ξ est l'angle entre l'axe principal de l'ellipse 1 et celui de l'ellipse résultant du produit de convolution, la matrice de changement de repère s'exprime de la façon suivante :

$$\begin{pmatrix} x_3 \\ y_3 \end{pmatrix} = \begin{pmatrix} +\cos(\theta_1+\xi) & -\sin(\theta_1+\xi) \\ +\sin(\theta_1+\xi) & +\cos(\theta_1+\xi) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

3.2.2 Ellipse, matrice de variance-covariance et changement de repère

Les erreurs de positionnement en astronomie sont presque toujours données sous la forme d'une ellipse de semi grand axe a, de semi petit axe b et d'angle θ entre le pôle Nord et le semi grand axe se trouvant dans la partie Est, orienté positivement du Nord vers l'Est. On a donc $\theta \in [0, 180[$.

a Passage d'une ellipse d'erreur à une matrice de variance-covariance

Dans le repère pour lequel l'ellipse a un angle nul, elle est sous sa forme canonique : l'ellipse équivaut à une matrice de variance-covariance diagonale (covariance nulle) portant sur sa diagonale le demi-grand axe a et le demi-petit axe b. Si l'ellipse est centrée en X, on passe du repère où l'ellipse est sous sa forme canonique au repère où l'axe des x est l'axe XN' par une rotation d'angle $-\theta$. On peut déduire la matrice V de variance-covariance de l'ellipse à partir de la matrice diagonale D.
Soit U la matrice de rotation d'angle θ :

$$U = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{II.24}$$

La rotation d'angle θ d'une matrice R s'exprime :

$$R' = U^{-1}RU = {}^tURU \tag{II.25}$$

Remarque : $U^{-1} = {}^tU$ car la matrice d'une (base orthonormée, ou 'BON', est unitaire ($U^tU = I_n$ avec I_n la matrice identité),sa rotation donne une nouvelle BON donc la matrice résultante reste unitaire. Par définition de la matrice inverse $UU^{-1} = I_n$, comme U unitaire, $U^tU = I_n$ et donc $U^{-1} = {}^tU$).
Dans notre cas, la rotation est d'angle $-\theta$, on a donc :

$$V = UD^tU$$

Le calcul direct nous donne le résultat voulu :

$$\begin{aligned}
V = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} &= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \\
&= \begin{pmatrix} a\cos^2\theta + b\sin^2\theta & (a-b)\cos\theta\sin\theta \\ (a-b)\cos\theta\sin\theta & a\sin^2\theta + b\cos^2\theta \end{pmatrix}
\end{aligned}$$

D'où :

$$\begin{aligned} \sigma_x^2 &= a\cos^2\theta + b\sin^2\theta && \text{(II.26)} \\ \sigma_y^2 &= a\sin^2\theta + b\cos^2\theta && \text{(II.27)} \\ \rho\sigma_x\sigma_x &= (a-b)\cos\theta\sin\theta && \text{(II.28)} \end{aligned}$$

b Changement de repère d'une matrice de variance-covariance

Soit la matrice de variance-covariance d'une loi de probabilité normale en dimension 2 $f(x,y)$ (voir Pelat 1998, p. 92). Cette matrice s'écrit sous forme générale :

$$V = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \quad \text{(II.29)}$$

On a pour variances :

$$\begin{aligned} \text{Var}(x) &= \sigma_x^2 \\ \text{Var}(y) &= \sigma_y^2 \end{aligned}$$

Et pour covariance :

$$\text{Cov}(x,y) = \rho\sigma_x\sigma_y$$

On fait subir au repère (O,x,y) une rotation d'angle θ. On cherche la matrice de variance-covariance dans ce nouveau repére. Pour cela, j'expose deux solutions différentes. La première utilise les formules de changement de repère et les propriétés de la variance. La deuxième, la plus immédiate, la matrice de rotation.

Propriétes de la variance On rappelle quelques propriétés sur l'espérance, la variance et la covariance. Soit a un réel :

$$\begin{aligned} \text{E}(ax) &= a\text{E}(x) \\ \text{Var}(ax) &= a^2\text{Var}(x) \\ \text{Var}(x+y) &= \text{Var}(x) + \text{Var}(y) + 2\text{Cov}(x,y) \\ \text{Cov}(x,y) &= \text{E}(xy) - \text{E}(x)\text{E}(y) = \text{E}[(x-\text{E}(x))(y-\text{E}(y))] \end{aligned}$$

Faisons subir au repère (O,x,y) une rotation d'angle θ. Les formules de changements de repères donnent :

$$\begin{aligned} X &= x\cos\theta + y\sin\theta \\ Y &= -x\sin\theta + y\cos\theta \end{aligned}$$

D'où :

$$\text{Var}(X) = \text{Var}(x\cos\theta + y\sin\theta) = \text{Var}(x)\cos^2\theta + \text{Var}(y)\sin^2\theta + 2\text{Cov}(x\cos\theta, y\sin\theta)$$

Or :

$$\text{Cov}(x\cos\theta, y\sin\theta) = \cos\theta\sin\theta(\text{E}(xy) - \text{E}(x)\text{E}(y)) = \cos\theta\sin\theta\text{Cov}(x,y)$$

D'où au final :

$$\boxed{\sigma_X^2 = \sigma_x^2\cos^2\theta + \sigma_y^2\sin^2\theta + 2\cos\theta\sin\theta\rho\sigma_x\sigma_y} \quad \text{(II.30)}$$

De même, on trouve :

$$\boxed{\sigma_Y^2 = \sigma_x^2\sin^2\theta + \sigma_y^2\cos^2\theta - 2\cos\theta\sin\theta\rho\sigma_x\sigma_y} \quad \text{(II.31)}$$

Enfin on s'intéresse à la covariance du couple (X,Y). On a :

$$\text{Cov}(X,Y) = \frac{1}{2}\left(\text{Var}(X+Y) - \text{Var}(X) - \text{Var}(Y)\right)$$

Or :

$$X + Y = x(\cos\theta - \sin\theta) + y(\cos\theta + \sin\theta)$$

D'où :

$$\mathrm{Cov}(X,Y) = \frac{1}{2}\left((\cos\theta - \sin\theta)^2\mathrm{Var}(x) + (\cos\theta + \sin\theta)^2\mathrm{Var}(y) + 2(\cos^2\theta - \sin^2\theta)\mathrm{Cov}(x,y) - \mathrm{Var}(X) - \mathrm{Var}(Y)\right)$$

On développe, et on trouve au final :

$$\boxed{\rho\sigma_X\sigma_Y = \cos\theta\sin\theta(\sigma_y^2 - \sigma_x^2) + (\cos^2\theta - \sin^2\theta)\rho\sigma_x\sigma_y} \quad \text{(II.32)}$$

Matrice de rotation Soit U la matrice de rotation d'angle θ. Cette matrice est la même que celle de l'équation II.24.
Soit R la matrice de variance covariance dans la nouvelle base, on a la relation, similaire à l'Eq. II.25 :

$$R = U^{-1}VU = U^tVU$$

On pose le calcul :

$$R = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

Ce qui nous donne :

$$R = \begin{pmatrix} \sigma_x^2\cos^2\theta + \sigma_y^2\sin^2\theta + 2\cos\theta\sin\theta\rho\sigma_x\sigma_y & \cos\theta\sin\theta(\sigma_y^2 - \sigma_x^2) + (\cos^2\theta - \sin^2\theta)\rho\sigma_x\sigma_y \\ \cos\theta\sin\theta(\sigma_y^2 - \sigma_x^2) + (\cos^2\theta - \sin^2\theta)\rho\sigma_x\sigma_y & \sigma_x^2\sin^2\theta + \sigma_y^2\cos^2\theta - 2\cos\theta\sin\theta\rho\sigma_x\sigma_y \end{pmatrix}$$

On retrouve bien les formules II.30, II.31 et II.32.

c Passage d'une matrice de variance-covariance à une ellipse d'erreur

On dispose ici d'une matrice de variance-covariance V et on voudrait en déduire le semi grand axe, le semi petit axe et l'angle de rotation par rapport à l'axe des x du repère. On sait que sous sa forme canonique la matrice de variance-covariance est une matrice diagonale contenant les valeurs du demi grand axe et du demi petit axe. Il suffit donc[1] de diagonaliser la matrice V. Ce qui donne après calcul :

$$\boxed{\sigma_{M,m}^2 = \frac{1}{2}\left(\sigma_x^2 + \sigma_y^2 \pm \sqrt{(\sigma_x^2 - \sigma_y^2)^2 + 4\rho^2\sigma_x^2\sigma_y^2}\right)} \quad \text{(II.33)}$$

Remarque : ces formules présentent l'avantage de ne pas comporter d'opérations sur les sinus et cosinus, opérations qui sont coûteuses en temps CPU

L'angle de rotation peut être calculé à partir de l'équation II.32. En effet, si la matrice est diagonale, la covariance est nulle. On a donc :

$$\cos\theta\sin\theta(\sigma_x^2 - \sigma_y^2) = (\cos^2\theta - \sin^2\theta)\rho\sigma_x\sigma_y$$

En se souvenant des deux formules de trigonométrie :

$$\begin{aligned} \sin 2\theta &= 2\sin\theta\cos\theta \\ \cos 2\theta &= \cos^2\theta - \sin^2\theta \end{aligned}$$

on obtient :

$$\boxed{\theta = \frac{1}{2}\arctan\frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}} \quad \text{(II.34)}$$

[1] c.f. p. 95 (6.2.4 Forme Quadratique Associée) du cours « *Base et Méthode pour le Traitement des Données (Bruit et Signaux)* » de Didier Pelat

3.2.3 Approche matricielle

Soient deux variables aléatoires Gaussiennes centrées Z_1 et Z_2 en dimension 2 . Z_1 et Z_2 sont indépendantes. La variable aléatoire :

$$Z = Z_1 + Z_2 \tag{II.35}$$

est définie par le produit de convolution de Z_1 et Z_2. Le résultat est une Gaussienne dont on veut la matrice de variance-covariance.
Soit V_{Z_1} la matrice de variance-covariance de Z_1 et V_{Z_2} celle de Z_2 :

$$\begin{aligned} V_{Z_1} &= \begin{pmatrix} \sigma^2_{x_1} & \rho_1\sigma_{x_1}\sigma_{y_1} \\ \rho_1\sigma_{x_1}\sigma_{x_1} & \sigma^2_{y_1} \end{pmatrix} \\ V_{Z_2} &= \begin{pmatrix} \sigma^2_{x_2} & \rho_2\sigma_{x_2}\sigma_{y_2} \\ \rho_2\sigma_{x_2}\sigma_{y_2} & \sigma^2_{y_2} \end{pmatrix} \end{aligned}$$

On note $\sigma_{x,y}$ la covariance entre x et y.
La matrice de variance-covariance de Z est définie par :

$$V = \begin{pmatrix} \sigma^2_{x_1+x_2} & \sigma_{x_1+x_2,y_1+y_2} \\ \sigma_{x_1+x_2,y_1+y_2} & \sigma^2_{y_1+y_2} \end{pmatrix}$$

Or

$$Var(x_1 + x_2) = Var(x_1) + Var(x_2) + 2Cov(x_1, x_2)$$

Mais Z_1 et Z_2 et donc x_1 et x_2 sont indépendantes. Leur covariance est donc nulle. D'où :

$$\sigma^2_{x_1+x_2} = \sigma^2_{x_1} + \sigma^2_{x_2}$$

De même :

$$\sigma^2_{y_1+y_2} = \sigma^2_{y_1} + \sigma^2_{y_2}$$

Enfin, la covariance :

$$\sigma_{x_1+x_2,y_1+y_2} = E\{(x_1 + x_2)(y_1 + y_2)\} - E\{(x_1 + x_2)\}E\{(y_1 + y_2)\}$$

Or l'espérance de x_1, y_1 , x_2, y_2 est nulle (Gaussiennes centrées). Il en est donc de même pour $E\{(x_1 + x_2)\}$ et $E\{(y_1 + y_2)\}$. D'où :

$$\begin{aligned} \sigma_{x_1+x_2,y_1+y_2} &= E(x_1 + x_2)(y_1 + y_2) \\ &= E(x_1y_1) + E(x_1y_2) + E(x_2y_1) + E(x_2y_2) \end{aligned}$$

Or x_1 et y_2 sont indépendantes. De même pour x_2 et y_1. Donc leurs covariances (espérances de leur produits) sont nulles. D'où au final :

$$\sigma_{x_1+x_2,y_1+y_2} = \sigma_{x_1,y_1} + \sigma_{x_2,y_2}$$

On trouve donc que la matrice de variance-covariance de Z est la somme des matrices de variances-covariances de Z_1 et Z_2 :

$$\boxed{V_Z = \begin{pmatrix} \sigma^2_{x_1} + \sigma^2_{x_2} & \rho_1\sigma_{x_1}\sigma_{y_1} + \rho_2\sigma_{x_2}\sigma_{y_2} \\ \rho_1\sigma_{x_1}\sigma_{y_1} + \rho_2\sigma_{x_2}\sigma_{y_2} & \sigma^2_{y_1} + \sigma^2_{y_2} \end{pmatrix}} \tag{II.36}$$

3.2.4 Distance en σ de deux sources

La distance en σ entre deux sources est donnée par le rapport entre la distance angulaire séparant ces deux sources et la distance entre le centre de l'ellipse de convolution et son périmètre à 1σ.
Cette dernière distance varie selon le point du périmètre considéré et donc selon l'angle de l'ellipse par rapport au segment joignant les deux sources.

Si le centre de l'ellipse de convolution est placé sur la position de la source X et qu'on se place dans le référentiel où l'axe des x est défini par la distance angulaire entre la source X et la source d'archive, la distance en sigma vaut directement :

$$d_{1\sigma} = \frac{a}{\sigma_{x_c}\sqrt{1-\rho_c^2}}$$

où a est la distance angulaire, σ_{x_c}, σ_{y_c} et ρ_c les paramètres de l'ellipse.

De façon plus générale, la distance d'un point (x, y) à l'ellipse de matrice de variance-covariance V vaut :

$$\boxed{d_{1\sigma} = \frac{1}{\sqrt{(1-\rho)^2}}\left(\frac{x^2}{\sigma_{x_c}^2} - \frac{2\rho xy}{\sigma_{x_c}\sigma_{y_c}} + \frac{y^2}{\sigma_{y_c}^2}\right)^{1/2}} \quad \text{(II.37)}$$

3.3 Origines du rapport de vraisemblance et probabilité d'identification

3.3.1 Origines du rapport de vraisemblance

On considère la cas où une source X a seulement un candidat optique et on formule deux hypothèses :

H_1 **:** la source candidate est la contrepartie, elle n'est pas là par hasard

H_0 **:** la source candidate est une source aléatoire provenant d'une distribution Poissonienne de source de densité λ

L'union de ces deux hypothèses forme l'ensemble des posibilités : $P(H_1) + P(H_0) = 1$, i.e. $H_1 = \bar{H_0}$. La probabilité de trouver un candidat à une distance r de la source X peut donc être écrit :

$$P(r) = P(r|H_1) \cdot P(H_1) + P(r|H_0) \cdot P(H_0) \quad \text{(II.38)}$$

A partir des probabilités conditionnelles :

$$\begin{aligned} P(H_1|r) &= \frac{P(H_1 \cap r)}{P(r)} \\ P(r|H_1) &= \frac{P(H_1 \cap r)}{P(H_1)} \end{aligned}$$

On déduit :

$$P(H_1|r) = \frac{P(r|H_1) \cdot P(H_1)}{P(r)}$$

Qui peut être réécrit (voir Eq. II.38) :

$$P(H_1|r) = \frac{P(r|H_1) \cdot P(H_1)}{P(r|H_1) \cdot P(H_1) + P(r|H_0) \cdot P(H_0)}$$

On transforme cette équation en :

$$P(H_1|r) = \frac{1}{1 + \frac{1}{\frac{P(H_1)}{P(H_0)} LR}} \quad \text{(II.39)}$$

Pour faire apparaître le rapport de vraisemblance :

$$LR(r) = \frac{P(r|H_1)}{P(r|H_0)} \tag{II.40}$$

Le terme restant, $\frac{P(H_1)}{P(H_0)}$, doit être estimé pour calculer $P(H_1|r)$, la probabilité que la candidat soit la vraie contrepartie. Si on considère l'estimation du nombre de contreparties réelles divisé par le nombre de fausses associations, on obtient le facteur $\theta/(1-\theta)$ de De Ruiter *et al.* (1977).

Si on ajoute une connaissance a priori basée sur la magnitude, la formule II.39 est inchangée sauf qu'elle exprime maintenant $P(H_1|r \cap m)$ et le LR devient :

$$LR(r \cap m) = \frac{P(r \cap m|H_1)}{P(r \cap m|H_0)} = \frac{P(r|H_1)P(m|H_1)}{P(r|H_0)P(m|H_0)} \tag{II.41}$$

Le terme $P(r \cap m|H_0) = P(r|H_0)P(m|H_1)$ exprime la probabilité d'avoir une source aléatoire de magnitude m. C'est la densité Poissonienne locale de sources de magnitude m. Le terme $P(r|H_1)P(m|H_1)$ est la distribution Gaussienne 2D classique fois la probabilité d'avoir, parmi toutes les contreparties, une contrepartie de magnitude m. On peut le voir comme la distribution de sources de contreparties en fonction de leurs magnitudes et il correspond au facteur $q(m)$ dans Brusa *et al.* (2007) et $q(m, c)$ dans Sutherland & Saunders (1992), si on ignore les types d'objets.

3.3.2 Généralisation des probabilité Bayesiennes

Comme précédemment, on suppose qu'on a M candidats et que au plus un correspond à la vraie contrepartie. On remplace les deux hypothèses H_1 et H_0 par les $M+1$ hypothèses :

H_i **:** le candidat i est la vraie contrepartie, $i \in [1, M]$

H_0 **:** aucun des candidats est la vraie contrepartie λ

L'union des hypothèses vaut toujours 1 : $\sum_{i=1}^{M} H_i + H_0 = 1$

On note $\cap r_i = r_1 \cap r_2 \cap ... \cap r_M$. La probabilité de trouver le candidat 1 à la distance r_1 et le candidat 2 à la distance r_2, ..., s'écrit :

$$P(\cap r_i) = \sum_{i=1}^{M} P(\cap r_i|H_i)P(H_i) + P(\cap r_i|H_0)P(H_0) = \sum_{i=0}^{M} P(\cap r_i|H_i)P(H_i)$$

On a toujours les probabilités conditionnelles :

$$\begin{aligned} P(H_i| \cap r_i) &= \frac{P(H_i \cap \cap r_i)}{P(\cap r_i)} \\ P(\cap r_i|H_i) &= \frac{P(H_i \cap \cap r_i)}{P(H_i)} \end{aligned}$$

D'où :

$$\begin{aligned} P(H_i| \cap r_i) &= \frac{P(\cap r_i|H_i)P(H_i)}{P(\cap r_i)} \\ &= \frac{P(\cap r_i|H_i)P(H_i)}{\sum_{i=1}^{M} P(\cap r_i|H_i)P(H_i) + P(\cap r_i|H_0)P(H_0)} \end{aligned}$$

Si on divise en haut et en bas par $P(\cap r_i|H_0)P(H_0)$, on obtient :

$$P(H_i| \cap r_i) = \frac{\frac{P(\cap r_i|H_i)P(H_i)}{P(\cap r_i|H_0)P(H_0)}}{1 + \sum_{j=1}^{M} \frac{P(\cap r_j|H_{cp_j})P(H_{cp_j})}{P(\cap r_j|H_0)P(H_0)}}$$

Les distances sont indépendantes les unes des autres. On peut donc écrire :

$$\begin{aligned} P(\cap r_i|H_i) &= \prod_{j=1}^{M} P(r_j|H_{cp_j}) = Gauss_i \prod_{j=1;j\neq i}^{M} Poisson_j \\ P(\cap r_i|H_0) &= \prod_{j=1}^{M} P(r_j|H_0) = \prod_{j=1}^{M} Poisson_j \end{aligned}$$

Ce qui permet finalement d'écrire :

$$\boxed{P(H_i|\cap r_i) = \frac{LR_i \frac{P(H_i)}{P(H_0)}}{1 + \sum_{j=1}^{M} LR_j \frac{P(H_j)}{P(H_0)}}} \quad \text{(II.42)}$$

3.3.3 Probabilité d'identification de Rutledge

La fiabilité donne la probabilité pour qu'un couple source/contrepartie ayant une valeur de LR donnée soit une association juste, non due au hasard. On trouvera dans l'article de Rutledge *et al.* (2000) comment traiter les cas où les associations sont multiples.
Ce résultat vient de la généralisation des probabilités Bayesiennes abordée dans le paragraphe précédent. On simplifie le problème en supposant que sur M candidats au plus un est la vraie contrepartie. On se retrouve alors avec $M+1$ solutions :

- M solutions : le candidat i est la vraie contrepartie, $i \in [1, M]$
- 1 solution : aucun des candidats est la vraie contrepartie

Nous reprenons ici les notations introduites par Rutledge *et al.* (2000). La somme des probabilités de ces $M+1$ hypothèses vaut 1 :

$$P_{no_id} + \sum_{i=1}^{M} P_{id,i} = 1$$

On a donc, avec Rj la fiabilité de l'identification du candidat j :

$$P_{no_id} = \frac{\prod_{j=1}^{M}(1-R_j)}{S}$$

$$P_{id,i} = \frac{R_i \prod_{j\neq i}(1-R_j)}{S} = \frac{\frac{R_i}{1-R_i}\prod_{j=1}^{M}(1-R_j)}{S}$$

Avec :

$$S = \prod_{j=1}^{M}(1-R_j) + \sum_{i=1}^{M} \frac{R_i}{1-R_i} \prod_{j=1}^{M}(1-R_j)$$

D'où au final la grandeur qui nous intéresse :

$$P_{id,i} = \frac{\frac{R_i}{1-R_i}\prod_{j=1}^{M}(1-R_j)}{\prod_{j=1}^{M}(1-R_j) + \sum_{i=1}^{M}\frac{R_i}{1-R_i}\prod_{j=1}^{M}(1-R_j)}$$

Qui peut se simplifier en :

$$\boxed{P_{id,i} = \frac{\frac{R_i}{1-R_i}}{1 + \sum_{i=1}^{M}\frac{R_i}{1-R_i}}} \quad \text{(II.43)}$$

Voir la première partie de ce chapitre pour faire le lien entre cette formule (Eq. II.43) et la formule précedente (Eq. II.42).

3.4 Estimation du paramètre de la loi de Poisson par k-NN

On estime la densité locale en faisant l'hypothèse qu'elle suit une loi de Poisson de paramètre λ. On estime λ par la méthode du maximum de vraisemblance. On suppose que dans une surface S autour d'une source X on trouve N sources dans le catalogue des contreparties qui sont au moins aussi brillantes que la source candidate. L'estimation suivante est valide pour $N > 10$, valeur à partir de laquelle on peut commencer à approximer la loi de Poisson par une loi normale. En pratique on augmente la surface S jusqu'à trouver un nombre minimum fixé de sources.
Pour une loi de Poisson, la fonction de vraisemblance est donnée par la formule :

$$L(\lambda) = \frac{\lambda S^N}{N!} e^{-\lambda S}$$

On cherche le maximum de cette fonction. La fonction $ln(x)$ étant strictement monotone sur le domaine de définition de $L(\lambda)$, on peut rechercher le maximum de $L(\lambda)$ en cherchant le maximum de son logarithme $\ln(L(\lambda))$:

$$\ln\left(L(\lambda)\right) = \ln\left(\frac{\lambda S^N}{N!} e^{-\lambda S}\right) = -\lambda S + N\ln(\lambda S) - \ln(N!)$$

Pour trouver le maximum, on étudie la dérivée qui s'annule ici en un seul point :

$$\frac{\partial \ln\left(L(\lambda)\right)}{\partial \lambda} = S + \frac{N \cdot S}{\lambda S} = 0$$

On s'assure que ce point est bien un maximum et on en déduit finalement que :

$$\boxed{\lambda = \frac{N}{S}} \tag{II.44}$$

On estime de même la variance à l'aide du maximum de vraisemblance :

$$Var(\hat{\lambda}) = -\left(\frac{\partial^2 \ln\left(L(\lambda)\right)}{\partial^2 \lambda}|_{\hat{\lambda}=\lambda}\right)^{-1}$$

Or :

$$\frac{\partial^2 \ln\left(L(\lambda)\right)}{\partial^2 \lambda} = \frac{\partial}{\partial \lambda}\left(\frac{N}{\lambda}\right) = \frac{-N}{\lambda^2} \tag{II.45}$$

On déduit l'expression finale des formules (II.45) et (II.44) :

$$\boxed{\sigma^2 = Var(\hat{\lambda}) = \frac{N}{S^2}}$$

On estime donc le paramètre de la loi de Poisson représentant notre densité locale par par une loi normale :

$$\boxed{\lambda \rightsquigarrow \mathcal{N}(\frac{N}{S}; \frac{\sqrt{N}}{S})} \tag{II.46}$$

4 Recalibration des positions des sources XMM

Nous abordons dans cette partie une méthode générale qui permet une première recalibration des positions des sources d'un champ de vue à partir d'un catalogue de référence. Nous proposons ensuite une méthode plus fine.

4.1 MÉTHODE GÉNÉRALE DE RECALIBRATION DE POSITIONS

4.1.1 Le problème

On cherche à recalibrer les positions des sources issues d'une observation en fonction d'un catalogue de référence. On commence par projeter les sources sur un plan euclidien tangent, au point central de l'observation, à la sphère unité. On prend en compte trois biais qui consistent uniquement en des opérations linéaires. On ne tient pas compte ici de distorsions d'ordre supérieur. Ces trois biais sont :

- une échelle différente (homothétie) : les distances ne sont pas évaluées correctement dans notre observation, il y a un rapport s (pour *scale*) entre les distances des sources de notre observation et celles du catalogue de référence.
- un erreur d'orientation (rotation) : l'axe Nord-Sud de notre observation est décalé d'un angle θ par rapport à l'axe Nord-Sud réel
- un décalage systématique (translation) : les positions des sources de l'observation, sur le plan de projection, présentent un décalage systématique (dx, dy) par rapport à celles des sources du catalogue de référence

La solution proposée ici est basée sur celle de Kaiser *et al.* (1999) avec cependant quelques différences : nous n'utilisons pas le même système de projection et, contrairement à Kaiser *et al.* (1999), nous prenons en compte et propageons les erreurs à chaque étape du calcul.

4.1.2 Le système

On appelle ici « paire » un couple de deux sources distinctes venant d'un même catalogue ou liste de sources. On appelle $\{N_o\}$ l'ensemble des points $(\log(r), \alpha)$ issu des paires de notre observation, où r est la distance qui sépare les deux sources de la paire, et α l'angle entre l'axe des x et la droite passant par les deux sources. A chaque paire peuvent être associés deux angles : α et $\alpha - \pi$. On choisit, par convention, de ne garder que les angles $\alpha \in [0 - \pi]$. Si notre observation contient n sources, on obtient ainsi un nombre de paires $n_p = \sum_{i=1}^{n-1} k = \frac{(n-1)n}{2}$. De même, on appelle $\{N_{ref}\}$ l'ensemble des points $(\log(r), \alpha)$ issu des paires du catalogue de référence. (Remarque : si on connaît les valeurs limites des trois biais – l'échelle, l'angle et l'offset – on peut se contenter d'extraire dans le catalogue de référence les sources dans un rayon variable autour des sources de notre observation).

On identifie ensuite toutes les associations possibles entre $\{N_o\}$ et $\{N_{ref}\}$. Là encore, si on connaît les valeurs limites des trois biais, on peut contraindre les distances de corrélation. On considère donc pour chaque point de $\{N_o\}$ les points de $\{N_{ref}\}$ situés dans une boîte définie par les limites maximales de la valeur d'échelle s et de l'angle α.

Le résultat de cette corrélation croisée est un ensemble de points $\{N_c\}$ dans l'espace $(\Delta \log(r), \Delta\alpha)$. Le maximum de cette distribution nous donne l'erreur d'orientation $\theta = (\Delta\alpha)_{max}$ et le logarithme du facteur d'échelle $\log(s) = (\Delta \log(r))_{max}$.
Pour trouver ce maximum, il suffit de faire, par exemple, un *Mean Shift* où l'on associe à chaque point un noyau Gaussien dont la largeur de bande correspond à la convolution des ellipses d'erreur sur les points $(\log(r), \alpha)$. Cette technique est une montée de gradient sur une somme de Gaussiennes. Elle est par conséquent équivalente à un Levenberg-Marquart qui est une technique de descente de gradient sur le maximum de vraisemblance de cette même distribution.

En pratique, même en optimisant les zones de recherche, le nombre de points dans $\{N_c\}$ peut être très élevé : plusieurs dizaines de millions et plus. Stocker tous ces points dépasse en géneral les limites de nos machines. On peut alors calculer une grille et mettre à jour au fur et à mesure la densité sur les points de la grille. Cette méthode est plus élégante qu'un histogramme 2D (ou image où chaque pixel est un bin 2D) du nombre de points comme proposé dans Kaiser *et al.* (1999).

Une fois s et θ obtenus, on peut corriger une première fois les positions des sources de l'observation. On fait alors une corrélation croisée classique entre les sources de l'observation, qui ont leurs nouvelles coordonnées, et les sources de référence. De même que pour s et θ, on cherche le maximum de la distribution pour obtenir le décalage (dx, dy). En pratique, on n'a pas besoin pour ce calcul de passer par une grille, le nombre de points étant plus faible que pour le cas précédent.

4.2 CORRECTION DU POSITIONNEMENT DES SOURCES XMM

La solution qu'on propose ici peut être utilisée pour améliorer la correction obtenue avec la méthode précédente. Elle est basée sur la minimisation d'un χ^2 par la méthode de Levenberg-Marquardt. On a généralisé l'algorithme disponible dans Press *et al.* (2007) afin de prendre en compte des erreurs 2D elliptiques et non 1D. Cette méthode peut également être appliquée de façon indépendente si il n'y a pas de biais d'échelle et si on a une idée de l'ordre de grandeur de la valeur de la rotation et du décalage systématique. Dans ce cas, on calcule les solutions à partir d'une grille de points dont la distance typique entre deux points vaut environ la moyenne de la variance des ellipses d'erreur convoluées. La meilleure solution est celle associée au plus petit χ^2.

4.2.1 Erreur de positionnement des sources XMM

Le centre et l'orientation d'une observation XMM sont définis avec une certaine incertitude. Les positions des sources XMM, définies à partir du centre d'une observation, peuvent subir les effets d'une rotation et d'une translation. La translation correspond au vecteur entre l'estimation et la vraie position du centre de l'observation. La rotation correspond à l'angle entre l'estimation de l'orientation des axes sud-nord et ouest-est et la vraie orientation de ces axes. On suppose ici qu'il n'y a pas de biais liée à une échelle différente. Autrement dit, le paramètre de l'homothétie vaut 1.

On se place dans le plan cartésien 2D du FOV d'une observation XMM. Soit, dans ce plan, (x_{obs}, y_{obs}) la position d'une source définie à partir du repère issu de la calibration de l'instrument. Soit (x_{true}, y_{true}) la vraie position de cette source.
Soit Δx et Δy les coordonnées du centre calibré dans le repère constitué du vrai centre et de la bonne orientation des axes. Δx et Δy repésentent donc les écarts entre les vraies coordonnées d'une source et ses coordonnées dans le « faux »repère (voir IF. II.9). En l'absence de rotation on a

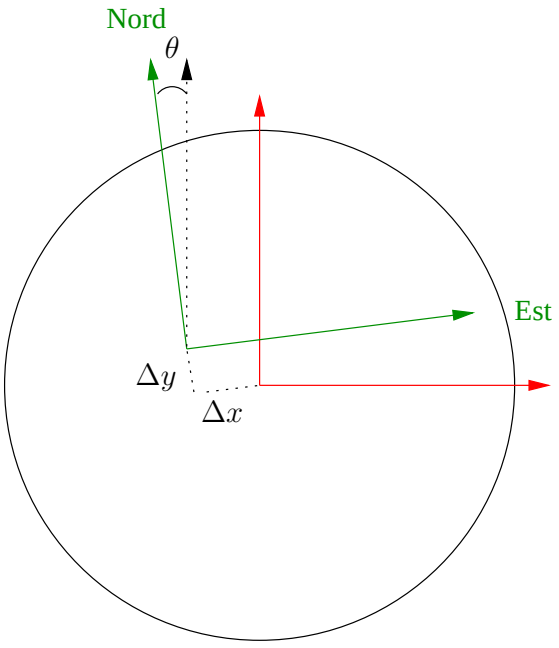


FIG. II.9 – Schéma représentant le système de correction du référentiel d'un FOV.

donc :

$$\Delta x = x_{true} - x_{obs} \quad \text{(II.47)}$$
$$\Delta y = y_{true} - y_{obs} \quad \text{(II.48)}$$

Soit θ l'angle entre l'orientation des vrais axes Sud-Nord Ouest-Est et ceux issus de la calibration. En l'absence d'*offset*, c'est-à-dire $\Delta x = \Delta y = 0$, on trouve le "faux" repère en faisant subir aux vrais axes une rotation d'angle θ. On a donc dans ce cadre :

$$\begin{pmatrix} x_{obs} \\ y_{obs} \end{pmatrix} = \begin{pmatrix} +\cos\theta & +\sin\theta \\ -\sin\theta & +\cos\theta \end{pmatrix} \begin{pmatrix} x_{true} \\ y_{true} \end{pmatrix} \tag{II.49}$$

La position observée se déduit donc de la vraie position et des paramètres θ, Δx et Δy par la formule exprimant une translation suivie d'une rotation :

$$\begin{pmatrix} x_{obs} \\ y_{obs} \end{pmatrix} = \begin{pmatrix} +\cos\theta & +\sin\theta \\ -\sin\theta & +\cos\theta \end{pmatrix} \begin{pmatrix} x_{true} - \Delta x \\ y_{true} - \Delta y \end{pmatrix} \tag{II.50}$$

On réécrit la formule pour obtenir l'expression des vraies coordonnées en fonction des fausses :

$$\begin{pmatrix} x_{true} \\ y_{true} \end{pmatrix} = \begin{pmatrix} +\cos\theta & -\sin\theta \\ +\sin\theta & +\cos\theta \end{pmatrix} \begin{pmatrix} x_{obs} \\ y_{obs} \end{pmatrix} + \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \tag{II.51}$$

Ce qui nous permet de définir deux fonctions :

$$\boxed{f_x(x_{obs}, y_{obs}, \theta, \Delta x, \Delta y) = x_{obs}\cos\theta - y_{obs}\sin\theta + \Delta x} \tag{II.52}$$

$$\boxed{f_y(x_{obs}, y_{obs}, \theta, \Delta x, \Delta y) = x_{obs}\sin\theta + y_{obs}\cos\theta + \Delta y} \tag{II.53}$$

On a un échantillon de N positions (x_{obs_i}, y_{obs_i}), $i \in [[0, N]]$. On cherche à déterminer les 3 paramètres θ, Δx et Δy.

4.2.2 Généralisation de la méthode de Levenberg-Marquardt

a Principe de base

La méthode de Levenberg-Marquardt[2] est basée sur une méthode des moindres carrés pondérés (ou méthode du moindre χ^2)[3]. Elle décrit une façon de trouver le maximum d'une probabilité issue du produit de N gaussiennes de dimension 1. Cette probabilité est le produit des probabilités d'obtenir un point de mesure à une certaine distance d'une courbe théorique, dépendant de un ou plusieurs paramètres (le vecteur $\mathbf{p}$). A chaque point de mesure est associée une variance différente.
On cherche donc à maximiser la probabilité :

$$P = \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(f(x_i, \mathbf{p}) - y_i)^2}{\sigma_i^2}\right) \tag{II.54}$$

Ce qui revient à minimiser l'opposé de son logarithme et donc la fonction, débarassée des termes constants, définie par :

$$\chi^2(\mathbf{p}) = \sum_{i=1}^{N} \frac{(y_i - f(x_i, \mathbf{p}))^2}{\sigma_i^2} \tag{II.55}$$

C'est cette fonction que l'on va généraliser. On gardera exactement le reste de la méthode (c.f. Numerical Recipes) pour trouver le minimum de notre nouvelle fonction.

b Généralisation

On généralise ici cette méthode pour des Gausiennes 2D dont les matrices de variance-covariance ne sont pas forcément diagonales. On peut appliquer cette méthode dans le cas d'XMM car les erreurs de position XMM sont circulaires, donc leur orientation ne change pas en fonction de θ.

[2] voir « Numerical Recipes in C », Sec. 15.5, p. 683

[3] Voir "Base et Méthodes pour le Traitements des Données", Didier Pelat, Sec. 16.2.1, p. 272

Nouveau χ^2 La fonction dont on cherche le minimum est alors définie par :

$$\chi^2(\mathbf{p}) = \sum_{i=1}^{N} \frac{(x_i - f_x(x_i, y_i; \mathbf{p}))^2}{\sigma_{x_i}^2} + \frac{(y_i - f_y(x_i, y_i; \mathbf{p}))^2}{\sigma_{y_i}^2} - \frac{2\rho_i (x_i - f_x(x_i, y_i; \mathbf{p}))(y_i - f_y(x_i, y_i; \mathbf{p}))}{\sigma_{x_i}^2 \sigma_{y_i}^2} \tag{II.56}$$

Soit en termes matriciels :

$$\chi^2(\mathbf{p}) = \sum_{i=1}^{N} (\mathbf{v}_i - \mathbf{f}(\mathbf{v}_i, \mathbf{p}))^t V^{-1} (\mathbf{v}_i - \mathbf{f}(\mathbf{v}_i, \mathbf{p})) \tag{II.57}$$

Calcul du nouveau Gradient Le gradient associé à cette nouvelle fonction χ^2 devient :

$$\frac{\partial \chi^2}{\partial p_k} = -2 \sum_{i=1}^{N} \frac{(x_i - f_x)}{\sigma_{x_i}^2} \frac{\partial f_x}{\partial p_k} + \frac{(y_i - f_y)}{\sigma_{y_i}^2} \frac{\partial f_y}{\partial p_k} - \frac{\rho_i}{\sigma_{x_i}^2 \sigma_{y_i}^2} \left(\frac{\partial f_x}{\partial p_k}(y_i - f_y) + \frac{\partial f_y}{\partial p_k}(x_i - f_x) \right) \tag{II.58}$$

Soit, après factorisation :

$$\frac{\partial \chi^2}{\partial p_k} = -2 \sum_{i=1}^{N} \left(\frac{x_i - f_x}{\sigma_{x_i}^2} - \frac{\rho_i (y_i - f_y)}{\sigma_{x_i}^2 \sigma_{y_i}^2} \right) \frac{\partial f_x}{\partial p_k} + \left(\frac{(y_i - f_y)}{\sigma_{y_i}^2} - \frac{\rho_i (x_i - f_x)}{\sigma_{x_i}^2 \sigma_{y_i}^2} \right) \frac{\partial f_y}{\partial p_k} \tag{II.59}$$

Jacobienne pour notre cas particulier A l'application $F : \mathbb{R}^3 \to \mathbb{R}^2$ dont les composantes sont $f_x(\theta, \Delta x, \Delta y)$ et $f_y(\theta, \Delta x, \Delta y)$, on associe la matrice Jacobienne :

$$J_F = \begin{pmatrix} \nabla f_x \\ \nabla f_y \end{pmatrix} = \begin{pmatrix} \frac{\partial f_x}{\partial \theta} & \frac{\partial f_x}{\partial \Delta x} & \frac{\partial f_x}{\partial \Delta y} \\ \frac{\partial f_y}{\partial \theta} & \frac{\partial f_y}{\partial \Delta x} & \frac{\partial f_y}{\partial \Delta y} \end{pmatrix} = \begin{pmatrix} -x \sin\theta - y \cos\theta & 1 & 0 \\ +x \cos\theta - y \sin\theta & 0 & 1 \end{pmatrix} \tag{II.60}$$

Implémentation On utilise comme algorithme de base le package Java de J. Holopainen[4]. On a introduit dans ce package la prise en compte des données d'entrée à deux dimensions. On a essentiellement modifié le calcul du χ^2 et des coefficients α et β, ces derniers faisant intervenir le nouveau gradient.

c Détermination des coordonnées corrigées

La position approximative du vrai centre par rapport au centre issu de la calibration se déduit par les formules :

$$\alpha_{c_t} = \alpha_{c_{obs}} - \frac{\Delta x}{\cos \delta_{c_{obs}}} \tag{II.61}$$

$$\delta_{c_t} = \delta_{c_{obs}} - \Delta y \tag{II.62}$$

Ces formules, approximatives, peuvent poser problème si on se trouve près des pôles.

On propose une autre méthode pour trouver la position exacte, aux erreurs numériques près, du vrai centre de l'observation.
Soit Q_1 le quaternion permettant de passer de l'axe 3D $\vec{X}$, défini par la coordonnée du point vernal $(1, 0, 0)$, au centre observé de l'observation $\vec{X_{Q_1}}$. Si $\theta = 0$, les axes 3D $\vec{Y_{Q_1}}$ et $\vec{Z_{Q_1}}$ correspondent respectivement aux axes $\vec{x}$ et $\vec{y}$ du repère local.

On transforme Δx et Δy en coordonnées polaires (ρ_c, θ_c). La transformation permettant de passer du centre mesuré au vrai centre est une rotation d'angle $-\rho_c$ et d'axe défini par le produit vectoriel de $\vec{X_{Q_1}}$ avec $\vec{Y'_{Q_1}}$. $\vec{Y'}$ étant l'axe obtenu par rotation de $-\theta^\circ$ de l'axe $\vec{Y}$ autour de $\vec{X}$. On associe à cette transformation un quaternion Q_2. L'image du vecteur $\vec{X_{Q_1}}$ par le quaternion (voir annexe A) Q_2 donne la position du vrai centre de l'observation, définie par le vecteur $\vec{X_{Q_1 Q_2}}$.

On utilise un processus similaire pour retrouver les positions corrigées de toutes les sources en fonction de leurs coordonnées (x, y) dans le bon repère.

[4] http://users.utu.fi/jaolho/

5 Estimation de densités locales

On s'intéresse ici aux densités locales de sources sur le ciel. Plus exactement, on s'intéresse aux densités locales de sources projetées sur un plan euclidien tangent à la sphère céleste. On se concentre donc sur le cas 2D mais toutes les méthodes abordées dans la suite sont généralisables en dimension n, $n \in \mathcal{N}^{*+}$.

5.1 HISTOGRAMMES ET BOITES GLISSANTES

5.1.1 Généralités

La méthode la plus facile pour estimer une densité locale consite à regrouper les points dans des boîtes. On crée une image dont chaque pixel est un bin et a pour valeur le rapport entre le nombre de points contenus dans ce pixel et la surface du pixel. Cette image, ou histogramme 2D (à une constante multiplicative[5] près pour des pixels de taille fixe), est une carte de densité. La taille des pixels peut être calculée en fixant l'erreur moyenne que l'on désire avoir sur l'estimation de la densité locale. On considère ici des pixels carrés. Soit ρ le nombre moyen de sources par unité de surface et $E\{N\}$ le nombre moyen de sources dans un pixel. Soit a la valeur du côté d'un pixel de surface S. Le côté des pixels, a, est obtenu à partir de l'erreur désirée ϵ sur la densité par la formule :

$$\epsilon = \frac{\sqrt{E\{N\}}}{S} = \frac{\sqrt{\rho a^2}}{a^2} = \frac{\sqrt{\rho}}{a} \tag{II.63}$$

Avec cette méthode, on obtient des valeurs de densités discontinues. En effet, deux points très proches mais dans deux pixels différents peuvent être associés à des densité locales assez différentes. On peut lisser artificiellement ces valeurs en associant la densité en chaque pixel au point central de ces pixels et en interpolant – linéairement, avec des fonctions splines, ... – les valeurs entre les centres. Bien que visuellement agréable, cette technique n'est pas satisfaisante sur le plan statistique. L'autre problème, de taille, est le comportement de ce genre de méthode dans le cas où la densité locale peut varier fortement entre les différentes zones de l'espace. Certains pixels peuvent se retrouver vides (densité locale nulle) ou avec très peu de source. Dans ce cas l'erreur faite sur l'estimation de la densité peut être grande ($\epsilon_{\hat{\rho}}/\hat{\rho} = 100\%$ si une seule source se trouve à l'intérieur du pixel).

Une autre méthode, qui dérive de la précédente, consiste à utiliser des boites glissantes adaptatives. Cette technique a été implémentée avant d'être abandonnée au profit du k-nn puis du kernel smoothing. Elle peut être vue comme une extension de l'histogramme qui vient d'être présenté. On crée, là aussi, une image. Mais on divise chaque pixel précédent en neuf pixels. Comme le calcul de la densité locale dépend aussi de la magnitude des sources, un système d'image 3D avait été mis en place. Ce système, abandonné depuis est décrit dans la suite de ce paragraphe. Il était notamment utilisé pour tester l'homogénéité d'un FOV, la densité en tout point étant alors directement le rapport entre le nombre de source et la surface du FOV.

5.1.2 Image 3D

Image 3D est le nom que nous avons donné à l'implémentation d'un système de calcul de densités locales rapide adapté aux FOV de XMM. Il est basé sur le principe de boîtes glissantes.

a Principe

L'image 3D (Fig. II.10) consiste en une matrice de pixels constituant un carré recouvrant le cercle d'observation. Chaque pixel élémentaire contient un tableau de valeur. A chaque case de ce tableau est associée une magnitude maximale. Chaque case contient comme valeur le nombre de sources dans la zone couverte par le pixel dont la magnitude vaut au plus la magnitude associée à la case.

[5] la surface des pixels

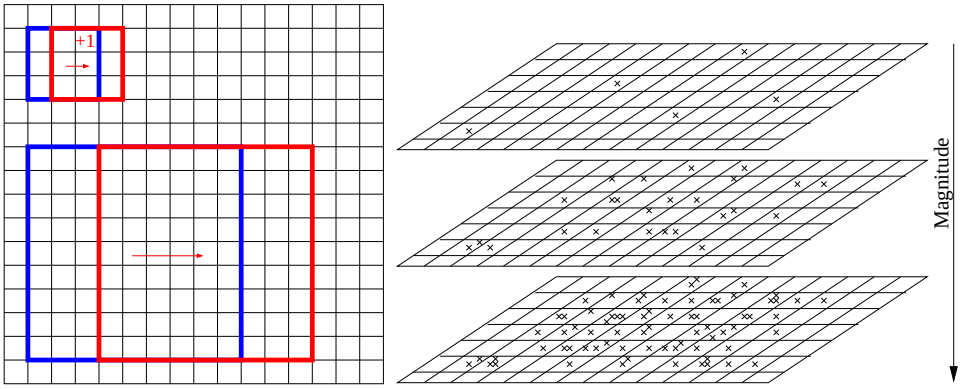


FIG. II.10 – Principe schématique de l'image 3D (droite) et des boîtes glissantes (gauche).

L'image 3D peut donc se résumer à une image composée de deux dimensions spatiales et possédant une dimension supplémentaire. Pour chaque pixel, cette dimension supplémentaire contient le nombre cumulé de sources par tranches de magnitudes croissantes. Le plan correspondant à la magnitude la plus forte, donc l'intensité lumineuse perçue la plus faible, contient le maximum de sources.

b Masque

Un masque (Fig. II.11) est généré pour ne pas prendre en considération dans les calculs les pixels qui ne recouvrent pas complètement la zone d'observation. Pour cela, on vérifie pour chaque pixel que son coin le plus éloigné du centre de l'image se trouve à une distance inférieure au rayon du cône de recherche. Soit R le rayon du champ de vue ee (*arcsec*). Soit d_{pix} la taille d'un

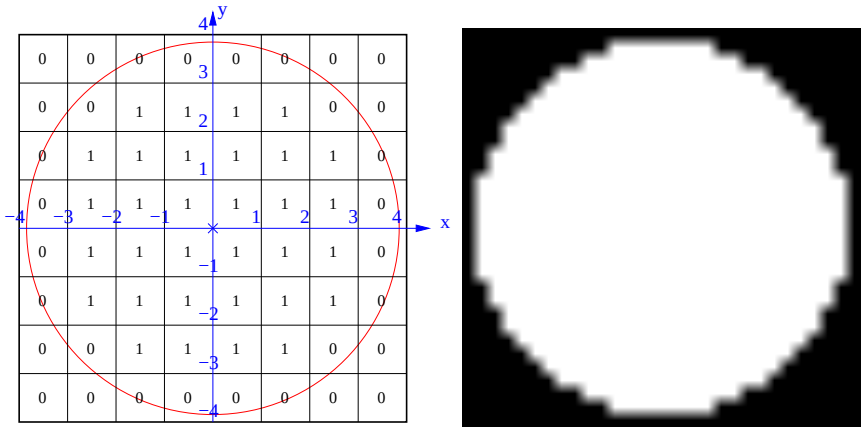


FIG. II.11 – Gauche : Schéma synthétisant la construction du masque (0=rejeté, 1=accepté). Droite : Exemple de masque d'une image 30x30 pixels

pixel en $arcsec$. Soit $O(0,0)$ le centre de l'image qui est au milieu de 4 pixels. La distance entre le centre O et le coin le plus éloigné d'un pixel vaut $\sqrt{x^2+y^2}$. Pour qu'un pixel soit "valide", il suffit qu'il vérifie l'inégalité :

$$d_{pix}\sqrt{x^2+y^2} \leq R \tag{II.64}$$

Avec x le numéro de colonne du pixel ($\in \mathbb{N}^*$) et y le numéro de ligne du pixel ($\in \mathbb{N}^*$) en commençant la numérotation au centre de l'image.

c Détermination de la taille des pixels et des boîtes glissantes

Pour pouvoir utiliser sur cette image un système de boîtes glissantes qui se déplacent par pas de $1/3$, un pixel élémentaire vaut $1/9$ de la surface d'une boîte glissante élémentaire. On fixe

le nombre moyen N_{B_1} de sources, toutes magnitudes confondues, que l'on veut observer dans une boîte glissante élémentaire. La valeur par défaut est fixée de façon arbitraire, par exemple $N_{B_1} = 100$. Dans ce cas, on aura en moyenne 11 sources par pixel. Selon l'homogénéité du champ d'observation, les écarts à la moyenne peuvent être plus ou moins importants.
Pour l'estimation de la densité locale, on fixe deux valeurs :

- N_{moy_λ} : le nombre moyen de sources utilisées pour le calcul de la densité locale
- N_{min_λ} : le nombre minimum de sources qui doivent être utilisées pour le calcul de la densité locale

Soit N_{tot} le nombre total de sources toutes magnitudes confondues dans notre champ de vue. Soit R le rayon du champ de vue en $arcsec$. On veut en moyenne N_{B_1} sources par pixel. Un pixel étant carré, et en écrivant l'égalité de la densité moyenne, son coté a est donné par :

$$a = R\sqrt{\pi \frac{N_{B_1}}{9 \cdot N_{tot}}} \tag{II.65}$$

Pour une magnitude donnée on calcule la taille de la boîte glissante de la façon suivante :

- on cherche pour cette magnitude la valeur médiane N_{med} du nombre de sources par pixel.
- si $N_{med} = 0$ on recommence en prenant non pas un pixel mais une boîte glissante de taille 1 (9 pixels). Dans ce cas N_{med} peut ne plus être un entier, ce qui n'est pas gênant pour les calculs.
- le nombre moyen de pixels dont on a besoin vaut alors $N_{moyMag} = N_{moy_\lambda}/N_{med}$
- l'échelle (ou taille) de la boîte glissante est alors donnée par : $scale = \frac{\sqrt{N_{moyMag}}-1}{2}$ où $scale$ est arrondi au chiffre supérieur.
- si lors d'une recherche on trouve un nombre de sources dans la boîte inférieur à N_{min_λ}, on augmente la taille de la boîte en fonction d'une première estimation de la densité locale et on itère.

Taille	0	1	2	3	4	...	n
Nbr Pixels	1	9	25	49	81	...	$(2n+1)^2$

TAB. II.1 – Nombre de pixel en fonction de la taille d'une boîte glissante

d Détermination de l'homogénéité

Pour chaque plan/index de magnitude on recherche dans toute l'image les nombres minimum N_{min} et maximum N_{max} de sources contenues dans une boîte glissante de taille n. L'image est considérée homogène pour cet index de magnitude et cette échelle (taille de boîte glissante) si l'inégalité suivante est respectée :

$$\boxed{\frac{N_{max} - N_{min}}{\sqrt{N_{max} + N_{min}}} < 3} \tag{II.66}$$

Cette équation traduit le fait que la différence maximum ne doit pas dépasser trois fois la somme quadratique des erreurs de comptage à 1σ.

Si le champ est homogène la densité locale est directement obtenue par le nombre de sources présentes dans le champ, au dessus de la magnitude considérée, et par la surface globale de la zone. Cette surface est calculée à partir du masque.

e Calcul de la densité locale

Dans le cas où le champ n'est pas homogène, la taille de la boîte de recherche dépend de la zone d'homogénéité locale. On prend pour taille de la boîte de recherche de départ la taille qui satisfait au mieux la condition $N_{src} \approx N_{moy_\lambda}$ et $N_{src} \geq N_{min_\lambda}$, avec N_{src} le nombre de sources dans la boîte. On regarde pour toutes les boîtes voisines si le nombre de sources est homogène via l'inégalité (II.66). Si oui, on augmente la taille de la boîte de recherche et on itère. Si non, on reprend la taille de boîte précédente. Si pour la première taille de boîte la zone est inhomogène, on

garde cette valeur mais on la signale par un $flag$. La densité locale est obtenue à partir du nombre de sources présentes dans la zone de recherche et à partir de la taille de la zone de recherche.

f Critiques de la méthode

Les zones d'inhomogénéités peuvent varier en fonction de la taille des pixels imposée au départ, même si cette taille dépend de la densité globale du champ. Elles peuvent également varier en fonction du nombre de sources défini pour le calcul de la densité locale. On remarque que la surface d'une boîte glissante augmente rapidement en fonction de sa taille (le nombre de pixels contenu dans la boîte n'augmente pas linéairement).
On a donc abandonné cette méthode au profit du kernel smoothing.

5.2 LES K PLUS PROCHES VOISINS

La méthode des k plus proches voisins (KNN pour "K Nearest Neighbours") est bien connue. Elle consiste à calculer la densité locale $\hat{g}$ en un point O de l'espace en divisant k par la surface du disque centré en O et de rayon d_k où d_k est la distance entre O et le $k^{\text{ème}}$ plus proche voisin :

$$g = \frac{k}{\pi d_k^2} \tag{II.67}$$

L'erreur associée à cette estimation de la densité locale vaut (voir paragraphe 3.4) :

$$\sigma_g = \frac{\sqrt{k}}{\pi d_k^2} \tag{II.68}$$

On a donc pour tous les points le même rapport signal sur bruit :

$$\frac{S}{N} = \frac{k}{\pi d_k^2} \frac{\pi d_k^2}{\sqrt{k}} = \sqrt{k} \tag{II.69}$$

Mais, comme indiqué dans la thèse de Matthieu Petremand[6] p. 129, "*Les méthodes de type k-plus proches voisins* [...] *sont peu robustes au bruit lorsque k est petit (lorsque k est choisi trop grand, la densité de probabilité estimée est lissée)*". Notre cas est un peu différent du sien. Pour nous, le principal inconvénient réside dans l'aspect discontinu de ce « knn averaging ».

Une solution possible, que nous avons utilisée, pour effectuer des recherches de type k-nn de façon rapide consiste à avoir recours au kd-tree (Press *et al.* 2007). Nous avons développé un code spécifique en Java dans lequel le kd-tree consiste uniquement en un tableau trié.

5.3 LA TESSELATION DE VORONOÏ

La tesselation de Voronoï (Aurenhammer *et al.* 2000) consiste à isoler chaque source dans une cellule. La cellule de Voronoï associée à une source s_i est définie comme étant l'ensemble des points ayant pour plus proche voisin cette même source. Elle est délimitée par les médiatrices entre s_i et ses sources voisines. La densité locale en un point est alors souvent définie comme valant l'inverse de la surface de la cellule de Voronoï dans laquelle se trouve le point. Comme précédemment, l'estimation des densités locales est discontinue et à chaque estimation est associée une erreur relative de 100%. Des études théoriques et basées sur des simulations (Heinrich *et al.* 1998) portent sur les propriétés des diagrammes de Voronoï (l'ensemble des cellules de Voronoï associé à une distribution de points) appliqués sur des distributions poissonniennes de points , soit en anglais les *poisson-voronoï diagrams*. Ainsi, on peut retrouver le paramètre λ d'une distribution poissonnienne 2D en étudiant la surface moyenne (ou médiane) des cellules de Voronoï. On trouvera un exemple d'application de la tesselation de Voronoï pour chercher des surdensités dans Panko & Flin (2004). On trouvera un autre exemple d'utilisation dans Ebeling & Wiedenmann (1993).

[6] http://picabia.u-strasbg.fr/lsiit/perso/collet/ftp/PhD/PhD_Petremand.pdf

Cependant, dans notre cas, on cherche à déterminer des densités locales. Il faudrait pouvoir définir le nombre de cellules nécessaire pour avoir une « bonne » estimation de ces densités locales, ce qui reviendrait à fusionner les k plus proches cellules voisines. On se retrouverait alors avec un knn amélioré, mais discontinu.

5.4 LES ONDELETTES ET LES SOM

La piste des ondelettes a été assez vite abandonnée par manque de temps et en raison de l'utilisation beaucoup plus répandue du *kernel smoothing*. Cependant, une analyse détaillé de Kolaczyk & Dixon (2000) pourrait être intéressante.

On propose ici une méthode non supervisée pour regrouper dans des boîtes (ou *binner* en Franglais) des données en nD. On a remarqué dans les SOM (*Self Organizing Maps*, cf. Chapitre 3) que le nombre de points associé à chaque neurone est à peu près constant. Or l'ensemble des points de l'espace associé à un neurone correspond à la cellule de Voronoi associée à ce neurone dans l'espace des paramètres. En fixant le rapport f entre nombre de neurones et nombre de points dans la distribution qu'on cherche à regrouper, on fixe approximativement le rapport S/N de chaque boîte ou *bin* ($1/\sqrt{f}$). Les boîtes sont les cellules de Voronoi associées à chaque neurone dans l'espace des paramètres.

Les caractéristiques des boîtes peuvent facilement être visualisées sur la carte auto-organisatrice. La CVT, pour *Centroidale Voronoi Tessellation*, qui consiste en un regroupement en boîtes 2D adaptatif est traité dans Cappellari & Copin (2002) et la WVT, pour *Weighted Voronoi Tesselations* dans Diehl & Statler (2006).

Nous n'avons pas poussé les investigations sur ce sujet car comme pour tout système de regroupement en boîtes, la densité locale est discontinue au niveau des bords de ces boîtes.

5.5 LE KERNEL SMOOTHING

5.5.1 Description

Le *kernel smoothing* (KS), aussi appelé méthode des noyaux de Parzen (*Parzen window technique*) ou encore estimation de densité par noyau (*kernel density estimation*), voir Comaniciu & Meer (2002), est une méthode qui permet d'estimer une densité de probabilité à partir d'une distribution de N points en dimension d. On obtient une estimation de la densité en chaque point de l'espace en multipliant la distribution par N. Cette méthode nous permet donc de calculer des densités locales pour tout point de l'espace.

On trouvera une description claire du *kernel smoothing* dans la thèse de Matthieu Petremand[7] au paragraphe 5.3.2.1 "*Les noyaux de Parzen pour l'estimation de la densité de probabilité* $\hat{f}(x)$". On introduit ici le produit de convolution.

Le KS permet d'estimer une densité de probabilité grâce à une fonction dite noyau.
Fixons d'abord les notations. Soit :

- $\vec{x}_i$, $i \in [1, .., N]$, $\vec{x}_i \in \mathbb{R}^d$ les N points de dimension d de notre distribution X ;
- $f(\vec{x})$ la densité de probabilité, que l'on ne connaît pas, dont est issue la distribution X ;
- $\hat{f}(\vec{x})$ une estimation de la densité de probabilité $f(\vec{x})$;
- $K(\vec{x})$ une fonction noyau dont l'intégrale vaut 1 ;
- h une distance appelée « largeur de bande »ou encore « paramètre de lissage ».

Le KS revient à considérer tous les points de la distribution X comme des fonctions δ de Dirac, aussi appelées pics de Dirac. Ainsi, cette distribution peut s'écrire sous forme d'une fonction somme :

$$s(\vec{x}) = \sum_{i=1}^{N} \delta(\vec{x} - \vec{x}_i) \tag{II.70}$$

Où $\delta(\vec{x} - \vec{x}_i)$ est la fonction δ de Dirac centrée en $\vec{x}_i$. L'estimation de la densité de probabilité par le

[7] http://picabia.u-strasbg.fr/lsiit/perso/collet/ftp/PhD/PhD_Petremand.pdf

kernel smoothing peut alors se comprendre comme étant le produit de convolution de la fonction noyau par la somme des pics de Dirac placés en chaque point de notre distribution. Ce résultat est normalisé par le nombre de points, N, et par la valeur du nouveau volume élémentaire induit par la constante de lissage, h, de sorte que l'intégrale sur tout l'espace vaut 1.

$$\hat{f}(\vec{x}) = \frac{1}{Nh^d}(K * s)(\frac{\vec{x}}{h}) = \frac{1}{Nh^d}\sum_{i=1}^{N} K(\frac{\vec{x} - \vec{x}_i}{h}) \tag{II.71}$$

Pour résumer, on place sur chaque point de notre distribution une impulsion de forme K. La somme normalisée de toutes ces impulsions nous donne la densité de probabilité en tout point de l'espace (voir Fig. II.12).

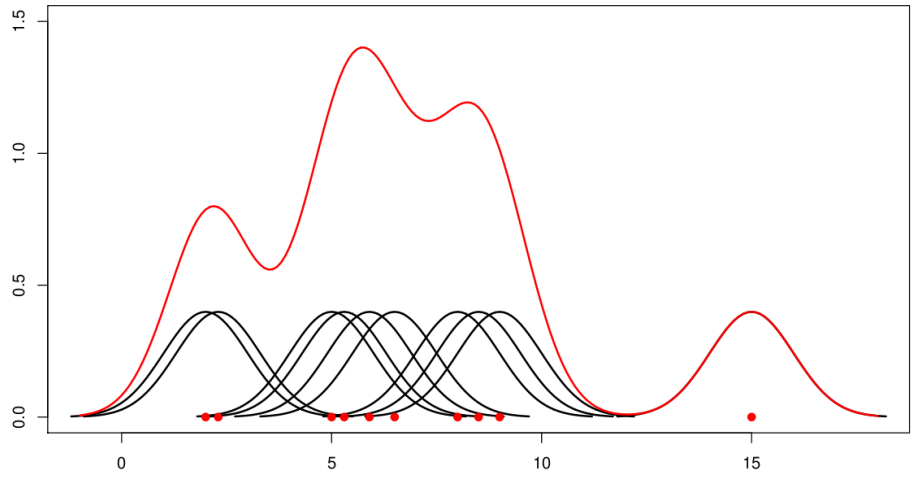


FIG. II.12 – Exemple de kernel smoothing avec noyau Gaussien en dimension 1 (Crédits : Stefanie Scheid). Ici est représentée la densité locale. Il faut la normaliser, soit la diviser par le nombre de points, pour obtenir la densité de probabilité.

Cette formulation (Eq. II.71) correspond à un KS à largeur de bande fixe (*fixed bandwidth kernel smoothing*), puisque h est la même en tout point. On peut compliquer/améliorer le kernel smoothing à largeur de bande fixe en remplaçant h par une matrice de variance-covariance H de dimension d. Les éléments diagonaux – les variances – de H sont alors les largeurs de bande selon chaque axe et les éléments non diagonaux sont les covariances associées. La formule Eq. II.71 se transforme alors en :

$$\hat{f}(\vec{x}) = \frac{1}{N|H|^{1/2}}\sum_{i=1}^{N} K(H^{-1/2}(\vec{x} - \vec{x}_i)) \tag{II.72}$$

Qui peut encore s'écrire, c.f Sain (2002) :

$$\hat{f}(\vec{x}) = \frac{1}{N}\sum_{i=1}^{N} K_H(\vec{x} - \vec{x}_i) \tag{II.73}$$

Avec :

$$K_H(\vec{x}) = |H|^{-1/2}K(H^{-1/2}\vec{x}) \tag{II.74}$$

Le seul paramètre à fixer est la largeur de bande h (ou la matrice H). Il existe pour cela plusieurs méthodes dont une est abordée paragraphe 5.5.2 ci-après.

Une autre alternative consiste à faire varier la largeur de bande h en fonction de $\vec{x}$ ou de $\vec{x}_i$: ce sont des méthodes de KS dites à largeur de bande variable (*variable bandwidth kernel smoothing*). Par exemple, l'estimateur de balloon, ou *balloon estimator* en anglais, présenté dans Sain (2002),

voit sa largeur de bande varier avec $\vec{x}$:

$$\hat{f}(\vec{x}) = \frac{1}{N}\sum_{i=1}^{N} K_{H(\vec{x})}(\vec{x} - \vec{x}_i) \tag{II.75}$$

Le k-NN, pour k^{th} *nearest neighbour*, décrit plus haut est un estimateur de balloon particulier où la largeur de bande correspond à la distance d_k au $k^{\text{ème}}$ plus proche voisin et le noyau est la fonction :

$$K_H(\vec{x}) = \begin{cases} \frac{1}{\pi d_k^2} & \text{si} \quad ||\vec{x}|| < d_k \\ 0 & \text{sinon} \end{cases} \tag{II.76}$$

Dans les cas où on a utilisé cet estimateur, pratique car rapide et facile à implémenter, on a choisi un noyau d'Epanechnikov (voir plus loin) avec pour largeur de bande la distance d_k au $k^{\text{ème}}$ plus proche voisin.

Les KS où la largeur de bande varie en fonction de $\vec{x}_i$ sont appelés *sample-point estimators*. Leur forme générale est :

$$\hat{f}(\vec{x}) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{|H(\vec{x}_i)|^{1/2}} K(H(\vec{x}_i)^{-1/2}(\vec{x} - \vec{x}_i)) = \frac{1}{N}\sum_{i=1}^{N} K_{H(\vec{x}_i)}(\vec{x} - \vec{x}_i) \tag{II.77}$$

En simplifiant au cas d'une largeur de bande égale pour toutes les dimensions :

$$\hat{f}(\vec{x}) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{h(\vec{x}_i)^d} K(\frac{\vec{x} - \vec{x}_i}{h(\vec{x}_i)}) \tag{II.78}$$

Souvent (voir Comaniciu *et al.* (2001)) la largeur de bande en $\vec{x}_i$ est choisie comme valant :

$$h(\vec{x_i}) = h_o \left[\frac{\lambda}{f(\vec{x_i})}\right]^{1/2} \tag{II.79}$$

Avec :

- h_o une largeur de bande fixe ;
- λ une constante de proportionnalité qui joue sur le degré de lissage.

Bien évidemment f n'est pas connue est doit être remplacée par une fonction pilote qui consiste en une première estimation $\tilde{f}$ de f. Comaniciu *et al.* (2001) mentionne qu'un bon choix initial de λ est de prendre la moyenne géometrique $E\{\tilde{f}(\vec{x_i})\}$ et que la méthode est insensible aux petites variations (« fine detail ») de la fonction pilote. La méthode classique consiste à déterminer la fonction pilote par un KS à largeur de bande fixe. Ce qui nécessite une première utilisation de méthodes *plug-in* ou *cross-validation* ou encore *leave-one-out* pour déterminer cette largeur de bande fixe.

Par souci de vitesse d'exécution, et de facilité d'implémentation, nous avons choisi comme fonction pilote le résultat d'un moyennage par les k plus proches voisins : $f(\vec{x_i}) = \frac{k}{\pi d_k^2}$ où d_k est la distance au $k^{\text{ème}}$ plus proche voisin. L'aspect non lisse du moyennage par k-NN peut être considéré comme résultant de petites variations. Comme suggéré précédemment, on choisit λ comme la moyenne de la fonction pilote $\lambda = \frac{N}{\pi R_{fov}^2}$ avec R le rayon de la zone totale (du champ de vue XMM dans notre cas). La largeur de bande h_o choisie correspond à la surface dans laquelle on trouve, en moyenne k points : $h_o = \sqrt{\frac{k}{N}} R_{fov}$. Au final, avec ces choix, on trouve $h(\vec{x_i}) = d_k(\vec{x_i})$, la distance du $k^{\text{ème}}$ plus proche voisin du point $\vec{x_i}$. On n'a donc qu'un seul paramètre à fixer : k.

Revenons un instant sur la fonction noyau. La fonction noyau K, c'est-à-dire le "kernel", doit satisfaire les conditions suivantes :

1 :	$\forall \vec{x} \in \mathbb{R}^d, K(\vec{x}) \geq 0$	positive
2 :	$\exists S \in \mathbb{R}^d, \forall \vec{x} \in \mathbb{R}^d, K(\vec{x}) \leq S$	bornée
3 :	$\int_{\mathbb{R}^d} K(\vec{x}) d\vec{x} = 1$	normalisée
4 :	$\lim_{\|\|\vec{x}\|\| \to +\infty} \|\|\vec{x}\|\|^d K(\vec{x}) = 0$	décroissance rapide

Les fonctions noyaux les plus couramment utilisées, c.f. Fig. II.13 sont :
- noyau gaussien : $(2\pi)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}||\vec{x}||^2\right)$
- noyau exponentiel : $2^{-d} \exp\left(-||\vec{x}||\right)$
- noyau d'Epanechnikov : $K_E(\vec{x})$

Avec :

$$K_E(\vec{x}) = \begin{cases} \frac{1}{2} C_d^{-1}(d+2)(1-||\vec{x}||^2) & \text{si} \quad ||\vec{x}|| < 1 \\ 0 & \text{sinon} \end{cases} \tag{II.80}$$

Où C_d est le volume de l'hypersphère unité de dimension d et vaut :

$$C_d = \begin{cases} \frac{1}{(d/2)!}\pi^{d/2} & \text{si} \quad d \text{ pair} \\ \frac{2^{d+1}(\frac{d+1}{2})!}{(d+1)!}\pi^{(d-1)/2} & \text{si} \quad d \text{ impair} \end{cases} \tag{II.81}$$

Ce qui donne en dimension 2 :

$$K_E(\vec{x}) = \begin{cases} \frac{2}{\pi}(1-||\vec{x}||^2) & \text{si} \quad ||\vec{x}|| < 1 \\ 0 & \text{sinon} \end{cases} \tag{II.82}$$

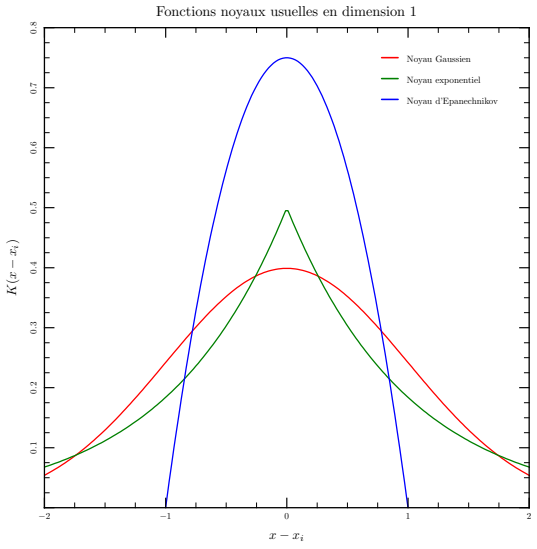


FIG. II.13 – Les trois fonctions noyaux usuels en dimension 1.

A chaque fois qu'on a eu recours au kernel smoothing, on a choisi le noyau d'Epanechnikof, cf. Eq. II.80. Le choix de la largeur de bande est plus important que le choix de la forme du noyau pour la qualité de la reconstruction de f. Cependant le noyau d'Epanechnikof est connu (Comaniciu & Meer 2002) pour être celui qui minimise l'erreur quadratique moyenne intégrée asymptotique (AMISE pour « Asymptotic Mean Integrated Square Error » en anglais).

L'erreur quadratique moyenne intégrée (MISE ou EQMI en français) est définie par :

$$MISE\hat{f}(.;H) = \int_{\mathbb{R}^d} MSE\hat{f}(\vec{x};H)d\vec{x} = \int_{\mathbb{R}^d} \mathbb{E}[\hat{f}(\vec{x};H) - f(\vec{x})]d\vec{x} \tag{II.83}$$

Dans le cas du kernel smoothing à largeur de bande fixe, il existe deux grandes familles de méthodes pour trouver les largeurs de bande optimale :
- Les méthodes dites classiques, basées sur la validation croisée (*cross-validation*) dont la méthode *leave-one-out* est un cas particulier (voir section suivante) ;
- Les méthodes *plug-in* basées sur la minimisation de l'EQMI et qui nécessitent une fonction pilote.

D'après Loader (1999), les méthodes *plug-in* peuvent être fortement biaisées par l'utilisation de fonctions pilotes somme toute assez arbitraires. On a donc préféré une méthode basée sur l'estimateur *leave-one-out* que nous présentons dans la section suivante. Il ne faut cependant pas perdre de vue que dans le cas où la distribution de points contient de grandes zones peu denses, le *kernel smoothing* à largeur de bande fixe n'est pas adapté. On doit lui préférer le *kernel smoothing* à largeur de bandes variables. Le principal problème de ce dernier et qu'il semble impossible de déterminer les paramètres optimisant le résultat (Devroye & Lugosi 1998) tant les possibilités sont grandes.

Les ouvrages suivants sont souvent cités en référence pour les questions liées au *kernel smoothing* :
- D.W. Scott, *Multivariate Density Estimation*. Wiley, 1992 ;
- M.P. Wand and M. Jones, *Kernel Smoothing*. Chapman and Hall, 1995.

5.5.2 Détermination de la largeur de bande optimale par méthode de Monte Carlo Markov Chain

Dans le cas du kernel smoothing à largeur de bande fixe, le seul paramètre à fixer est la largeur de bande ou, tout au plus, les coefficients de la matrice H. La question est alors de savoir comment obtenir la matrice H optimale, c'est-à-dire celle qui permettra la meilleure reconstruction de la densité de probabilité. La méthode utilisée pour résoudre ce problème est tirée de l'article Zhang *et al.* (2006) qui nous a été suggéré par M. André Jalobeanu[8] du LSIIT. Cette section en retrace les grandes lignes.

Le principe est de chercher la matrice H qui minimise la dissimilarité entre la vraie densité de probabilité, non connue, et son estimation par le kernel smoothing. La dissimilarité, mesurée par la distance Kullback-Leibler, est calculée sur un échantillon qui n'est autre que la distribution X de points. La densité estimée en chacun de ces points est obtenue par un kernel smoothing calculé en retirant le point concerné : c'est l'estimateur *leave-one-out*. Ce problème de minimisation revient à utiliser comme critère pour trouver H le maximum de vraisemblance sur l'échantillon X avec l'estimateur *leave-one-out*. Ce critère est appelé critère de validation croisée de la vraisemblance. Pour trouver la solution optimale, on parcourt l'espace des paramètres par une méthode de chaîne de Markov Monte Carlo, ou MCMC pour *Monte Carlo Markov Chain*. Plus précisément, les coefficients de H sont échantillonnés par un algorithme de Metropolis-Hastings à marche aléatoire (*random walk*). Le résultat final est obtenu en faisant la moyenne ergodique des coefficients de H.

a Pré-requis : la distance de Kullback-Leibler

La distance (ou divergence) de Kullback-Leibler (K-L) ou encore l'entropie relative est une mesure de la dissimilarité entre deux densités de probabilité. Soit $f_1(\vec{x})$ et $f_2(\vec{x})$ deux densités de probabilité en dimension d. La distance de Kullback-Leibler qui sépare ces deux distributions s'exprime comme la moyenne par rapport à $f_1(\vec{x})$ du logarithme de leur rapport :

$$D_{KL}(f_1, f_2) = E_{f_1}\left\{\log\left[\frac{f_1(\vec{x})}{f_2(\vec{x})}\right]\right\} = \int_{\mathbb{R}^d} \log\left[\frac{f_1(\vec{x})}{f_2(\vec{x})}\right] f_1(\vec{x}) d\vec{x} \tag{II.84}$$

On peut encore la mettre sous la forme d'une différence d'espérances :

$$\begin{aligned} D_{KL}(f_1, f_2) &= \int_{\mathbb{R}^d} f_1(\vec{x}) \log f_1(\vec{x}) d\vec{x} - \int_{\mathbb{R}^d} f_1(\vec{x}) \log f_2(\vec{x}) d\vec{x} \\ &= E_{f_1}\{\log f_1(\vec{x})\} - E_{f_1}\{\log f_2(\vec{x})\} \end{aligned}$$

Cette mesure n'est pas symétrique entre f_1 et f_2, c'est pourquoi elle n'est pas à proprement parler une "distance" et on la nomme souvent "divergence" de Kullback-Leibler. La distance de K-L est toujours positive et s'annule quand les deux densités de probabilité sont égales.

Démonstration de la positivité de la divergence de Kullback-Leibler : On sait que $\log(x) \leq x - 1$ et que $f_1(\vec{x})$ est toujours positive puisque c'est une densité de probabilité. On peut donc

[8] http://lsiit-miv.u-strasbg.fr/paseo/jalobeanu.php

écrire, à partir de l'équation Eq. II.84 :

$$-D_{KL}(f_1, f_2) = \int_{\mathbb{R}^d} \log\left[\frac{f_2(\vec{x})}{f_1(\vec{x})}\right] f_1(\vec{x})d\vec{x} \leq \int_{\mathbb{R}^d} f_2(\vec{x})d\vec{x} - \int_{\mathbb{R}^d} f_1(\vec{x})d\vec{x} = 0$$

b L'estimateur leave-one-out *et le critère de validation croisée de la vraisemblance*

Dans le cas idéal où la densité de probabilité estimée $\hat{f}(\vec{x})$ vaut exactement la densité réelle $f(\vec{x})$, la distance de K-L entre les deux est nulle. En pratique, pour se rapprocher le plus possible de $f(\vec{x})$, on va chercher à minimiser la distance de K-L $D_{KL}(f, \hat{f})$. Comme $D_{KL}(f, \hat{f})$ est positive et que $E_f\left\{\log f(\vec{x})\right\}$ est une constante, on va chercher à maximiser $E_f\left\{\log \hat{f}(\vec{x})\right\}$.

On ne dispose pas de la fonction continue $\hat{f}(\vec{x})$, mais on peut l'estimer en tout point par le *kernel smoothing*. On ne connaît pas non plus $f(\vec{x})$, mais on a un échantillon X issu de cette densité de probabilité. On peut donc estimer l'espérance du logarithme de la densité de probabilité estimée :

$$E_f\left\{\log \hat{f}(\vec{x})\right\} = \int_{\mathbb{R}^d} f(\vec{x}) \log \hat{f}(\vec{x})d\vec{x}$$

par :

$$\begin{aligned}\hat{E}_f\left\{\log \hat{f}_H(\vec{x})\right\} &= \frac{1}{N}\sum_{i=1}^{N} \log \hat{f}_H(\vec{x}_i) \\ &= \frac{1}{N}\sum_{i=1}^{N} \log\left[\frac{1}{N}\sum_{i=1}^{N} K_H(\vec{x}_i - \vec{x}_j)\right]\end{aligned}$$

En pratique, si on cherche à maximiser cette expression, on obtient une matrice H composée uniquement de zéros. Autrement dit les noyaux se transforment en δ de Dirac. Pour éviter ce problème, on utilise l'estimateur *leave-one-out* où on enlève le point central du *kernel smoothing* :

$$\hat{f}_{H,i}(\vec{x}_i) = \frac{1}{N-1}\sum_{j=1;i\neq j}^{N} K_H(\vec{x}_i - \vec{x}_j) \tag{II.85}$$

Le problème revient donc, au coefficient n^{-1} près, à trouver H par la méthode du maximum de vraisemblance :

$$L(\vec{x}_1, ..., \vec{x}_N|H) = \log\left(\prod_{i=1}^{N} \hat{f}_{H,i}(\vec{x}_i)\right) = \sum_{i=1}^{N} \log\left(\hat{f}_{H,i}(\vec{x}_i)\right) \tag{II.86}$$

Le critère consistant à sélectionner H qui maximise $n^{-1}L(\vec{x}_1, ..., \vec{x}_N|H)$ s'appelle, en anglais, le *likelihood cross-validation criterion*, soit validation croisée de la vraisemblance.

c Utilisation de Chaînes de Markov Monte Carlo (MCMC)

Bien que nous ayons implémenté le cas général, on ne traite ici que le cas simple où la matrice H est diagonale. Pour les détails du cas général, nous renvoyons le lecteur à l'article Zhang *et al.* (2006).

On suppose (voir Zhang *et al.* 2006) que la densité a priori de chaque élément diagonal h_k de H est de la forme :

$$\pi(h_k) \propto \frac{1}{1+h_k^2} \tag{II.87}$$

La probabilité a posteriori s'exprime, selon le théorème de Bayes directement tiré des probabilités conditionelles, par :

$$\pi(H|\vec{x}_1, ..., \vec{x}_n) \propto P(H) \times P(X|H) \propto \left[\prod_{k=1}^{d} \frac{1}{1+h_k^2}\right] \times \prod_{i=1}^{N} \hat{f}_{H,i}(\vec{x}_i) \tag{II.88}$$

Pour des raisons de stabilité numérique, on a utilisé dans notre code le logarithme de cette formule :

$$\log\left(\pi(H|\vec{x}_1,...,\vec{x}_n)\right) \propto \sum_{i=1}^{N} \log\left(\hat{f}_{H,i}(\vec{x}_i)\right) - \sum_{k=1}^{d} \log(1+h_k^2) \tag{II.89}$$

La méthode des MCMC consiste à créer une chaîne contenant un échantillon de matrices H où la $i^{\text{ème}}$ valeur ne dépend que de la $(i-1)^{\text{ème}}$ (propriété Markovienne) en satisfaisant l'égalité :

$$\pi(H_{i-1})p(H_i|H_{i-1}) = \pi(H_i)p(H_{i-1}|H_i) \tag{II.90}$$

Cette égalité est connue sous le nom de *detailed balance equation*, soit équation d'équilibre exhaustive.
La chaîne est un échantillonnage aléatoire. La probabilité pour que la chaîne « visite » un élément H est proportionnelle à $\pi(H|\vec{x}_1,...,\vec{x}_n)$. De fait, les MCMC échantillonnent de façon privilégiée la région où $\pi(H|\vec{x}_1,...,\vec{x}_n)$ est grand.

Il reste à définir $p(H_i|H_{i-1})$ satisfaisant l'équation Eq. II.90. Pour cela, Zhang *et al.* (2006) proposent l'algorithme de Metropolis-Hastings qui revient à définir cette probabilité conditionnelle comme valant :

$$p(H_i|H_{i-1}) = q(H_i|H_{i-1})\alpha(H_i,H_{i-1}) \tag{II.91}$$

où q est une distribution que l'on choisit et α est une probabilité d'acceptation qui vaut :

$$\alpha(H_i,H_{i-1}) = \min\left(1, \frac{\pi(H_{i-1}|\vec{x}_1,...,\vec{x}_n)q(H_{i-1}|H_i)}{\pi(H_i|\vec{x}_1,...,\vec{x}_n)q(H_i|H_{i-1})}\right) \tag{II.92}$$

En pratique, la méthode d'échantillonnage de l'algorithme de Metropolis-Hastings se déroule comme suit :

1. Choisir H_i à partir de la distribution q ;
2. Calculer $\alpha(H_i,H_{i-1})$;
3. Faire un tirage aléatoire uniforme entre 0 et 1. La valeur obtenue est notée u ;
4. Si $\alpha(H_i,H_{i-1}) < u$, alors rejeter la valeur H_i calculée à partir de q : $H_i \leftarrow H_{i-1}$;
5. boucler en recommençant à l'étape 1.

On peut démontrer (voir Press *et al.* 2007, p. 827) que ce processus satisfait l'équation Eq. II.90 avec l'équation Eq. II.91.

Comme suggéré dans Zhang *et al.* (2006), nous avons pris pour q la densité de probabilité Gaussienne multidimensionnelle : $q(H_i|H_{i-1}) = G(\vec{\mu} = H_{i-1}, \vec{\sigma})$. Ce choix permet de simplifier l'équation Eq. II.91 :

$$\alpha(H_i,H_{i-1}) = \min\left(1, \frac{\pi(H_{i-1}|\vec{x}_1,...,\vec{x}_n)}{\pi(H_i|\vec{x}_1,...,\vec{x}_n)}\right) \tag{II.93}$$

Enfin, le $\vec{\sigma}$ est un vecteur de paramètres ajustable. Dans notre implémentation, on le modifie au cours de l'échantillonnage pour avoir un taux d'acceptation des nouvelles valeurs H_i compris entre 0.2 et 0.3.

Les valeurs des coefficients de la matrice H s'obtiennent en faisant la moyenne, pour chaque coefficient, des valeurs (éventuellement multiples) contenues dans la chaîne. Plus la chaîne est longue, plus la moyenne converge vers la valeur optimale. Afin d'assurer un bon compromis entre taille de l'échantillon (de la chaîne) et précision du résultat, on a choisi une chaîne contenant 20 000 échantillons.
Au départ, il faut fixer une valeur initiale qui peut être éloignée du résultat final. Les premiers éléments de la chaîne, peu probables, vont successivement se diriger vers la zone de plus forte probabilité. Pour ne pas biaiser la moyenne finale, il faut donc rejeter ces premiers échantillons. L'ensemble des échantillons rejetés est appelé *burn-in sample*. Sa taille a été fixée à 5 000.

Nous avons implémenté l'algorithme qui vient d'être décrit. Celui-ci est applicable en dimension n.

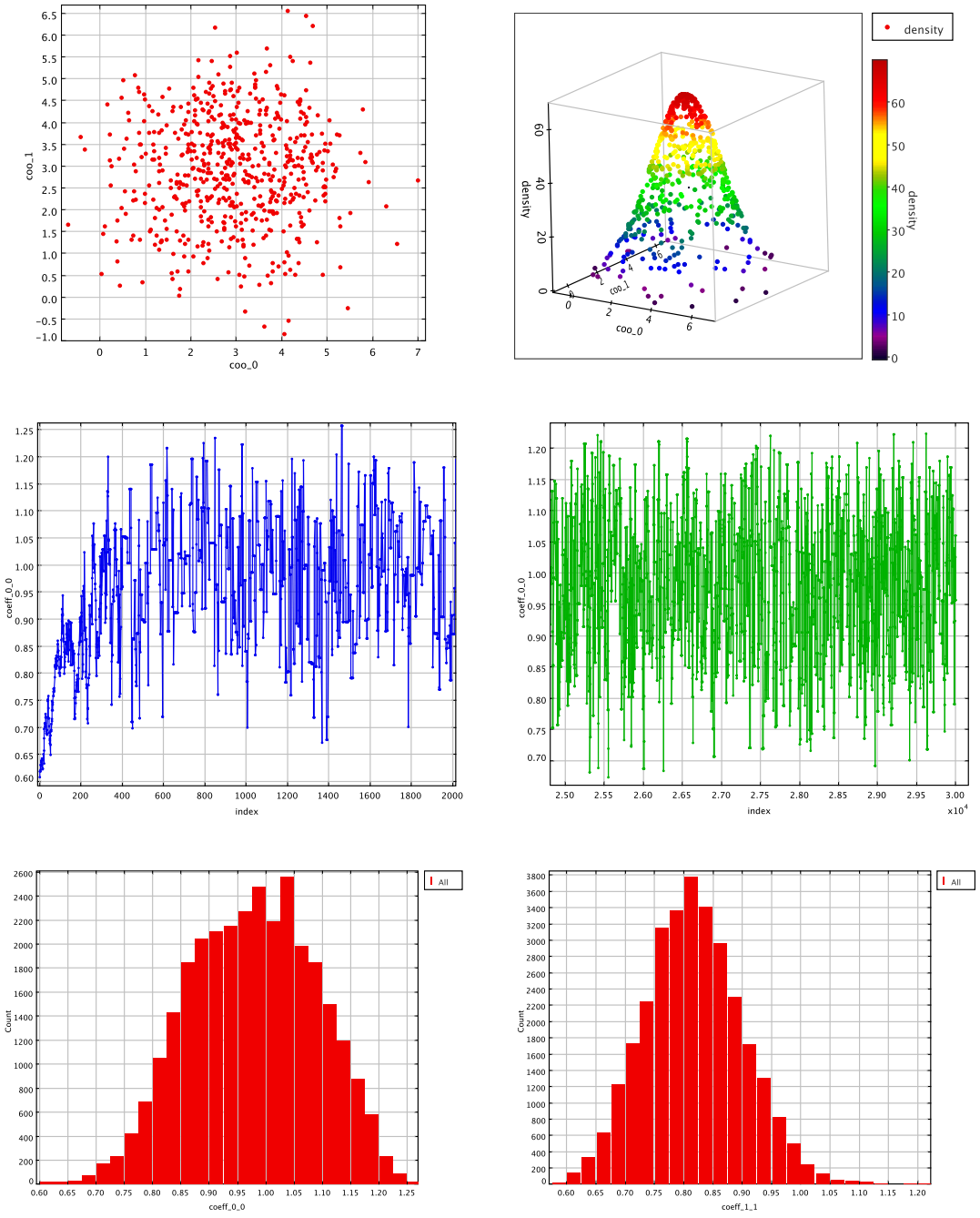


FIG. II.14 – Coin supérieur gauche : distribution test constituée de 500 points issus d'une distribution gaussienne et 100 points issus d'une distribution uniforme entre 0 et 5 pour les deux dimensions ; Coin supérieur gauche : reconstruction de la densité locale ($\propto$ densité de probabilité) par kernel smoothing à largeur de bande fixe calculé par MCMC ; Milieu gauche : les 1 500 premières valeurs du premier coefficient de H de la chaîne. Ici, l'échantillon « burn-in » aurait pu se limiter aux 1 000 premières valeurs ; Milieu droit : les 5 000 dernières valeurs de la même chaîne ; Coin inférieur gauche : histogramme des valeurs échantillonnées du premier coefficient diagonal ; Coin inférieur droit : histogramme des valeurs échantillonnées du second coefficient diagonal.

Références

AURENHAMMER, F., GRAZ, T. U., KLEIN, R., HAGEN, F. & VI, P. I., 2000 ; *Voronoi diagrams*.

Dans *Handbook of Computational Geometry*. Elsevier Science Publishers B.V. North-Holland

BOLLER, T., BERTOLDI, F., DENNEFELD, M. & VOGES, W., 1998 ; *ROSAT All-Sky Survey observations of IRAS galaxies. I. Soft X-ray and far-infrared properties*. A&AS, **129**, 87

BRUSA, M., ZAMORANI, G., COMASTRI, A., HASINGER, G., CAPPELLUTI, N., CIVANO, F., FINOGUENOV, A., MAINIERI, V., SALVATO, M., VIGNALI, C., ELVIS, M., FIORE, F., GILLI, R., IMPEY, C. D., LILLY, S. J., MIGNOLI, M., SILVERMAN, J., TRUMP, J., URRY, C. M., BENDER, R., CAPAK, P., HUCHRA, J. P., KNEIB, J. P., KOEKEMOER, A., LEAUTHAUD, A., LEHMANN, I., MASSEY, R., MATUTE, I., MCCARTHY, P. J., MCCRACKEN, H. J., RHODES, J., SCOVILLE, N. Z., TANIGUCHI, Y. & THOMPSON, D., 2007 ; *The XMM-Newton Wide-Field Survey in the COSMOS Field. III. Optical Identification and Multiwavelength Properties of a Large Sample of X-Ray-Selected Sources*. ApJS, **172**, 353

BUDAVARI, T. & SZALAY, A. S., 2007 ; *Probabilistic Cross-Identification of Astronomical Sources*. ArXiv e-prints, **707**

CAPPELLARI, M. & COPIN, Y., 2002 ; *Adaptive Spatial Binning of 2D Spectra and Images Using Voronoi Tessellations*. Dans *Galaxies : the Third Dimension*, M. Rosada, L. Binette & L. Arias, réds., tm. 282 de *Astronomical Society of the Pacific Conference Series*

COMANICIU, D. & MEER, P., 2002 ; *Mean Shift : A Robust Approach Toward Feature Space Analysis*. IEEE Trans. Pattern Anal. Mach. Intell., **24** (5), 603

COMANICIU, D., RAMESH, V. & MEER, P., 2001 ; *The variable bandwidth mean shift and data-driven scale selection*. Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, **1**, 438

CONDON, J. J., ANDERSON, E. & BRODERICK, J. J., 1995 ; *Radio Identifications of Extragalactic IRAS Sources*. AJ, **109**, 2318

DE RUITER, H. R., ARP, H. C. & WILLIS, A. G., 1977 ; *A Westerbork 1415 MHz survey of background radio sources. II - Optical identifications with deep IIIA-J plates*. A&AS, **28**, 211

DEVROYE, L. & LUGOSI, G., 1998 ; *Variable Kernel Estimates : On The Impossibility Of Tuning The Parameters*. Dans *in : High-Dimensional Probability II, (edited by*. Springer-Verlag

DIEHL, S. & STATLER, T. S., 2006 ; *Adaptive binning of X-ray data with weighted Voronoi tessellations*. MNRAS, **368**, 497

EBELING, H. & WIEDENMANN, G., 1993 ; *Detecting structure in two dimensions combining Voronoi tessellation and percolation*. Phys. Rev. E, **47**, 704

HEINRICH, L., MEHLHORN, N. & MUCHE, L., 1998 ; *Numerical and analytical computation of some second-order characteristics of spatial Poisson-Voronoi tessellations. Statistics*

KAISER, N., WILSON, G., LUPPINO, G. & DAHLE, H., 1999 ; *A Photometric Study of the Supercluster MS0302 with the UH8K CCD Camera : Image Processing and Object Catalogs*. ArXiv Astrophysics e-prints

KOLACZYK, E. D. & DIXON, D. D., 2000 ; *Nonparametric Estimation of Intensity Maps Using Haar Wavelets and Poisson Noise Characteristics*. ApJ, **534**, 490

LOADER, C. R., 1999 ; *Bandwidth selection : classical or plug-in ?* Ann. Statist., **27** (2), 415

MONET, D. G., LEVINE, S. E., CANZIAN, B., ABLES, H. D., BIRD, A. R., DAHN, C. C., GUETTER, H. H., HARRIS, H. C., HENDEN, A. A., LEGGETT, S. K., LEVISON, H. F., LUGINBUHL, C. B., MARTINI, J., MONET, A. K. B., MUNN, J. A., PIER, J. R., RHODES, A. R., RIEPE, B., SELL, S., STONE, R. C., VRBA, F. J., WALKER, R. L., WESTERHOUT, G., BRUCATO, R. J., REID, I. N., SCHOENING, W., HARTLEY, M., READ, M. A. & TRITTON, S. B., 2003 ; *The USNO-B Catalog*. AJ, **125**, 984

OYABU, S., YUN, M. S., MURAYAMA, T., SANDERS, D. B., KAWARA, K., TANIGUCHI, Y., VEILLEUX, S., OKUDA, H., MATSUHARA, H., COWIE, L. L., SATO, Y., WAKAMATSU, K. & SOFUE, Y., 2005 ; *Optical Identification of Infrared Space Observatory Far-Infrared Sources in the Lockman Hole Using a Deep Very Large Array 1.4 GHz Continuum Survey*. AJ, **130**, 2019

PANKO, E. & FLIN, P., 2004 ; *Application of the Voronoi tessellation technique for galaxy cluster search in the Münster Red Sky Survey*. Dans *IAU Colloq. 195 : Outskirts of Galaxy Clusters : Intense Life in the Suburbs*, A. Diaferio, réd.

PELAT, D., 1998 ; *Bases et Méthodes pour le Traitement des Données (Bruits et Signaux)*. Obspm, Paris

PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T. & FLANNERY, B. P., 2007 ; *Numerical Recipes 3rd Edition : The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA

RUTLEDGE, R. E., BRUNNER, R. J., PRINCE, T. A. & LONSDALE, C., 2000 ; *XID : Cross-Association of ROSAT/Bright Source Catalog X-Ray Sources with USNO A-2 Optical Point Sources*. ApJS, **131**, 335

SAIN, S. R., 2002 ; *Multivariate locally adaptive density estimation*. Comput. Stat. Data Anal., **39** (2), 165

SAPORTA, G., 2006 ; *Probabilités, analyse des données et statistique*. Technip, Paris

SEIDELMANN, P. K. & KOVALEVSKY, J., 2002 ; *Application of the new concepts and definitions (ICRS, CIP and CEO) in fundamental astronomy*. A&A, **392**, 341

SINNOTT, R. W., 1984 ; *Virtues of the Haversine*. S&T, **68**, 158

STEPHEN, J. B., BASSANI, L., MOLINA, M., MALIZIA, A., BAZZANO, A., UBERTINI, P., DEAN, A. J., BIRD, A. J., LEBRUN, F., MUCH, R. & WALTER, R., 2005 ; *Using the ROSAT Bright Source Catalogue to find counterparts for IBIS/ISGRI survey sources*. A&A, **432**, L49

SUTHERLAND, W. & SAUNDERS, W., 1992 ; *On the likelihood ratio for source identification*. MNRAS, **259**, 413

WOLSTENCROFT, R. D., SAVAGE, A., CLOWES, R. G., MACGILLIVRAY, H. T., LEGGETT, S. K. & KALAFI, M., 1986 ; *The Identification of IRAS Point Sources - Part One - a 304-DEGREE Field Centred on the South Galactic Pole*. MNRAS, **223**, 279

ZHANG, X., KING, M. L. & HYNDMAN, R. J., 2006 ; *A Bayesian approach to bandwidth selection for multivariate kernel density estimation*. Computational Statistics & Data Analysis, **50** (11), 3009. URL `http://ideas.repec.org/a/eee/csdana/v50y2006i11p3009-3031.html`

Chapitre III

Analyse de données et classification de sources X

1 L'Analyse en Composantes Principales (ACP)

L'analyse en composantes principales, ACP ou *PCA* pour *Principal Component Analysis*, est une méthode qui permet l'étude de la structure de données multidimensionnelles. C'est également un outil qui fournit des critères pour réduire la dimension d'un espace de paramètres en conservant le plus d'informations possible. L'ACP est très appréciée des statisticiens et trouve des applications dans divers problèmes de domaines aussi variés que l'astronomie, la biologie, la chimie, le traitement d'image, ou encore l'économie.

Cette méthode nous a servi, par exemple, à rechercher parmi plusieurs paramètres observés ceux qui étaient les plus discriminants, c'est-à-dire ceux qui séparaient le mieux différentes populations d'objets. En réduisant l'espace des paramètres d'entrée de nos algorithmes de classification, l'ACP nous a également permis à la fois une optimisation des temps de calcul, mais aussi la suppression d'informations redondantes et de bruits résiduels.

Après une présentation des grands principes de l'ACP, nous aborderons le domaine de l'ACP robuste. Nous verrons ensuite les modifications apportées aux algorithmes classiques afin de prendre en compte les erreurs de mesures associées à chaque paramètre observé. Le but est d'obtenir la meilleur estimation possible des composantes principales. Nous présenterons également la *Kernel PCA (KPCA)*, version non linéaire de l'ACP, accompagnée d'exemples de résultats obtenus avec cette dernière méthode.

1.1 L'ACP CLASSIQUE

Nous abordons ici l'algorithme classique de l'ACP. L'entrée comporte un certain nombre d'observations, ou de sources, pour lesquelles on dispose d'un certain nombre de paramètres. L'ACP consiste juste en un changement de base de ces observations dans l'espace des paramètres. Les axes de la nouvelle base dans cet espace, appelés composantes principales, sont donc des combinaisons linéaires des paramètres d'entrée. Les premiers de ces nouveaux axes portent le maximum d'informations. On peut donc réduire la dimension de la sortie en projetant les observations sur le sous-espace constitué des premières composantes principales.

Dans ce qui suit, nous présentons plus en détail le principe de l'ACP. Nous décrivons ensuite l'algorithme classique, suivi de critères permettant de choisir le nombre de dimensions à conserver en sortie. Nous verrons enfin comment relier les paramètres d'entrée aux composantes principales.

1.1.1 Principe

Soit $\mathbb{R}^m$ l'espace vectoriel de m paramètres mesurés pour n sources (ou observations). Les données d'entrée consistent en n vecteurs dans la base orthonormée de $\mathbb{R}^m$. Le but de l'ACP est de trouver une nouvelle base dans $\mathbb{R}^m$ dont les premiers axes portent le maximum d'information et les derniers le minimum. Autrement dit, la projection des données sur le premier axe de la nouvelle base, appelé premier axe principal, doit présenter la variance maximale. Cette variance est la variance maximale que présente le nuage de points sur une droite passant par son centre de gravité, qui est aussi le centre du repère (comme on le verra la première étape de l'ACP consiste à centrer les données). Le premier axe principal correspond à la direction de cette droite.

Nous proposons une autre manière d'appréhender ce problème : l'ACP revient à assimiler le nuage de points à une hyper-ellipse de dimension m. Le premier axe principal correspond à l'axe majeur de cette ellipse. Le second axe principal correspond au 2ème plus grand axe de l'ellipse, etc. Enfin le dernier axe correspond à l'axe mineur.
Pour une ellipse de dimension 2, le demi grand axe et le demi petit axe ont pour valeur la racine carrée des valeurs propres issues de la diagonalisation de la matrice de variance-covariance associée à l'ellipse (voir paragraphe 6.2.4 de Pelat (1998)). La base dans laquelle l'axe des x correspond à l'axe majeur et l'axe des y correspond à l'axe mineur est la base dans laquelle l'ellipse se trouve sous sa forme canonique. Dans cette base l'axe des x, c'est à dire l'axe sur lequel se trouve l'axe majeur, correspond au premier axe principal et l'axe des y au second axe principal. On passe de l'ancienne base à la base canonique par une matrice de rotation dont les colonnes correspondent aux vecteurs propres de la diagonalisation.
L'algorithme de l'ACP consiste donc principalement en la diagonalisation de la matrice de variance-covariance des paramètres d'entrée. On projette ensuite les données d'entrée, préalablement centrées, dans la nouvelle base par la matrice de rotation issue de la diagonalisation. Le premier axe principal est le vecteur propre associé à la plus grande valeur propre, etc. Pour des questions de mise à échelle, et parce qu'il convient souvent de travailler sur des données sans unité, on se sert souvent des données centrées et réduites. Dans ce cas on ne diagonalise pas la matrice de variance-covariance mais la matrice de corrélation, son équivalent dans ce cas.

De fait, les valeurs propres représentent la variance des données projetées sur le vecteur propre associé. Le pourcentage d'inertie totale représenté par les k premières composantes principales vaut :

$$I_k = \sum_{p=1}^{k} \lambda_p / \sum_{q=1}^{n} \lambda_q \tag{III.1}$$

où les λ_p sont les valeurs propres triées par ordre croissant. Cette grandeur, I_k, est importante car elle permet d'évaluer la perte d'information suite à la réduction de dimensionnalité.

1.1.2 Algorithme

Je donne ici l'algorithme de l'ACP classique tel que je l'ai implémenté. Dans la suite, nous utiliserons les indices i et j pour désigner les sources (lignes de la matrice) et les indices p et q pour désigner les paramètres (colonnes de la matrice).

Entrée : 1 entrée :
- X : matrice de données. Elle contient n observations (lignes) de m paramètres (colonnes). Dans notre cas, n sources et leurs m caractéristiques.

Étape 1 : on centre et, éventuellement, on réduit les données :

$$\forall i \in [1,n], \forall p \in [1,m] : x_{ip} \leftarrow \frac{x_{ip} - \mu_p}{\sigma_p}$$

Étape 2 : on calcule la matrice V de variance-covariance des paramètres d'entrée ; cette matrice est symétrique ($v_{pq} = v_{qp}$) et contient m lignes et m colonnes :

$$\forall p,q \in [1,m]^2 : v_{pq} \leftarrow \sum_{i=1}^{n} x_{ip} x_{iq}$$

Étape 3 : on diagonalise V, les valeurs propres et les vecteurs propres sont respectivement notés λ_p et $\vec{u}_p$

Étape 4 : les couples $(\lambda_p, \vec{u}_p)$ sont triés par ordre décroissant des valeurs propres

Étape 5 : on extrait les projections de nos données centrées réduites sur les N premières composantes principales (cp), c'est-à-dire les N axes qui contiennent le plus d'informations :

$$\forall i \in [1, n], \forall p \in [1, N] : cp_p(\vec{x}_i) \leftarrow \vec{x}_i.\vec{u}_p$$

Sortie : CP : matrice qui contient les n observations centrées et éventuellement réduites (lignes), projetées sur les N premières composantes principales (colonnes).

$$\forall i \in [1, n], \forall p \in [1, N] : cp_{ip} \leftarrow cp_p(\vec{x}_i)$$

1.1.3 Nombre de composantes principales à conserver

Un des intérêts de l'ACP est la réduction de dimensionalité. Des lors, la question du nombre de composantes principales à conserver s'impose. Un premier choix peut se porter sur le critère de Kaiser : si on travaille sur des données centrées-réduites, on conservera les axes principaux associés aux valeurs propres > 1. « *L'explication de ce critère est que l'on conserve les axes dont la variance, c'est-à-dire la valeur propre, est supérieure à celle des variables analysées, ici égale à 1* » (Tufféry 2007).

Une autre solution consiste à fixer le seuil « d'informations perdues ». On choisit alors le nombre k de composantes principales à conserver en fixant le pourcentage d'inertie minimum (Eq. III.1) qu'elles doivent représenter.

Un dernier critère consiste à ne garder que les valeurs propres qui dépassent leur moyenne de plus de 2 écart-types (Saporta 2006) :

$$\lambda > 1 + 2\sqrt{\frac{p-1}{n-1}} \tag{III.2}$$

où n est le nombre de sources et p le nombre de valeurs propres.

Aucun de ces critères n'est absolu. On peut, par exemple, détecter une cassure, un coude, permettant de séparer les valeurs propres en deux groupes dont on ne conservera que celui contenant les valeurs les plus grandes.

1.1.4 Lien entre vecteurs propres et vecteurs d'origine

« *La méthode la plus naturelle pour donner une signification à une composante principale c'est de la relier aux variables initiales* [...] *en calculant les coefficients de corrélation linéaire* [...] *et en s'intéressant aux plus forts coefficients en valeur absolue* » (Saporta 2006).

Dans le cas où on travaille sur des données centrées réduites, le coefficient de corrélation c_{pk} entre le $p^{\text{ème}}$ paramètre d'entrée x_p, appelé variable initiale dans l'extrait ci-dessus, et la $k^{\text{ème}}$ composante principale de valeur propre λ_k et vecteur propre $\vec{u}_k$ vaut :

$$c_{pk} = \sqrt{\lambda_k} u_{kp} \tag{III.3}$$

Pour un couple de composantes principales on peut alors synthétiser leurs corrélations avec les paramètres d'entrée par un "cercle de corrélation" (voir exemple figure Fig. III.1). Il y a autant de points que de paramètres d'entrée. Pour chaque point, l'abscisse vaut la valeur du coefficient de corrélation avec la première composante principale du couple et l'ordonnée celle du coefficient de corrélation avec la seconde.

1.2 L'ACP MODIFIÉE

Le principal inconvénient de l'algorithme classique de l'ACP est qu'il n'est pas robuste et ne prend pas en compte les erreurs de mesures associées à chaque valeur. La présence de valeurs extrêmes, d'*outliers*, peut fortement biaiser les estimations de moyennes et de variances nécessaires

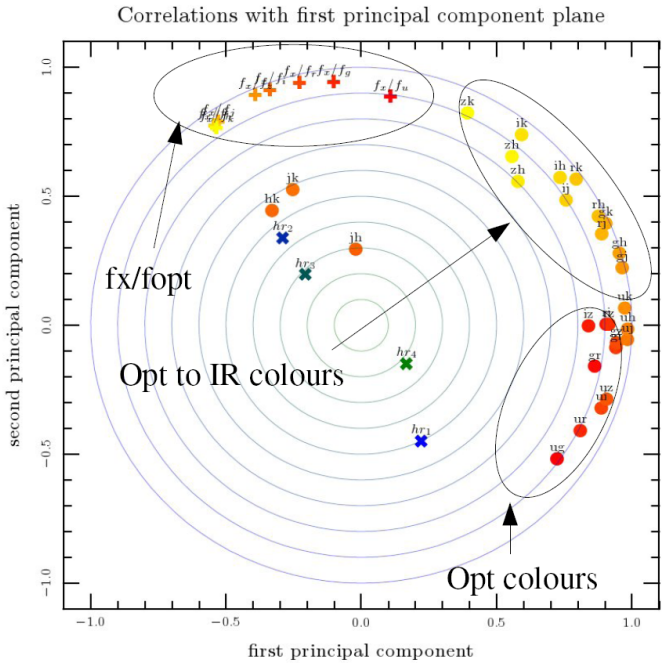


FIG. III.1 – Corrélations entre les deux premiers axes principaux et les paramètres d'entrée issues de l'identification croisée entre les sources du 2XMMi, du SDSS DR6 et du 2MASS. On voit ici que les rapports de dureté – ou *hardness ratio* – XMM (hr1, hr2, hr3 et hr4) et les couleurs infrarouges issues du catalogue 2MASS (jh, hk et jk) n'interviennent presque pas dans le premier plan principal.

pour calculer la matrice de corrélation entre paramètres d'entrée. Ainsi, un seul point de mesure déviant, auquel est associée une grande erreur, peut radicalement changer les coefficients de corrélation et donc les axes principaux de l'ACP.
Pour chaque paramètre, toutes les valeurs dont nous disposons sont entachées d'erreurs. On souhaite utiliser cette information pour obtenir une meilleure estimation des coefficients de corrélation et, de ce fait, rendre plus robuste l'ACP sur nos données. C'est le principe de base de notre ACP améliorée.

Cette partie débute avec un aperçu des méthodes existantes pour rendre robuste l'ACP. Ces algorithmes sont robustes dans la mesure où ils cherchent essentiellement à minimiser, voire à supprimer, l'influence des *outliers*. Sera ensuite exposée la méthode qu'on propose et qu'on a utilisée. Enfin, on parlera d'une autre solution proposée pour prendre en compte les erreurs de mesures dans l'ACP.

1.2.1 ACP robustes

La littérature qui traite des ACP robustes est abondante en traitement d'image, chimie, économie,etc. On peut séparer deux types d'approches :

- une classique, basée sur la diagonalisation de la matrice de variance-covariance ou de corrélation ;
- une itérative, basée sur les algorithmes dits de *projection-pursuit*.

De nombreuses solutions ont été, et sont encore, proposées. Toute une famille est basée sur des estimations robustes de la matrice de variance-covariance (ou de corrélation), sans prendre en compte les erreurs de mesure. D'autres semblent les prendre en compte et, bien que plus complexes, mériteraient qu'on s'y attarde.

a Algorithmes de projection-pursuit

Les algorithmes de *projection-pursuit* (PP) cherchent à déterminer le sous espace, souvent de dimension 1, qui maximise un index de projection, noté S^2 dans la suite. L'ACP peut donc être vue comme un cas particulier d'algorithme de PP où la variance est choisie comme index de projection. En effet, on obtient la première CP en choisissant le sous-espace de dimension 1 – une droite de direction $\vec{u}_1$ – maximisant la variance :

$$\vec{u}_1 = \underset{||\vec{u}_1||=1}{\arg\max} S^2({}^t\vec{u}_1.\vec{x}_1, ..., {}^t\vec{u}_1.\vec{x}_n) \tag{III.4}$$

On projette ensuite les données dans le sous-espace de dimension $m-1$ orthogonal à $\vec{u}_1$. Il suffit pour cela de soustraire à $\vec{x}_i$ sa composante selon $\vec{u}_1$, c'est-à-dire sa projection sur $\vec{u}_1$:

$$\vec{x}_i^2 = \vec{x}_i^1 - ({}^t\vec{u}_1.\vec{x}_i^1)\vec{u}_1 \tag{III.5}$$

et on trouve les $\vec{u}_k$ suivants itérativement :

$$\vec{u}_k = \underset{||\vec{u}_k||=1,\, \vec{u}_k \perp \vec{u}_1, ..., \vec{u}_k \perp \vec{u}_{k-1}}{\arg\max} S^2({}^t\vec{u}_k.\vec{x}_1^k, ..., {}^t\vec{u}_k.\vec{x}_n^k) \tag{III.6}$$

où :

$$\vec{x}_i^k = \vec{x}_i^{k-1} - ({}^t\vec{u}_{k-1}.\vec{x}_i^{k-1})\vec{u}_{k-1} \tag{III.7}$$

Une caractéristique des algorithmes de PP, qui peut s'avérer être un avantage, et qu'ils ne nécessitent pas le calcul de toutes les composantes principales. Comme ils sont itératifs, on peut s'arrêter n'importe quand. Cette caractéristique semble notamment recherchée en chimiométrie[1] où le nombre de paramètres peut être élevé, mais n'est pas d'une grande importance dans notre cas. Croux *et al.* (2006) fournit des références pour des algorithmes de PP dont un est basé sur le *simulated annealing*. L'algorithme qu'ils proposent, nommé GRID, est basé sur des estimations robustes de la moyenne et de la variance. Il consiste à :

- centrer les données à partir d'un estimateur robuste de la moyenne (médiane L_1 ou la *coordinatewise* médiane)
- utiliser un estimateur robuste de la variance (MAD pour *Median Absolute Deviation* ou Q, le premier quartile de la différence deux à deux entre tous les points) comme index de projection S^2
- utiliser un système de grille adaptative en parcourant des plans pour trouver la direction optimale satisfaisant la formule Eq. III.6

Une définition de la médiane L1 ainsi qu'un algorithme permettant de la déterminer peuvent être trouvés dans l'article Vardi & Zhang (2000). D'autres, dans Hössjer & Croux (1995). Dans le cas où tous les poids valent 1, la médiane L_1 est équivalente à la médiane géométrique définie par :

$$\vec{\text{med}}_{L_1}(\vec{x}_1, ..., \vec{x}_n) = \underset{\vec{y}\in\mathbb{R}^m}{\arg\min} \sum_{i=1}^{n} ||\vec{x}_i - \vec{y}|| \tag{III.8}$$

La *coordinatewise* médiane consite à traiter chaque coordonnée individuellement. C'est donc simplement le vecteur dont chaque coordonnée est la médiane du paramètre correspondant. Dans les deux équations qui suivent, $\forall i \in [\![1, ..., n]\!]$, $x_i \in \mathbb{R}$:

$$MAD(x_1, ..., x_n) = 1.48 \underset{j}{\text{med}} |x_j - \underset{i}{\text{med}}\, x_i| \tag{III.9}$$

$$Q(x_1, ..., x_n) = 2.22\{|x_i - x_j|; 1 \le i \le j \le n\}_{(\binom{n}{2}/4)} \tag{III.10}$$

Le papier de Chen *et al.* (2006) explore une autre approche basée sur le *kernel smoothing* et l'algorithme *mean shift*. La PP est ici utilisé « à l'envers » : les auteurs recherchent d'abord les directions sur lesquelles les données projetées présentent un pic de densité maximum, et donc implicitement une variance minimum. Le pic de densité est calculé à l'aide de l'algorithme *mean*

[1] « *La chimiométrie* [ou *chemometric* en anglais,] *est l'application des outils mathématiques, en particulier statistiques, pour obtenir le maximum d'informations à partir des données chimiques* »(wikipédia)

shift qui est une montée de gradient sur une densité de probabilité estimée à l'aide d'un *kernel smoothing*. Les points les plus éloignés du pic maximum sont considérés comme des *outliers*. Le résultat final est obtenu par une ACP classique sur l'ensemble des *inliers*, c'est-à-dire des entrées qui n'ont pas été marquées comme *outliers* pendant l'algorithme.

b Estimation robuste de la matrice de variance-covariance

Les approches mentionnées ici sont équivalentes à ce qui peut être pratiqué avec les algorithmes de PP. Une PCA classique est appliquée sur une estimation robuste de la matrice de variance-covariance.

Les M-estimateurs L'article de Gentiane *et al.* (2000) traite d'une approche répandue où la PCA robuste est basée sur des estimations robustes des matrices de variance-coariance ou de corrélation.

Dans le cas des estimations basées sur les M-estimateurs, on associe à chaque observation un poids. Ce poids est défini par une fonction de poids qui dépend de la distance en sigma entre l'observation et le centre d'une hyper-ellipse. Cette hyper-ellipse est définie itérativement par l'estimation du vecteur des moyennes et par l'estimation de la matrice de variance-covariance. Les fonctions de poids sont généralement d'abord croissantes avec l'écart à la moyenne estimée puis constante ou décroissante. Cf. Saporta (2006) p. 320, *« les estimateurs du maximum de vraisemblance sont des cas particuliers des M-estimateurs »*.

Autres méthodes D'autres méthodes se contentent, par exemple, de rejeter les *outliers* paramètre par paramètre. Une solution, par exemple, consiste à rejeter tous les points se trouvant en dehors d'une boîte à moustaches (ou *box plot* en anglais) : cette méthode (cf. p. 48 et p. 487 de Tufféry (2007)) consiste à considérer une valeur comme extrême si elle dépasse $q_3 + 1.5(q_3 - q_1)$ ou est plus petite que $q_1 - 1.5(q_3 - q_1)$, où q_i sont les quartiles de la distribution avec q_1 la valeur située à $1/4$ des valeurs triées par ordre croissant, q_2 la valeur médiane,...

c Prise en compte des erreurs de mesure

Une solution pour prendre en compte les erreurs de mesure est d'utiliser les formules des moindres carrés pondérés (cf. Pelat (1998) p. 272). On associe à chaque valeur un poids qui vaut l'inverse normalisé de la variance de l'erreur associée à cette valeur. Cette méthode est un cas particulier des M-estimateurs qui ne base pas le calcul des poids sur une fonction de poids, mais les relie directement aux erreurs de mesure. En pratique, 80% du poids total peut se répartir sur quelques points de mesure seulement. Le résultat est alors fortement biaisé.

Parmis les méthodes qui prennent en compte les erreurs de mesure et qu'il nous faudrait approfondir, on peut citer les méthodes de *Total Least Square* (TLS) (Branham 1999), (Branham 2001) et *Maximum Likelihood PCA* (MLPCA) (Wentzell *et al.* 1997). Ces deux approches semblent être équivalentes (Schuermans *et al.* 2005). L'approche TLS est également connue dans la littérature sous les noms de *orthogonal regression*, *errors-in-variables regression* ou encore *measurement error modeling*.

1.2.2 Bases statistiques de l'ACP modifiée

Notre ACP modifiée se base sur une estimation robuste de la matrice de variance-covariance en prenant en compte les erreurs de mesure. On s'inspire des formules des moindres carrés pondérés, qui font partie d'une classe de M-estimateurs, avec l'idée que la variance de chaque point est la somme quadratique de la variance de l'erreur de mesure et de la variance de la distribution de points. Notre but est surtout d'utiliser l'information supplémentaire apportée par les erreurs sur chaque valeur. On peut éventuellement se débarrasser préalablement des *outliers* via des coupures ou l'utilisation de la boîte à moustaches. Les *outliers* sont ici les valeurs extrêmes auxquelles ne sont pas associées de grandes erreurs de mesure.

On présente dans la section suivante les bases statistiques de cette ACP modifiée.

1.2.3 Estimation de la moyenne et de la variance

Dans la suite, on suppose que les erreurs sont indépendantes des valeurs auxquelles elles sont associées. On reviendra sur cette hypothèse dans la section suivante. Le but de cette section est d'introduire les estimateurs de la moyenne, de la variance et de la covariance qu'on propose. On démontre qu'ils sont non biaisés.

On introduit les notations utilisées dans la suite : on dispose d'une matrice X ayant n lignes et m colonnes ; les m colonnes correspondent à m variables différentes et les n lignes à n observations ; on notera i et j les indices portant sur les lignes et p et q ceux portant sur les colonnes ; soit Σ la matrice des erreurs σ_{ip} associée aux coefficients x_{ip} de la matrice X ; Σ est également de taille $n \times m$.

La variance d'un point de mesure est la somme de la variance liée à son erreur de mesure dûe à l'instrument (aux algorithmes, ...) et de la variance intrinsèque à la distribution qu'on ne connaît pas :

$$\sigma_{ip}^2 = \sigma_{\epsilon_{ip}}^2 + \sigma_p^2 \tag{III.11}$$

On utilise la formule des moindres carrés pondérés (c.f. Pelat (1998) p. 272) pour trouver une estimation γ de la moyenne :

$$\chi^2(\gamma_p) = \sum_{i=1}^{n} \frac{(x_{ip} - \gamma_p)^2}{\sigma_{ip}^2} = \sum_{i=1}^{n} \omega'_{ip}(x_{ip} - \gamma_p)^2 \text{ , avec } \omega'_{ip} = \frac{1}{\sigma_{ip}^2} \tag{III.12}$$

En utilisant l'estimateur du maximum de vraisemblance $\partial\chi^2(\gamma_p)/\partial\gamma_p = 0$, on trouve la relation :

$$\gamma_p = \frac{\sum\limits_{i=1}^{n} \omega'_{ip} x_{ip}}{\sum\limits_{i=1}^{n} \omega'_{ip}} \tag{III.13}$$

La moyenne pondérée de chaque ligne (aussi appelé M-estimateur, cf. Saporta (2006) p.320) s'écrit donc :

$$\gamma_p = \sum_{i=1}^{n} \omega_{ip} x_{ip} \tag{III.14}$$

Où ω_{ip} est le poids normalisé associé à la valeur x_{ip} tel que :

$$\sum_{i=1}^{n} \omega_{ip} = 1 \tag{III.15}$$

$$\omega_{ip} = \frac{\omega'_{ip}}{\sum\limits_{i=1}^{n} \omega'_{ip}} \tag{III.16}$$

$$\omega_{ip} = \frac{1}{\sigma_{ip}^2} \frac{1}{\sum\limits_{i=1}^{n} \frac{1}{\sigma_{ip}^2}} \tag{III.17}$$

Cet estimateur de la moyenne est un estimateur non biaisé :

$$E\{\gamma_p\} = \sum_{i=1}^{n} \omega_{ip} E\{x_{ip}\} = \mu_p \sum_{i=1}^{n} \omega_{ip} = \mu_p \tag{III.18}$$

L'erreur sur cette moyenne, calculée à partir de la variance (avec hypothèse d 'indépendance des x_{ip} pour avoir $Var(\sum) = \sum(Var)$), vaut :

$$\sigma_{\gamma_p} = \frac{1}{\sqrt{\sum\limits_{i=1}^{n} \omega'_{ip}}} = \frac{1}{\sqrt{\sum\limits_{i=1}^{n} \frac{1}{\sigma_{\epsilon_{ip}}^2 + \sigma_p^2}}} \tag{III.19}$$

Si $\forall i \in [[1...n]]$, $\sigma_{\epsilon_{ip}} = 0$, c'est à dire $\sigma_{ip} = \sigma_p$, on retrouve la formule bien connue $\sigma_\mu = \sigma_p/\sqrt{n}$.

Posons pour estimateur de la variance totale :

$$\hat{\sigma}_p^2 = \sum_{i=1}^{n} \omega_{ip}(x_{ip} - \gamma_p)^2 \tag{III.20}$$

On se rend compte que cette estimateur est biaisé :

$$E\{\hat{\sigma}_p^2\} = \sigma_p^2 - \sigma_{\gamma_p}^2 \tag{III.21}$$

On choisirait donc pour estimateur de la variance l'estimateur non biaisé suivant :

$$\hat{\sigma}_p^2 = \sigma_{\gamma_p}^2 + \sum_{i=1}^{n} \omega_{ip}(x_{ip} - \gamma_p)^2 \tag{III.22}$$

$$= \frac{1}{\sum_{i=1}^{n} \omega'_{ip}} \left(1 + \sum_{i=1}^{n} \omega'_{ip}(x_{ip} - \gamma_p)^2 \right) \tag{III.23}$$

Le défaut de cet estimateur et qu'il rend compte de la variance globale de la distribution, erreurs comprises. Si on veut estimer σ_p^2, il nous faut un autre estimateur. On propose ici un nouvel estimateur qui fonctionne bien en pratique. On sait que la variance des variables centrées réduites vaut 1. On cherche donc σ_p^2 tel qu'il satisfasse l'équation :

$$\frac{1}{n} \sum_{i=1}^{n} \frac{(x_{ip} - \gamma_p)^2}{\sigma_{\epsilon_{ip}}^2 + \sigma_p^2} = 1 \tag{III.24}$$

On résout cette équation par la méthode de Newton en recherchant les points où s'annule la fonction :

$$f(\sigma_p) = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_{ip} - \gamma_p)^2}{\sigma_{\epsilon_{ip}}^2 + \sigma_p^2} - 1 \tag{III.25}$$

Remarque : si toutes les erreurs de mesures $\sigma_{\epsilon_{ip}}$ sont nulles, la solution de l'équation nous donne la formule classique de la variance. A condition qu'on remplace γ_p par la valeur théorique μ_p, sinon, cet estimateur est légèrement biaisé :

$$E\left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{(x_{ip} - \gamma_p)^2}{\sigma_p^2} \right\} = 1 - \frac{\sigma_{\gamma_p}^2}{\sigma_p^2} = 1 - \frac{1}{\sqrt{n}} \tag{III.26}$$

Dans notre cas :

$$E\left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{(x_{ip} - \gamma_p)^2}{\sigma_{\epsilon_{ip}}^2 + \sigma_p^2} \right\} = \frac{var\{X_p\}}{n} \sum_{i=1}^{n} \frac{1}{\sigma_{\epsilon_{ip}}^2 + \sigma_p^2} - \frac{1}{n} \tag{III.27}$$

Si on remplace $var\{X_p\}$ par l'estimateur non biaisé $\hat{\sigma}_p^2$, on trouve :

$$E\left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{(x_{ip} - \gamma_p)^2}{\sigma_{\epsilon_{ip}}^2 + \sigma_p^2} \right\} = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_{ip} - \gamma_p)^2}{\sigma_{\epsilon_{ip}}^2 + \sigma_p^2} \tag{III.28}$$

Et on a choisi σ_p tel que cette somme vaille... 1 ! Cet estimateur est donc un estimateur non biaisé.

Rappel sur la méthode de Newton :

$$x_{i+1} = x_i - \frac{f'(x_i)}{f(x_i)} \tag{III.29}$$

Il nous faut donc calculer la dérivée de f :

$$\frac{\partial f(\sigma_p)}{\partial \sigma_p} = \frac{-2}{n} \left(\frac{\partial \gamma_p}{\partial \sigma_p} \sum_{i=1}^{n} \frac{x_{ip} - \gamma_p}{\sigma_{\epsilon_{ip}}^2 + \sigma_p^2} + \sigma_p \sum_{i=1}^{n} \frac{(x_{ip} - \gamma_p)^2}{(\sigma_{\epsilon_{ip}}^2 + \sigma_p^2)^2} \right) \tag{III.30}$$

Avec :

$$\frac{\partial \gamma_p}{\partial \sigma_p} = \frac{-2\sigma_p}{\sum_{i=1}^{n} \frac{1}{\sigma^2_{\epsilon_{ip}} + \sigma^2_p}} \sum_{i=1}^{n} \frac{x_{ip} - \gamma_p}{(\sigma^2_{\epsilon_{ip}} + \sigma^2_p)^2} \tag{III.31}$$

Maintenant que l'on a une estimation des σ_p on peut calculer la corrélation entre deux variables. On calcule pour cela la covariance entre les variables centrées réduites :

$$\rho_{p_1,p_2} = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_{ip_1} - \gamma_{p_1})(x_{ip_2} - \gamma_{p_2})}{\sqrt{(\sigma^2_{\epsilon_{ip_1}} + \sigma^2_{p_1})(\sigma^2_{\epsilon_{ip_2}} + \sigma^2_{p_2})}} \tag{III.32}$$

Plus l'erreur sur une grandeur est importante, moins elle aura d'effet sur la corrélation. On retrouve bien, voir formule (III.24), $\rho_{p_k,p_k} = 1$ et, une fois de plus, si toutes les erreurs de mesures sont nulles, on retrouve la formule classique (cf. Saporta (2006) p. 126) :

$$\rho_{p_1,p_2} = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_{ip_1} - \gamma_{p_1})(x_{ip_2} - \gamma_{p_2})}{\sigma_{p_1} \sigma_{p_2}} \tag{III.33}$$

1.2.4 Hétéroscédasticité

Définition : en statistique, un vecteur de variables aléatoires est hétéroscédastique si les variables aléatoires ont différentes variances. Dans au moins trois cas, les données peuvent être hétéroscédastiques (voir Fig. III.2 pour deux de ces cas) :

- si la variance de l'erreur (l'erreur de mesure) est différente pour chaque valeur mesurée ;
- si un même jeu de données contient des jeux de données avec des moyennes sur les erreurs de mesures différentes ; par exemple si on crée une liste de sources dont certaines ont une magnitude optique de l'USNO et les autres du SDSS, les erreurs typiques sur les magnitudes seront très différentes.
- si les erreurs de mesures sont corrélées avec les valeurs des mesures.

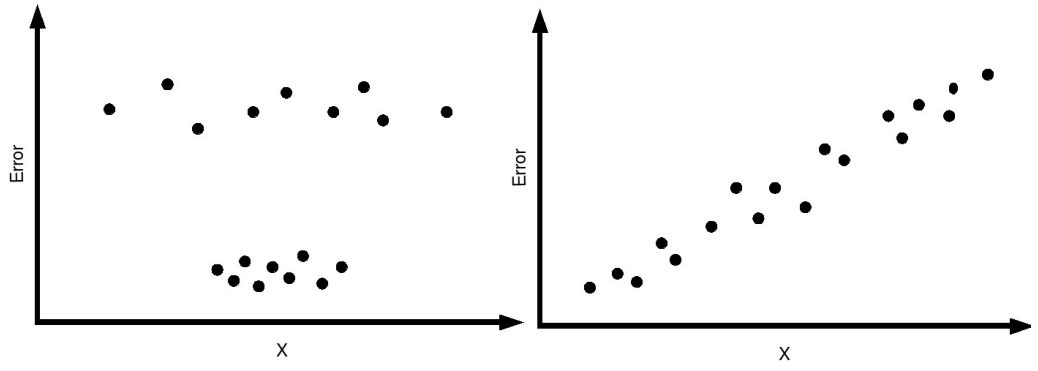


FIG. III.2 – Voici deux exemples de données hétéroscédastiques. Gauche : jeu de données contenant deux distributions d'erreur différentes impliquant une dispersion plus grande d'une partie des points autour de la valeur moyenne. Droite : erreurs corrélées aux valeurs de la grandeur mesurée.

Le cas qui nous dérange est le cas où l'hétéroscédasticité des erreurs entraîne une corrélation entre un paramètre et les erreurs associées. Ca peut par exemple être le cas avec les flux en rayons X. Plus le flux X est élevé, plus l'erreur sur ce flux est importante. En général, les erreurs sur le flux (F_X) sont $\propto \sqrt{F_X}$, mais faisont l'hypothèse que la dépendance est linéaire. Dans ce cas, on estime la droite de régression qui lie la grandeur physique x considérée aux erreurs associées :

$$\sigma_x = ax + b \tag{III.34}$$

On applique alors la transformation suivante :

$$x' = \frac{\ln(ax+b)}{a} \tag{III.35}$$

L'erreur associée à cette nouvelle grandeur vaut :

$$\ln(x') = \ln(\frac{\ln(ax+b)}{a}) \tag{III.36}$$

$$\frac{dx'}{x'} = \frac{d\frac{\ln(ax+b)}{a}}{x'} \tag{III.37}$$

$$\Delta x' = \frac{\Delta x}{ax+b} \tag{III.38}$$

$$\sigma_{x'} = \frac{\sigma_x}{ax+b} \tag{III.39}$$

Les erreur $\sigma_{x'}$ se répartissent donc autour d'une droite d'équation $y = 1$. Il n'y a donc pas de corrélation entre $\sigma_{x'}$ et x'.

1.2.5 Problèmes liés à l'ACP en vue d'une classification

L'ACP permet de trouver des structures dans des données multidimensionnelles. Or si le but est de différencier plusieurs classes d'objets, il faut que le nombre d'individus correspondant à chaque classe soit le même. Si, par exemple, une catégorie d'objets est dominante dans l'analyse, l'ACP aura tendance à extraire les structures à l'intérieur de cette classe. Une stratégie possible est de munir chaque objet d'un poids valant $\frac{M}{n_c M_i}$ où n_c est le nombre de classes et M_i le nombre d'objets dans la classe de l'objet considéré.

1.3 LA "KERNEL PCA"

L'ACP par noyau, ou *Kernel PCA* (Schölkopf *et al.* (1998), Schölkopf *et al.* (1999)), est une analyse en composantes principales non linéaire, basée sur le *kernel trick*. Comme pour l'ACP, le but est de rechercher des dépendances entre paramètres d'entrée. Mais, contrairement à l'ACP classique, la *Kernel PCA* (KPCA) ne se limite pas aux dépendances linéaires. Cette dernière est donc potentiellement plus puissante que la première.
Avant de décrire l'algorithme que nous avons implémenté, ainsi que sa validation, nous présentons le principe général de la KPCA. Nous exposons ensuite les résultats obtenus avec cette méthode sur un échantillon de données.

1.3.1 Principe

L'idée de la KPCA est de projeter l'espace des paramètres de départ dans un espace de plus grande dimension, le *feature space*, dont les axes contiennent des combinaisons non linéaires des paramètres d'entrées. On applique ensuite une PCA classique dans cet espace.
Imaginons une série de N observations de deux paramètres. Une observation i est décrite par un vecteur $\vec{x}_i \in \mathbb{R}^2$ dont la première coordonnée x_{ip_1} contient la valeur du premier paramètre – par exemple le rapport $\log_{10}(F_X/F_{opt})$ – et la deuxième x_{ip_2} la valeur du second paramètre. Imaginons ensuite que l'ensemble des observations présente, dans le plan $(\vec{x}_{p_1}, \vec{x}_{p_2})$ défini par nos deux seuls paramètres, des structures annulaires centrées en $(0,0)$. Il serait alors judicieux d'utiliser comme paramètre supplémentaire la distance ou le carré de la distance d'une observation par rapport au centre du plan $(\vec{x}_{p_1}, \vec{x}_{p_2})$. On projette alors nos observations dans un espace de dimension 3 par la fonction :

$$\begin{array}{rlcl} \Phi(\vec{x}): & \mathbb{R}^2 & \rightarrow & \mathbb{R}^3 \\ & \vec{x} & \rightarrow & \Phi(\vec{x}) \\ & (x_{ip_1}, x_{ip_2}) & \rightarrow & (x_{ip_1}, x_{ip_2}, x_{ip_1}^2 + x_{ip_2}^2) \end{array}$$

Le produit scalaire entre deux observations a et b dans ce nouvel espace s'écrit :

$$\begin{aligned}\Phi(\vec{x_a}).\Phi(\vec{x_b}) &= (x_{ap_1}, x_{ap_2}, x_{ap_1}^2 + x_{ap_2}^2).(x_{bp_1}, x_{bp_2}, x_{bp_1}^2 + x_{bp_2}^2) \\ &= x_{ap_1}x_{bp_1} + x_{ap_2}x_{bp_2} + x_{ap_1}^2x_{bp_1}^2 + x_{ap_1}^2x_{bp_2}^2 + x_{ap_2}^2x_{bp_1}^2 + x_{ap_2}^2x_{bp_2}^2 \\ &= \vec{x_a}.\vec{x_b} + (\vec{x_a}.\vec{x_a}) * (\vec{x_b}.\vec{x_b})\end{aligned}$$

On peut l'exprimer sous la forme d'une fonction de deux vecteurs de l'espace de départ :

$$k(\vec{x}, \vec{y}) = \Phi(\vec{x}).\Phi(\vec{y}) = \vec{x}.\vec{y} + (\vec{x}.\vec{x}) * (\vec{y}.\vec{y})$$

C'est cette fonction $k(\vec{x}, \vec{y})$ qu'on appelle « noyau », ou encore *kernel* dans la langue de Shakespeare, qui donne le nom à cette méthode.
L'ACP classique est basée sur la diagonalisation de la matrice de variance-covariance des paramètres d'entrée. En pratique, on peut démontrer que, pourvu qu'on dispose d'une fonction noyau, on peut calculer les valeurs propres et les vecteurs propres associés à la matrice de variance-covariance des paramètres dans l'espace de projection sans connaître $\Phi(\vec{x})$, donc sans avoir besoin de calculer les vecteurs $\Phi(\vec{x_i})$ associés aux observations $\vec{x_i}$, et donc sans calculer les nouveaux paramètres : cette astuce est appelée *kernel trick*. On a seulement besoin de la matrice de variance-covariance des observations projetées dans le nouvel espace, à savoir la matrice K des produits scalaires entre ces observations dans l'espace de projection :

$$\forall i, j \in [1, N]^2 : k_{ij} = k(\vec{x_i}, \vec{x_j})$$

Pour les démonstrations, nous renvoyons le lecteur à l'article fondateur de de la KPCA (Schölkopf *et al.* 1998).
Une fonction noyau peut correspondre à une projection dans un espace de très grande dimension, voire de dimension infinie. Dans ce cas, l'utilisation du *kernel trick* ne permet pas de trouver tous les axes principaux, mais, au pire, les N premiers (N étant le nombre d'observations en entrée).

Remarques : l'article Schölkopf *et al.* (1998) contient quelques erreurs que nous avons relevées. Les voici :

- page 3, après l'équation (15), lire "By virtue of (8) and (15)" au lieu de "By virtue of (8) and (12)" ;
- page 14, deuxième ligne de l'équation (48), il manque $\frac{1}{M}$ devant les deux sommes ;
- page 14, dernière ligne de l'équation (48), il manque $\frac{1}{M}$ devant l'avant dernier terme.

1.3.2 Algorithme

Une implémentation de la KPCA existe dans *RapidMiner*[2]. Les tests effectués ont montré que cette implémentation[3] est lente pour un nombre même réduit (~ 1000) d'observations en entrée. La lecture du code nous a appris que la partie diagonalisation est réalisée par le package *Jama*[4] faisant appel à la routine *tred2()* décrite dans Press *et al.* (2007), p. 582. Des versions plus rapides existent dans le cas où la matrice est très grande. De plus, toujours à la lecture du code, il semblerait que la matrice des produits scalaires ne soit pas recentré dans l'espace de projection (Étape 4 ci-après) et les vecteurs propres ne sont pas normalisés (Étape 5 ci-après). Les autres implémentations disponibles sont sous Matlab.
On a donc décidé de coder notre propre version de la KPCA. Voici l'algorithme que j'ai implémenté :

Entrée : 2 entrées :

- X : matrice de données. Elle contient n observations (lignes) de m paramètres (colonnes). Dans notre cas, n sources et leurs m caractéristiques.
- k : une fonction noyau

[2] `http://rapid-i.com/`

[3] En l'occurrence la version 1.3.4.1 2008/09/16 du KernelPCA.java

[4] `http://math.nist.gov/javanumerics/jama/`

Étape 1 : on centre et, éventuellement, on réduit les données :

$$\forall p \in [1,m], \forall i \in [1,n] : x_{ip} \leftarrow \frac{x_{ip} - \mu_p}{\sigma_p}$$

Étape 2 : on calcule la matrice K des nxn produits scalaires possibles entre les observations $\vec{x_i}$ dans l'espace de projection ; cette matrice est symétrique ($k_{ij} = k_{ji}$) et contient n lignes et n colonnes :

$$\forall i,j \in [1,n]^2 : k_{ij} \leftarrow k(\vec{x_i}, \vec{x_j})$$

Étape 3 : les données doivent aussi être centrées dans l'espace de projection ; la matrice des produits scalaires des données centrées dans cet espace peut se retrouver à partir de K :

$$\forall i,j \in [1,n]^2 : k_{ij} \leftarrow k_{ij} - \mu_{k_{.j}} - \mu_{k_{i.}} + \mu_{k_{..}}$$

Étape 4 : on diagonalise K, les valeurs propres et les vecteurs propres sont respectivement notés λ_i et $\vec{\alpha_i}$; ces couples sont triés par ordre décroissant des valeurs propres.

Étape 5 : on normalise les vecteurs $\vec{\alpha_i}$ en imposant l'égalité $\lambda_i(\vec{\alpha_i}.\vec{\alpha_i}) = 1$:

$$\forall i,j \in [1,n]^2 : \alpha_{ij} \leftarrow \frac{\alpha_{ij}}{\sqrt{\lambda_i \sum_{l=1}^{n} \alpha_{il}^2}}$$

Étape 6 : on extrait les projections de nos données centrées réduites sur les N premières composantes principales (*cp*), c'est-à-dire les N axes qui contiennent le plus d'informations :

$$\forall i \in [1,n], \forall l \in [1,N] : cp_l(\vec{x_i}) \leftarrow \sum_{j=1}^{n} \alpha_{l,j} k(\vec{x_j}, \vec{x_i})$$

Sortie : CP : matrice qui contient les n observations centrées et éventuellement réduites (ligne), projetées sur les N premières composantes principales (colonnes).

$$\forall i \in [1,n], \forall l \in [1,N] : cp_{il} \leftarrow cp_l(\vec{x_i})$$

1.3.3 Implémentation

L'étape 4 de l'algorithme, à savoir la diagonalisation de la matrice K, est le point critique de la KPCA. On peut en effet disposer de nombreuses observations, 10000 par exemple. Dans ce cas, la matrice K est de taille 10000x10000. Avec des matrices si grandes, il faut des algorithmes de diagonalisation qui soient optimisés, à la fois :

- pour minimiser les temps de calculs et l'espace mémoire requis.
- pour assurer la stabilité numérique des solutions.

Nous avons pour cela utilisé la version 0.8 du code *JLAPACK* (Doolin *et al.* 1999) disponible gratuitement sur le web[5]. En effet, comme mentionné précédemment, la routine du « *Numerical Recipes* » (Press *et al.* 2007) s'avère trop lente. On a fait appel à la classe *Dsyevr* dont la documentation[6] fait référence au manuscrit d'une thèse de doctorat (Dhillon 1998).

Au niveau des performances, le code de diagonalisation d'une matrice 10000x10000 utilisant *JLAPACK* nécessite 4 Go de mémoire virtuelle Java (option -Xmx4096M) et 1h30 de temps de calcul sur un processeur cadencé à 3.0 GHz. Ces résultats ont été obtenus avec la machine virtuelle JAVA6 de sun sur le serveur "pulsar" de l'équipe Hautes Énergies.

[5] http://www.netlib.org/java/f2j/

[6] http://icl.cs.utk.edu/projectsfiles/f2j/javadoc/org/netlib/lapack/Dsyevr.html

1.3.4 Validation de notre implémentation

Le premier test consiste à comparer les résultats d'une ACP classique avec ceux de la KPCA utilisant le noyau $k(\vec{x},\vec{y}) = \vec{x}.\vec{y}$, les deux méthodes étant équivalentes dans ce cas. Mais ce cas particulier ne permet pas pour autant de valider notre propre implémentation, et nous ne disposions pas d'algorithme de référence. Or on sait que la KPCA est équivalente à une PCA dans un espace de projection. La solution évidente pour valider notre algorithme est donc de créer un échantillon test, de le projeter dans un espace de dimension plus grande en connaissant la fonction noyau qui correspond au produit scalaire dans ce nouvel espace. Il suffit alors de comparer le résultat de la KPCA sur les données de départ avec celui de la PCA sur la projection de ces même données.

On a créé un échantillon de test synthétique composé de 3 classes. L' espace de départ est de dimension 3 et on a utilisé le noyau polynomial d'exposant 2 suivant : $k(\vec{v}_1,\vec{v}_2) = (\vec{v}_1.\vec{v}_2 + 1)^2$. Ce noyau correspond à la projection :

$$\begin{array}{rccl} \Phi(\vec{v}): & \mathbb{R}^3 & \rightarrow & \mathbb{R}^{10} \\ & (x,y,z) & \rightarrow & (\sqrt{2}x, \sqrt{2}y, \sqrt{2}z, \sqrt{2}xy, \sqrt{2}xz, \sqrt{2}yz, x^2, y^2, z^2, 1) \end{array}$$

On a ainsi pu vérifier qu'on obtenait bien les mêmes résultats avec les deux méthodes.

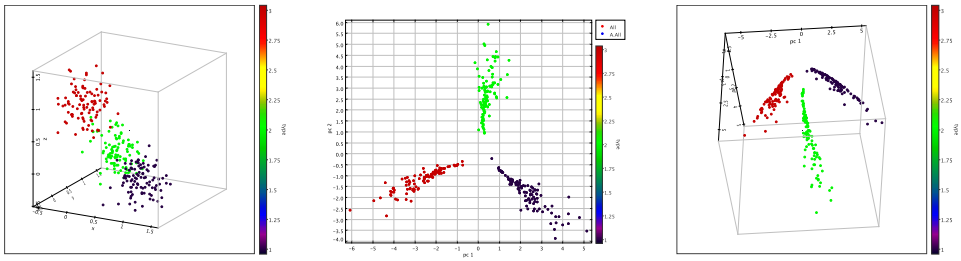


FIG. III.3 – Application de la KPCA avec un noyau polynomial de degré 2 sur un échantillon de test (gauche). Une KPCA a été appliquée sur l'échantillon de test, ainsi qu'un ACP classique après projection dans l'espace de dimension 10 défini par le noyau polynomial de degré 2. Centre : premier plan principal. Droite : trois premiers axes principaux. Les résultats se superposent parfaitement.

1.3.5 Résultats et analyse

Les noyaux les plus communs on été testés sur nos données, à savoir :

- les noyaux polynômiaux : $k(\vec{x},\vec{y}) = (\vec{x}.\vec{y} + 1)^d$
- les fonctions à base radiale : $k(\vec{x},\vec{y}) = \exp\left(-\frac{||\vec{x}-\vec{y}||^2}{2\sigma^2}\right)$
- les noyaux de type réseaux neuronaux : $k(\vec{x},\vec{y}) = \tanh((\vec{x}.\vec{y}) + b)$

Les résultats que nous avons obtenus en variant les paramètres de ces quelques noyaux ne nous ont pas semblé plus pertinents que ceux obtenus par l'ACP classique. Nous avons fait figurer trois exemples Fig. III.4.

Du fait des longs temps de calcul, la recherche expérimentale du noyau le plus adapté à un problème donné, avec des paramètres optimaux, peut être fastidieuse. C'est, nous le verrons, une des limitation des SVMs. Notons tout de même que la KPCA est un bon moyen de valider visuellement la pertinence d'une fonction noyau utilisée en entrée d'une SVM.

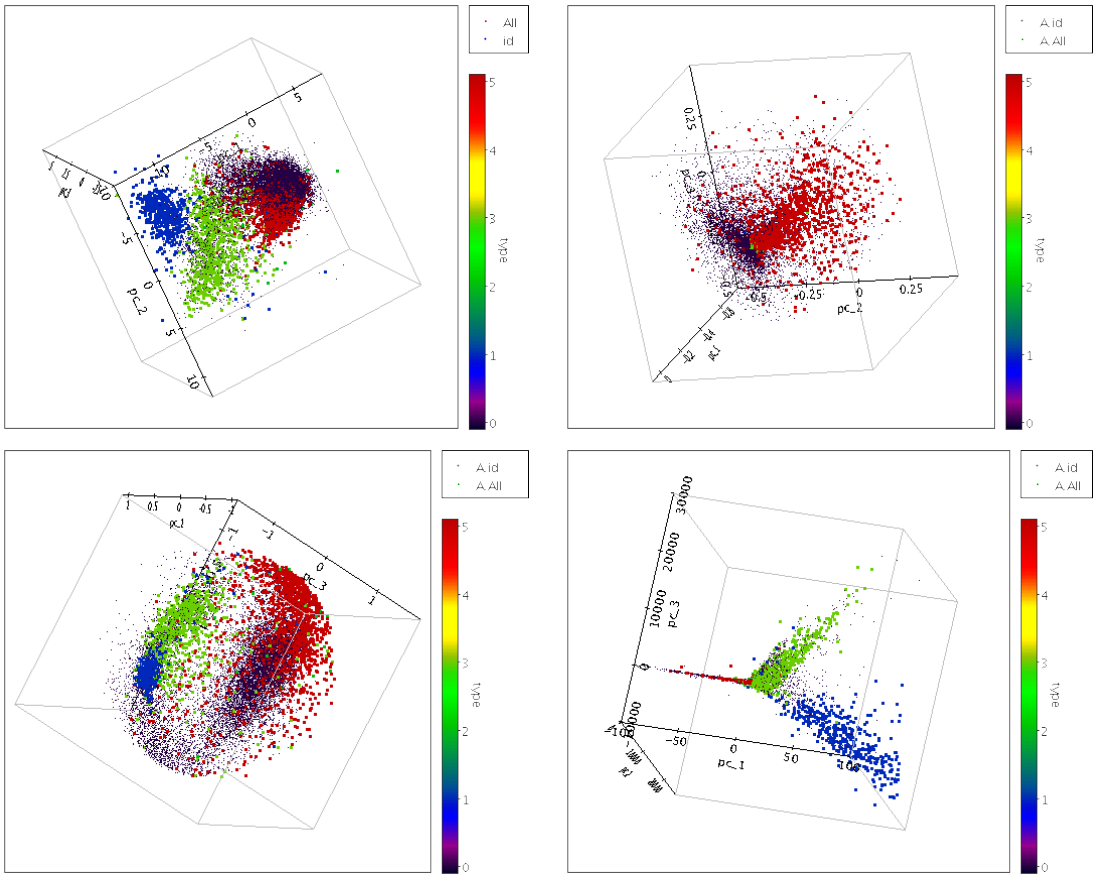


FIG. III.4 – Application de la KPCA sur le résultat de la corrélation du 2XMMi avec le SDSS DR7. Coin supérieur gauche : produit scalaire simple (même résultat que l'ACP classique) ; Coin supérieur droit : noyau à base radiale avec le paramètre $\sigma = 1$; Coin inférieur gauche : noyau de type réseaux neuronaux avec le paramètre $b = 1$; Coin inférieur droit : noyau polynômial de degré 4. La couleur code les différents types de sources : non connu (noir), étoile (bleu), binaire (vert foncé), galaxie (vert clair) et quasar (rouge). Les différents noyaux testés ici ne semblent pas mieux séparer les différentes classes d'objet que ne le fait l'ACP classique.

2 Réseaux de neurones

2.1 INTRODUCTION

Les réseaux de neurones sont des outils informatiques d'analyse de données. Ils sont apparus au milieu du XX^{e} siècle et s'inspirent du monde biologique. Le premier article parlant d'un modèle informatique de neurone date de 1943 (McCulloch & Pitts 1943) et un des articles fondateurs des réseaux de neurones en informatique date de 1959 (Lettvin *et al.* 1959). Les neurones informatiques modélisent, de façon plus ou moins élaborée, les neurones biologiques et certains réseaux cherchent à imiter la nature. Les cartes auto-organisatrices de Kohonen, qui font l'objet de cette partie, puisent leur inspiration dans le système visuel : les neurones de la carte correspondent aux neurones qui sont placés sur la rétine et les neurones de la couche d'entrée correspondent aux neurones du cortex visuel (cf. paragraphe 15.1.2 de Rojas (1996)).

Il existe aujourd'hui une multitude de réseaux de neurones distincts les uns des autres. Chacun a ses particularités propres et est adapté à divers types d'applications. Parmis les familles les plus courantes, on peut citer :

Les perceptrons multicouches La famille des perceptrons multicouches fait partie des types de

réseaux de neurones les plus connus. Une implémentation est par exemple disponible dans le logiciel *MLPfit*[7] du CERN. La structure de ces réseaux se divise souvent en trois couches : une couche d'entrée, une couche de sortie et une ou parfois plusieurs couches cachées. Chaque neurone n'est relié qu'aux neurones des couches précédentes, mais à tous les neurones de la couche précédente. Pour un exemple d'application d'un perceptron multicouche avec algorithme de retropropagation, voir par exemple Bazell & Peng (1998).

Les réseaux de Kohonen Les réseaux de Kohonen se divisent en trois sous-familles :

1. Les VQ (*Vector Quantization*) : méthode souvent qualifiée d'estimateur de densité non supervisé. C'est une méthode de *clustering* se rapprochant de l'algorithme des *k-means* (Kohonen 1988).
2. Les SOM (*Self Organizing Maps*) : permettent de projeter des données de dimension élevée sur une carte souvent 2D (Kohonen 1997), (Rojas 1996). On pourra trouver un exemple d'application de SOM au CDS (Poincot *et al.* 1998) et dans ce manuscrit.
3. Les LVQ (*Learning Vector Quantization*) : méthode qui permet une classification par « recherche du plus proche voisin »(Kohonen 1988). C'est un équivalent supervisé des SOM. Pour des exemples en astronomie, voir Bazell & Peng (1998), Cortiglioni *et al.* (2001) et Zhang & Zhao (2004).

Les deux premières de ces trois classes consistent en des réseaux à apprentissage non supervisé alors que la dernière est à apprentissage supervisé.

Les réseaux de Hopfield Les réseaux de Hopfield (Hopfield 1982) sont des réseaux composés de seulement deux couches (la couche d'entrée et celle de sortie) complètement bouclées : chaque neurone est relié à tous les autres. C'est un réseau asynchrone qui réagit jusqu'à obtenir un état stable. Ces réseaux sont particulièrement adaptés à la reconnaissance de formes.

Un des grand avantages des SOM est qu'elles sont non supervisées. Le résultat de l'apprentissage sera donc le même que l'on dispose d'un échantillon d'apprentissage parfait, d'un échantillon d'apprentissage biaisé ou d'aucun échantillon d'apprentissage. Se pose, en revanche, le problème de la classification à partir d'une SOM. Dans notre cas, les SOM ont surtout servis d'outils d'investigations.

Nous n'avons pas utilisé les LVQ car selon Zhang & Zhao (2004) les SVM (support vector machines) sont plus efficaces. Il en est de même pour les perceptrons multicouches.

2.2 Les cartes auto-organisatrices de Kohonen

Comme nous venons de le dire, les cartes auto-organisatrices de Kohonen (Kohonen 1988), ou SOM pour *Self-Organizing Map*, forment une classe particulière de réseaux de neurones . Une carte peut-être de dimension quelconque mais, pour des raisons de facilité de visualisation, elles sont généralement de dimension 2. Une carte représente une projection sur un nombre limité de dimension – souvent 2 – de la structure des données dans l'espace d'entrée, dont la dimension peut être très grande.

Pour une introduction un peu plus détaillée sur les réseaux de neurones et les cartes auto-organisatrices de Kohonen, nous renvoyons le lecteur aux chapitres 3 et 6 du manuscrit de thèse de Philippe Poinçot (Poinçot 1999) disponible sur internet[8].

Dans les sections suivantes, nous détaillons l'algorithme utilisé, et nous montrons les résultats sur un exemple de test et sur nos propres données.

Nous avons réalisé un logiciel qui réalise des cartes auto-organisatrices de Kohonen dont la documentation figure dans l'annexe B.

[7] `http://schwind.home.cern.ch/schwind/MLPfit.html`

[8] `http://astro.u-strasbg.fr/~fmurtagh/inform/these-philippe-poincot/index.html`

2.2.1 Description de l'algorithme

On suppose qu'on dispose de N neurones agencés spatialement selon une structure préétablie, hexagonale par exemple. On note $\{N\}$ cet ensemble de neurones. Le nombre N de neurones est choisi par l'utilisateur de manière empirique.

Entrée : 5 entrées :
- $\{N\}$: un ensemble de neurones dont les poids ont été initialisés aléatoirement.
- X : matrice de données. Elle contient n observations (lignes) de m paramètres (colonnes). Dans notre cas, n sources et leurs m caractéristiques. L'espace des observations est donc un espace de dimension m, ce qui signifie qu'à chaque neurone est attribué un vecteur de poids $\vec{w}$ de dimension m.
- $\alpha(t)$: une fonction d'apprentissage qui varie avec le nombre d'itérations accomplies t.
- $h(d,t)$: une fonction de voisinage qui varie en fonction de la distance d entre deux neurones et du nombre d'itérations accomplies t.
- T : le nombre d'itérations total fixé par l'utilisateur.

Étape 1 : initialisation :

Étape 1-1 : on centre et, éventuellement, on réduit les données :

$$\forall p \in [1,m], \forall i \in [1,n] : x_{ip} \leftarrow \frac{x_{ip} - \mu_p}{\sigma_p}$$

Étape 1-2 : on calcule les minimums et les maximums de chaque paramètre.

Étape 1-3 : pour chaque neurone du réseau et pour chaque paramètre d'un neurone, on génère des poids w_{ip} aléatoires uniformément répartis entre le maximum et le minimum du paramètre p.

Étape 2 : entraînement. Tant que le nombre d'itérations (ou epoch) $t \leq T$ faire :

Étape 2-1 : tant qu'il reste des observations (ou stimuli) :

Étape 2-2-1 : choisir aléatoirement une observation (ou stimulus) $\vec{x}_s$ parmi celles restantes

Étape 2-1-2 : déterminer le neurone gagnant, c'est-à-dire le neurone win dont le vecteur des poids $\vec{w}_{win}$ est le plus proche du stimuli $\vec{x}_s$.

Étape 2-1-3 : pour le neurone gagnant et tous les neurones de son voisinage, mettre à jour les poids w_{ip} :

$$w_{ip} \leftarrow w_{ip} + \alpha(t)h(d,t)(x_{sp} - w_{ip})$$

Où $d = ||\vec{w}_{win} - \vec{w}_i||$, distance entre le neurone gagnant win et le neurone voisin i.

Étape 2-1-3 : retirer l'observation de la liste des observations non testées.

Étape 2-2 : incrémenter t.

Sortie : $\{N\}$: l'ensemble de neurones d'entrée dont les poids ont été fixés par l'algorithme d'apprentissage.

Remarque : dans l'étape 1, on réduit éventuellement les données. L'intérêt est d'avoir des données sans dimension, ce qui permet de calculer la distance euclidienne dans un repère orthonormé.

2.2.2 Les traceurs T1 et T2

P. Poinçot définit dans sa thèse (paragraphe 6.2.1) deux traceurs. Nous les avons utilisé, comme lui, pour contrôler l'apprentissage du réseau. Ces deux traceurs sont :
- T1 : qui permet de contrôler l'homogénéité des classes
- T2 : qui permet de contrôler l'organisation de la carte

Soit $\vec{g}^{[j]}$ le centre de gravité, dans l'espace des paramètres, des $n^{[j]}$ observations associées au neurone j :

$$\vec{g}^{[j]} = \frac{1}{n^{[j]}} \sum_{i=1}^{n^{[j]}} \vec{x_i^{[j]}}$$

P. Poinçot utilise la quantité :

$$\frac{1}{n^{[j]}} \sum_{i=1}^{n^{[j]}} ||\vec{x_i}^{[j]} - \vec{g}^{[j]}||^2$$

pour quantifier la qualité d'une classe. Cette valeur n'est autre que la variance de la distance au centre de gravité $\vec{g}^{[j]}$ des observations associées au neurone j.
Le traceur $T1$ consiste en la moyenne de ces variances sur l'ensemble des N neurones :

$$T1 = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{n^{[j]}} \sum_{i=1}^{n^{[j]}} ||\vec{x_i}^{[j]} - \vec{g}^{[j]}||^2 \tag{III.40}$$

"*Une valeur faible de T1 indique une bonne homogénéité des classes*" (P. Poinçot). *Classe* désigne ici le type d'objet tombant, dans l'espace des paramètres, dans la zone délimitée par un neurone. Cette zone est définie par la cellule de Voronoï du neurone dans l'espace des paramètres. Des observations proches dans l'espace des paramètres ont des classes proches.

Le traceur T2 est défini selon la formule (différente mais équivalente à la formule de P. Poinçot) :

$$T2 = \frac{1}{\sum_{j=1}^{N} n^{[j]}} \sum_{j=1}^{N} \sum_{i=1}^{n^{[j]}} \frac{1}{\sum_{k=1}^{8} n^{[j,k]}} \sum_{k=1}^{8} n^{[j,k]} ||\vec{x_i}^{[j]} - \vec{g}^{[j,k]}||^2 \tag{III.41}$$

avec :
- $n^{[j,k]}$: le nombre d'observations associées au $k^{\text{ème}}$ neurone du voisinage immédiat du neurone j ($k \in [1, 8]$)
- $\vec{g}^{[j,k]}$: le centre de gravité des observations associées au $k^{\text{ème}}$ neurone du voisinage immédiat du neurone j

Autrement dit, T2 représente la moyenne sur toutes les observations de la moyenne pondérée des carrés des distances entre une observation i et les centres de gravité des neurones du voisinage du neurone associé à l'observation i.
« *Une valeur faible de T2 indique des classes voisines proches* » (P. Poinçot). *Classe* désigne, là aussi, le type d'objet tombant dans la cellule de Voronoi du neurone dans l'espace des paramètres.

2.2.3 Les fonctions d'apprentissage et de voisinage

Les fonctions de voisinage ($h(d,t)$) et d'apprentissage ($\alpha(t)$) utilisées dans l'algorithme peuvent avoir plusieurs formes différentes.

Voici quelques exemples de fonctions d'apprentissage courantes :

$$\begin{aligned}
\alpha(t) &= \alpha_0 \\
\alpha(t) &= \frac{\alpha_0}{1+t} \\
\alpha(t) &= \frac{\alpha_0}{1+t/\tau} \\
\alpha(t) &= \alpha_0 (\frac{\alpha_0}{\alpha_{t_{max}}})^{\frac{t}{t_{max}}}
\end{aligned}$$

Pour une discussion sur les performances relatives de ces différentes fonctions, voir paragraphe 3.4.2.2 de Poinçot (1999).
Nous avons quasi exclusivement utilisé la dernière formulation, proposée par Ritter, avec les coefficients suivants :
- $\alpha_0 = 0.9$

– $\alpha_{t_{max}} = 0.0005$
– $t_{max} \in [50, 100]$

Voici quelques exemples de fonctions de voisinage courantes :

$$\begin{aligned}
\text{Rectangulaire :} \quad & h(d,t) = && 1 \text{ si } d \leq d_{max} \text{ ; } 0 \text{ sinon} \\
\text{Gaussienne :} \quad & h(d,t) = && c_1 . \exp\left(-c_2 . d^2\right) \\
\text{Chapeau mexicain :} \quad & h(d,t) = && c_1 . \exp\left(-c_2 . d^2\right) - c_3 . \exp\left(-c_4 . d^2\right) \\
\text{Conique :} \quad & h(d,t) = && \frac{R(t) - r}{R(t)} \text{ si } r \leq R(t) \text{ ; } 0 \text{ sinon} \\
& \text{avec :} && R(t) = R_0(1 - 2\frac{t}{t_{max}})
\end{aligned}$$

Nous avons quasi exclusivement utilisé la fonction de voisinage conique avec $R_0 = 5$.

2.2.4 Exemple test

Pour illustrer le fonctionnement des SOM, nous présentons ici un exemple très simple, un *toy example*. Nous avons généré trois distributions gaussiennes 3D. A chacune nous avons associé un type d'objet et nous avons appliqué notre algorithme via le logiciel que nous avons développé :

```
java -cp ../extern/libs/kdtree.jar:../java/classes kohonen.Network
        3 toy.csv 0 Matrix2D 60x60 adaptativeRand 50
        RITTER "0.9 0.0005" CONE "5" toy.kohonen 0.1 0.7 -t1t2
```

Les résultats sont présentés sur les figures Fig. III.5, Fig. III.6 et Fig. III.7. Sur les figures III.6 et III.7, chaque pixel est un neurone. Les images sont de taille 60x60. Elles contiennent donc 3600 neurones. Les positions *pos_1* et *pos_2* sont les coordonnées des neurones sur la carte de Kohonen 2D.

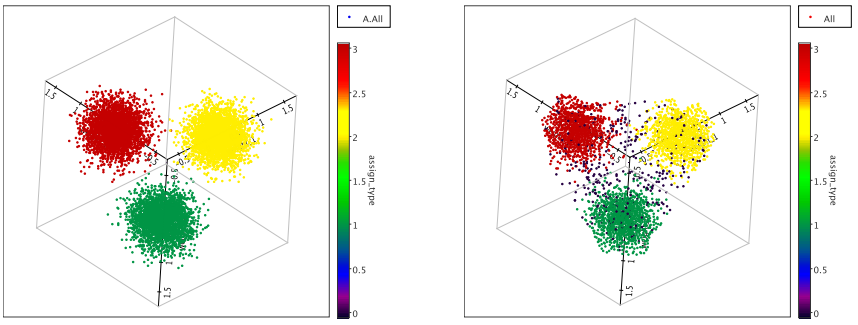


FIG. III.5 – Gauche : distributions Gaussiennes 3D générées, présentées en entrée ; la couleur code la classe associée à chaque point. Droite : positionnement des neurones dans l'espace des paramètres ; la couleur code la classe attribuée à chaque neurone (0=aucune classe attribuée).

2.2.5 Quelques résultats et exemples d'utilisation

Les SOM nous ont permis :
– d'isoler presque 200 sources X issues d'un même reste de supernova dans trois observations distinctes plus 200 autres SNR : ces sources, de caractéristiques spectrales similaires, formaient un îlot visible dans la carte de Kohonen issue de l'application de l'algorithme présenté sur les spectres EPIC ;

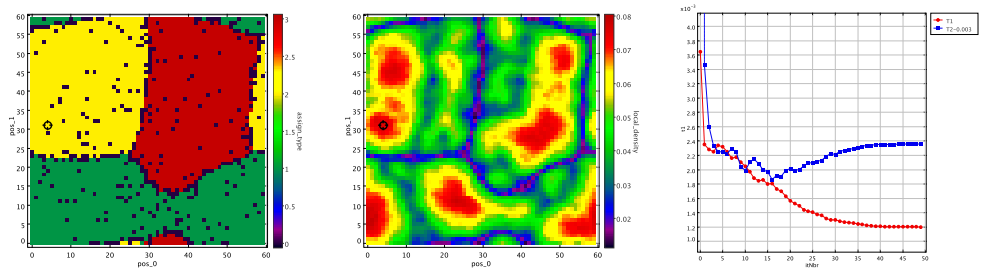


FIG. III.6 – Gauche : carte de Kohonen ; la couleur code la classe attribuée à chaque neurone (0=aucune classe attribuée). Centre : carte de Kohonen ; la couleur code la densité locale de point dans l'espace des paramètres. Droite : graphique représentant l'évolution au fil des itérations de marqueurs T1 et T2.

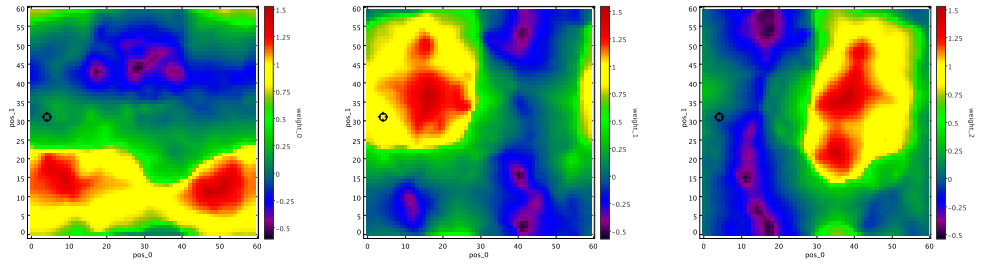


FIG. III.7 – Gauche : carte de Kohonen ; la couleur code la valeur selon l'axe des x. Centre : carte de Kohonen ; la couleur code la valeur selon l'axe des y. Droite : carte de Kohonen ; la couleur code la valeur selon l'axe des z.

– de retrouver rapidement de mauvaises identifications : visuellement, il est facile de détecter une source associée à un neurone dans une zone de la carte présentant des types d'objets différents du type de la source.

a Les SNR dans les spectres XMM

On a appliqué notre réseau de Kohonen sur les spectres de la caméra EPIC. Les spectres ont été « binés » puis normalisés, de sorte que la valeur moyenne soit égale à 1. Chaque bin contient une dizaine de canaux. Les résultats sont présentés figure Fig. III.8.

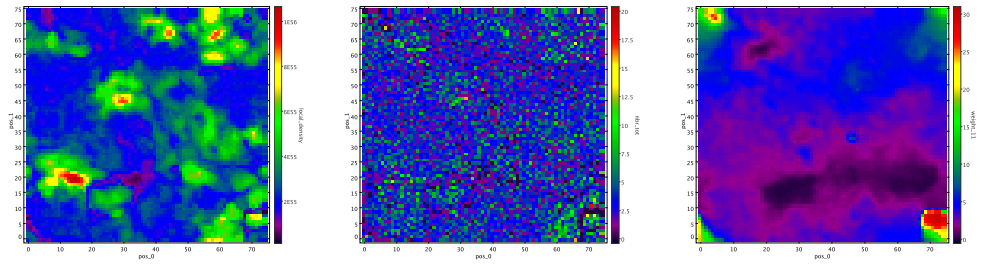


FIG. III.8 – Carte de Kohonen ayant pour entrée les spectres EPIC « binés ». Gauche : la couleur code la densité locale dans l'espace des paramètres. Centre : la couleur code le nombre d'entrées associé à chaque neurone. Droite : la couleur code la valeur selon le 11^{e} bin.

On voit que la région en bas à droite de la carte est une zone particulière : elle est entourée de neurones associés à aucune source d'entrée et elle correspond à une légère surdensité dans l'espace des paramètres. Il doit donc s'agir d'un groupe d'objets isolés. En sélectionnant les en-

trées associées aux neurones du centre de l'ilôt, on s'est rendu compte qu'elles correspondaient toutes à des sources différentes détectées dans deux observations du même reste de supernovae, la supernova de Tycho. Voici la liste des numéros (*obs_id*) de ces deux observations : 0310590201 et 0511180101.
La caractéristique principale de ces entrées, comme on peut le voir à droite de la Fig. III.8, est une valeur importante dans le 11$^{\text{ème}}$ bin.
En sélectionnant toutes les sources ayant des valeurs importantes dans le 11$^{\text{ème}}$ bin (couleurs rouge à verte sur la carte de droite), on trouve d'autres restes de supernovae. Seule une source diffère, mais elle a des caractéristiques X très proches : le système binaire Wolf-Rayet WR 147 décrit dans la littérature (Skinner *et al.* 2007) et qui met en jeu des collisions de gaz émis par les deux étoiles compagnons.
Grâce à cette analyse, on a pu extraire 450 sources du 2XMMi dont les spectres sont ceux de SNR.

b Visualisation de mauvaises identifications

Les pixels proches sur la carte de Kohonen sont proches dans l'espace des paramètres. Les objets de même type ont donc tendance à être proches sur la carte de Kohonen. On visualise sur la carte les lieux qu'occupent les différents types d'objets en codant la couleur des pixels en fonction du type d'objets qui domine dans chaque pixel. En affichant les pixels qui contiennent au moins un type de source et en les comparant au résultat précédent, on repère facilement les objets qui ont de fortes probabilités d'être de mauvaises identifications.

3 Les k-means

L'algorithme des k-moyennes (ou *k-means* en anglais) est un algorithme de classification non-supervisée. Son principe est très proche des SOM. Il peut même être vu comme un cas particulier de ces dernières pour lequel :

- la fonction de voisinage $h(d,t)$ ne dépend pas du temps, vaut 1 pour $d = 0$ et 0 ailleurs : $h(d,t) = h(d) = 1$ si $d = 0$, 0 sinon (=> réseau de type WTA pour *winner-takes-all*)
- le coefficient d'apprentissage n'est pas constant mais dépend du neurone i ($\alpha_i(t)$) et est modifié selon la loi $\alpha_i(t) = 1/t_i$ (décroissance hyperbolique) où t_i est le nombre de fois où le neurone i a été déclaré vainqueur
- parmi les deux algorithmes possibles pour les SOM, celui utilisé est l'algorithme adaptatif ($w_i(t+1) = w_i(t) + \alpha_i(t)(e_i - w_i(t))$) et non le *batch* (cf. Poinçot (1999) p. 31)

Avec l'algorithme des *k-means*, la valeur finale du neurone i est la moyenne de tous les points dont il a été le neurone vainqueur :

$$\begin{aligned} w_i(1) &= w_i(0) + \frac{1}{1}(e_1 - w_i(0)) = e_1 \\ w_i(2) &= e_1 + \frac{1}{2}(e_2 - e_1) \\ w_i(t) &= w_i(t-1) + \frac{1}{t}(e_t - w_i(t-1)) \\ &= \frac{e_1 + e_2 + ... + e_t}{t} \end{aligned}$$

D'où le nom de k-moyennes.

Contrairement aux SOM dont le but principal est d'offrir une représentation visuelle en 2D d'un nuage de points dans un espace de dimension potentiellement très élevé, les *k-means* ont pour but de trouver les N centres de gravité, où N est le nombre de neurones en entrée, de N classes d'objets supposées. Le nombre de neurones mis en jeu est donc beaucoup plus faible que pour les SOM. Le seul paramètre d'entrée, N, est le nombre de classes supposées être présentes dans le nuage de points.

On voit tout de suite qu'une des limites de cette méthode est de supposer que tous les éléments les plus proches du centre d'une classe appartiennent à cette classe, puisque le critère qui décide

du neurone vainqueur est basé sur la distance euclidienne. Ce genre de méthode n'est donc pas adapté si on à des distributions complexes, en forme de bananes par exemple.

Notons tout de même (Malyszko & Wierzchon 2007) que des critères, tel que l'index de Davies-Bouldin, permettent d'évaluer la qualité des groupes réalisés par l'algorithme des *k-means*. Les *k-means*, et plus généralement les SOM sont implémentés dans Matlab dans les package *somtoolbox* (Vesanto *et al.* 2000).

En raison des limitations mentionnées ci-dessus, on préférera aux *k-means* un algorithme basé sur le *mean shift* (voir section suivante).

4 Le *mean shift*

La méthode *mean shift* était très populaire avant que la mode ne passe aux SVM. Elle permet de trouver les « modes », c'est-à-dire les maximums de densité de probabilité, dans une distribution quelconque de points. Le principe est une fois de plus assez simple et, une fois encore, basé sur le kernel smoothing. Cette méthode est non supervisée.

A partir d'une distribution X de points en dimension p, on calcule la densité de probabilité correspondante par un kernel smoothing et on cherche les maximums locaux de cette distribution. Pour cela on réalise une montée de gradient avec pour points de départ chacun des points x_i de la distribution X. Le vecteur de déplacement, appelé *mean shift* est proportionnel au gradient local. Ce gradient local est calculé en dérivant $\hat{f}(\vec{x})$, l'estimation de la densité de probabilité locale obtenue par le kernel smoothing (voir formule Eq. II.71 dans le paragraphe traitant du kernel smoothing) :

$$\hat{f}(\vec{x}) = \frac{1}{Nh^d} \sum_{i=1}^{N} K(||\frac{\vec{x} - \vec{x}_i}{h}||^2) \tag{III.42}$$

La dérivée de $\hat{f}(\vec{x})$ s'écrit :

$$\hat{\Delta}\hat{f}(\vec{x}) = \frac{2}{Nh^{d+2}} \sum_{i=1}^{N} (\vec{x} - \vec{x}_i) K'(\frac{\vec{x} - \vec{x}_i}{h}) \tag{III.43}$$

où la fonction noyau $K(\vec{x})$ est appelée ombre (*shadow*) de la fonction noyau $G(\vec{x}) = -K'(\vec{x})$. Le noyau d'Epanechnikov est l'ombre du noyau uniforme, à savoir l'hypersphère de dimension d, alors que l'ombre du noyau Gaussien est une Gaussienne de même expression (Comaniciu & Meer 2002). Avec cette nouvelle notation et en développant la formule Eq. III.43, on peut écrire :

$$\hat{\Delta}\hat{f}(\vec{x}) = \frac{2}{Nh^{d+2}} \left[\sum_{i=1}^{N} G(\frac{\vec{x} - \vec{x}_i}{h}) \right] \left[\frac{\sum_{i=1}^{N} \vec{x}_i G(\frac{\vec{x} - \vec{x}_i}{h})}{\sum_{i=1}^{N} G(\frac{\vec{x} - \vec{x}_i}{h})} - \vec{x} \right] \tag{III.44}$$

Le deuxième terme correspond au *mean shift* :

$$\vec{m}_G(\vec{x}) = \frac{\sum_{i=1}^{N} \vec{x}_i G(\frac{\vec{x} - \vec{x}_i}{h})}{\sum_{i=1}^{N} G(\frac{\vec{x} - \vec{x}_i}{h})} - \vec{x} \tag{III.45}$$

Il représente la différence entre la moyenne des $\vec{x}_i$, pondérés par la fonction noyau G, et $\vec{x}$. Si on estime la densité locale à l'aide du noyau d'Epanechnikov, comme nous avons choisi de le faire systématiquement, le *mean shift* s'exprime simplement par :

$$\vec{m}_G(\vec{x}) = \frac{1}{n_{h,\vec{x}}} \left[\sum_{i=1}^{n_{h,\vec{x}}} \vec{x}_i \right] - \vec{x} \tag{III.46}$$

où $n_{h,\vec{x}}$ est le nombre de sources se trouvant dans l'hypersphère de rayon h autour de $\vec{x}$. Dans ce cas particulier le vecteur de *mean shift* se résume au vecteur allant de $\vec{x}$ au centre de gravité des vecteurs $\vec{x}_i$ se trouvant dans l'hypersphère de rayon h.

La montée de gradient s'effectue par déplacement de proche en proche où le vecteur à la $k^{\text{ème}}$ itération, $\vec{x}_k$, vaut :

$$\vec{x}_k = \vec{x}_{k-1} + \vec{m}_G(\vec{x}_{k-1}) = \frac{1}{n_{h,\vec{x}_{k-1}}} \sum_{i=1}^{n_{h,\vec{x}_{k-1}}} \vec{x}_i \tag{III.47}$$

Autrement dit le vecteur $\vec{x}_k$ est simplement le centre de gravité des points se trouvant dans l'hypersphère de rayon h autour de $\vec{x}_{k-1}$. En pratique, on pose une contrainte sur la valeur minimum du vecteur de *mean shift* pour arrêter l'itération. Le point final correspond à un maximum local.

Concernant la classification, on effectue une montée de gradient par les *mean shifts* à partir de chaque source. Les sources arrivant au même maximum local sont associées à une même classe. L'échantillon d'apprentissage sert alors à associer un type d'objet à chacun de ces maximums locaux, de ces "modes".

L'avantage de cette méthode, comparée par exemple au k-means, est qu'elle permet de prendre en compte des géométries de distributions complexes, en forme de banane par exemple. Dans ces cas, le critère de distance au centre du locus utilisé par les k-means n'est pas pertinent pour les sources situées aux extrémités de la banane.

5 Les arbres de décision

Les arbres de décision sont des algorithmes de classification supervisée qui séparent récursivement l'espace des paramètres, et les sous-espaces successifs, par des hyperplans. Dans le cas des arbres de décision classiques, chaque hyperplan séparateur est normal à un des axes de l'espace des paramètres. Dans le cas des arbres de décision obliques cette contrainte est relâchée, ce qui est conceptuellement plus élégant, présente de meilleurs résultats, mais complexifie les algorithmes. A chaque étape, le choix du meilleur plan séparateur se fait par maximisation d'un critère de pureté. Dans l'article Zhao & Zhang (2008), un certain nombre d'arbres de décision classiques ont été testés et succintement comparés. Contrairement à ces arbres, l'algorithme OC1 (Murthy *et al.* 1994) est un arbre de décision oblique. Dans les (très) grandes lignes :

- les coefficients d'un hyperplan sont perturbés et on garde ceux dont le plan résultant minimise un paramètre d'impureté
- des valeurs aléatoires sont utilisées pour éviter de tomber dans des minimums locaux
- le système d'élagage (*pruning*) utilisé est un système existant pour des arbres classiques

White (2008) fournit des exemples d'applications de l'algorithme OC1. Cet algorithme était au coeur du projet *ClassX* (McGlynn *et al.* 2004).

Au sujet des arbres de décisions, notons que, comme les SVM ou les réseaux de neurones, ils permettent d'établir des règles de classification mais sans probabilités associées.

6 Les Support Vector Machines

Les séparateurs à vaste marge ou machines à vecteurs de support (SVM pour *Support Vector Machines*), voir par exemple p. 359 de Tufféry (2007), p. 456 de Saporta (2006) ou p. 883 de Press *et al.* (2007), forment une classe d'algorithmes d'apprentissage et permettent de faire de la classification supervisée.

Leur principe conssite à rechercher un hyperplan optimal, c'est-à-dire un plan dans l'espace des paramètres qui sépare au mieux les données en deux catégories. La marge est la distance du point le plus proche à l'hyperplan. Les vecteurs de support sont les points par lesquels passent les deux plans qui définissent la marge. Ces deux plans sont parallèles à l'hyperplan et à égale distance de celui-ci (voir Fig. III.9). Les SVM utilisent le *kernel trick* pour projeter les données dans un espace de plus grande dimension (le *features space*) où elles ont plus de chance de pouvoir être séparées linéairement, c'est-à-dire par un hyperplan.

Dans la suite, on travaille avec n observations de m paramètres. Les paramètres peuvent être continus ou discrets. Les observations sont séparées en deux classes A et B auxquelles on associe

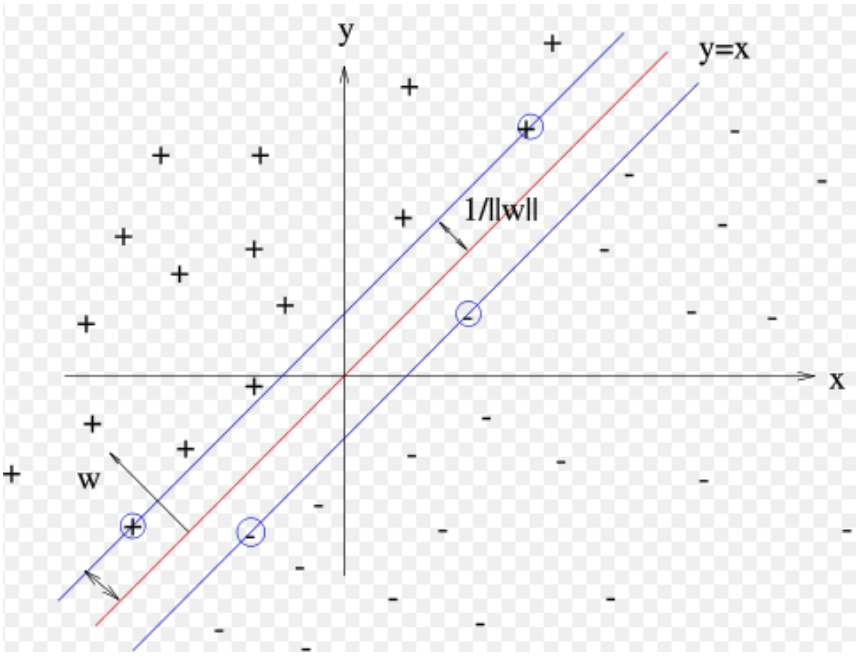


FIG. III.9 – Exemple de plan séparateur (rouge) en dimension 2. Les vecteurs de support sont les points entourés. (crédits wikipédia)

deux valeurs, respectivement +1 et -1. Chaque observation se voit donc associer une valeur y_i valant +1 ou -1.

6.1 CAS IDÉAL

Commençons par le cas idéal : les données peuvent être parfaitement séparées par un hyperplan défini par l'équation $f(\vec{x}) = 0$, où :

$$f(\vec{x}) = \vec{w}.\vec{x} + b \tag{III.48}$$

Avec $\vec{w}$ un vecteur normal à l'hyperplan et b un offset. Autrement dit, toutes les observations de type A sont d'un côté ($f(\vec{x}^A) > 0$) du plan et toutes les observations de type B sont de l'autre ($f(\vec{x}^B) < 0$). La distance d'un vecteur $\vec{x}$ quelconque au plan est donnée par la formule :

$$d(\vec{x}) = \frac{|\vec{w}.\vec{x} + b|}{||\vec{w}||} = \frac{|f(\vec{x})|}{||\vec{w}||} \tag{III.49}$$

Pour que toutes nos observations soient bien classées, il faut que $\forall i \in [1, ..., n]$, $y_i f(\vec{x}_i) > 0$. Comme le couple $(\vec{w}, b)$ est défini à un coefficient multiplicatif c près $(c\vec{w}, cb)$, on peut imposer, en jouant sur c, que $\forall i \in [1, ..., n]$, $y_i f(\vec{x}_i) \geq 1$. Dans ce cas, la distance perpendiculaire entre les deux plans délimitant le *no man's land*, qui vaut deux fois la marge vaut $2/||\vec{w}|| = 2/\sqrt{\vec{w}.\vec{w}}$. On cherche à maximiser cette distance pour trouver l'hyperplan optimal, ce qui revient à résoudre le problème de minimisation sous contraintes :

$$\begin{cases} \min \frac{1}{2}\vec{w}.\vec{w} \\ \forall i \in [1, ..., n],\ y_i(\vec{w}.\vec{x}_i + b) \geq 1 \end{cases} \tag{III.50}$$

Ce type de problème est un problème d'optimisation bien connu en programmation linéaire (*linear programming*). Il est dit *primal* en programmation quadratique (*quadratic programming*) et on peut le résoudre en faisant appel à son dual (cf. Press *et al.* (2007) p. 886). La résolution du problème dual ne fait plus directement appel aux observations $\vec{x}_i$ mais, comme pour la kernel PCA, à la matrice des produits scalaires de ces observations $(\vec{x}_i.\vec{x}_j)_{i,j}$. De fait, on va pouvoir utiliser le *kernel trick* à cette étape de la résolution du problème.

6.2 Cas moins idéal

Dans le cas réel, les observations ne sont pas parfaitement séparées par un hyperplan. Il y a toujours quelques observations de chaque classe qui se trouvent du « mauvais » côté du plan.

Pour assouplir les contraintes du problème de minimisation précédent (Eq. III.50), des variables « ressorts »ξ_i sont introduites. Si l'observation i d'un certain type peut être séparée des observations de l'autre type par un hyperplan, $\xi_i = 0$. Sinon $\xi_i > 0$ et représente un écart selon la formule :

$$\forall i \in [1, ..., n],\ y_i(\vec{w}.\vec{x}_i + b) \geq 1 - \xi_i \tag{III.51}$$

On cherche alors à obtenir les ξ_i les plus petits possibles avec le plus de ξ_i nuls possibles. D'un autre côté, on cherche à avoir la marge la plus grande possible. Toute la difficulté consiste à trouver un bon compromis entre ces deux aspects. Pour cela on pénalise la minimisation par les dépassements ξ_i des contraintes dans l'Eq. III.50, ce qui mène au nouveau problème de minimisation sous contraintes :

$$\begin{cases} \min \frac{1}{2}\vec{w}.\vec{w} + \lambda \sum_i \xi_i \\ \forall i \in [1, ..., n],\ y_i(\vec{w}.\vec{x}_i + b) \geq 1 - \xi_i \end{cases} \tag{III.52}$$

Ici, $\lambda \in \mathbb{R}^{*+}$ doit être défini par l'utilisateur. Au vu des formules, on voit bien que plus on choisit une valeur importante pour λ, plus on va chercher à minimiser les coefficients ξ_i au détriment de la taille de la marge. Et, inversement, plus λ sera petit, moins on se souciera des coefficients ξ_i et plus on cherchera à avoir une grande marge.
La résolution de ce problème de minimisation (Eq. III.52) est similaire au précédent (Eq. III.50) à peu de choses près (Press *et al.* 2007).

6.3 En pratique

Les SVM sont des algorithmes de classifications supervisées. La classification qu'elles fournissent est binaire. Dans le cas de plus de deux classes, il faut les utiliser itérativement. Au moins deux approches sont possibles :

- *one-versus-all* : on exécute l'algorithme autant de fois qu'il y a de classes et à chaque fois on oppose une classe à toutes les autres. Pour une nouvelle observation à classer, la classe attribuée est celle pour laquelle la marge est la plus grande.
- *one-versus-one* : on exécute l'algorithme pour chaque paire possible de classe qu'on oppose en ignorant les autres. Pour une nouvelle observation à classer, on comptabilise pour chaque paire possible la classe gagante. La classe qui a gagné le plus de fois est attribuée.

Pour un recensement des méthodes de classification dans le cas multi-classes, voir par exemple Rifkin & Klautau (2004).

Une des forces, mais aussi des faiblesses, de ces algorithmes est l'utilisation du *kernel trick*. Cet outil est puissant, mais trouver le « bon »kernel avec les « bons »paramètres peut se révéler difficile. Il n'y a pas, à l'heure actuelle, de critère permettant de choisir de façon optimale le paramètre λ. Enfin, les SVM ne fournissent pas de probabilités, juste des règles de classification.

6.4 Remarques

Au fond, les SVM peuvent être vus comme des arbres de décision obliques appliqués après projection des données dans un espace de plus grande dimension, et n'ayant qu'une seule règle de décision.
L'avantage des arbres de décision obliques, tel que l'algorithme OC1, par rapport aux SVM est la prise en compte naturelle de plus de deux classes.
Le principal avantage des SVM, outre l'élégance du problème de minimisation, est de permettre l'exploration d'espaces de projection (*features space*) très grands grâce au *kernel trick*.

7 Kernel density classification

La *kernel density classification* est un système de classification supervisée basé sur les probabilités Bayesiennes et le *kernel smoothing*. Le principe est simple, élégant, et pour chaque source on dispose en sortie des probabilités qu'elles appartiennent à chacune des classes.

7.1 PRINCIPE

On dispose en entrée d'un échantillon d'apprentissage contenant c classes d'objets distinctes.

A partir de cet échantillon d'apprentissage on peut estimer, via un kernel smoothing, les densités de probabilité $\hat{f}_i(\vec{x})$, $i \in [1, ..., c]$ de chacune de ces classes dans l'espace des paramètres. Pour une source de la classe i, la probabilité qu'elle se trouve à la position $\vec{x}$ dans l'espace des paramètres sachant qu'elle appartient à la classe i vaut directement $\hat{f}_i(\vec{x})$. On a donc :

$$p(\vec{x}|i) = \hat{f}_i(\vec{x})$$

On fixe les nombres N_i, $i \in [1, ..., c]$ de sources de chaque classe dans l'échantillon total, sources de l'échantillon d'apprentissage plus sources non-identifiées. En supposant que ces chiffres soient exacts, la densité de source de chaque classe dans l'espace des paramètres vaut $N_i \hat{f}_i(\vec{x})$. La probabilité qu'une source positionnée au point $\vec{x}$ soit de la classe i peut être défini comme le rapport en $\vec{x}$ de la densité locale de source de classe i sur la densité locale totale :

$$p(i|\vec{x}) = \frac{N_i \hat{f}_i(\vec{x})}{\sum_{j=1}^{c} N_j \hat{f}_j(\vec{x})}$$

Si on divise en haut et en bas par le nombre total de sources toutes classes confondues, $N_{tot} = \sum_{j=1}^{c} N_j$, et si on remplace la notation $\hat{f}_i(\vec{x})$ par $p(\vec{x}|i)$, on retrouve alors la formule Bayesienne classique à la base de la kernel density classification :

$$p(i|\vec{x}) = \frac{p(i)p(\vec{x}|i)}{\sum_{j=1}^{c} p(j)p(\vec{x}|j)} \tag{III.53}$$

7.2 PARAMÈTRES À FIXER

7.2.1 Les paramètres du kernel smoothing

Le calcul des densités de probabilité associées à chaque classe est basé, comme nous l'avons vu, sur le kernel smoothing. On peut choisir entre un kernel smoothing à largeur de bande fixe ou un *kernel smoothing* à largeur de bande variable. Pour l'instant, nous ne disposons d'un critère statistique pour estimer la largeur de bande optimale que dans le cas du kernel smoothing à largeur de bande fixe (5.5.2). On a donc choisi ce dernier avec la méthode MCMC sur l'estimateur *leave-one-out* pour déterminer la matrice H (Eq. II.72) de largeur de bande optimale.

7.2.2 Le nombre d'éléments de chaque classe

Comme vu précédemment, il faut fixer les nombres N_i, $i \in [1, ..., c]$ de sources de chaque classe dans l'échantillon total. Or ces nombres ne sont pas à priori connus. Plusieurs choix pour leurs valeurs initiales peuvent être faits :

- si on a de bonnes raisons de penser que les rapports entre les nombres de sources de chaque classe et le nombre total de sources dans l'échantillon d'apprentissage est représentatif de ces même rapports dans le résultat final, on peut directement les utiliser ;

– une inspection visuelle, quand c'est possible, du rapport entre sources de l'échantillon d'apprentissage et sources non identifiées dans les loci associés à chaque classe peut également servir de référence ;
– On peut également, avec plus ou moins de réussite, associer la même proportion de sources pour chaque classe.

Une fois la classification effectuée à partir de ces paramètres initiaux, on attribue à chaque source la classe où la probabilité $p(i|\vec{x})$ est la plus grande. On peut calculer le nombre de sources de chaque classe :

– soit en additionnant les sources associées à chaque classe (1 si $p(i|\vec{x}) = \max_j(p(j|\vec{x}))$, 0 sinon) ;
– soit en sommant sur toutes les sources les probabilités $p(i|\vec{x})$.

Nous avons choisi la deuxième solution, plus appropriée dans le cas où les recouvrements entre classes dans l'espace des paramètres peuvent être importants.

Si ces nombres sont différents de ceux choisis initialement, on recommence la classification en prenant ces nouveaux nombres en entrée. Après quelques itérations, les nombres de sources de chaque classe en entrée et en sortie doivent correspondre.

8 Conclusion

On vient de voir un certain nombre de techniques d'analyse de données multidimensionnelles, de classification supervisée et non supervisée.
Les techniques d'analyse de données multidimensionnelles, telles que l'ACP, la Kernel PCA ou encore les SOM ont, pour nous, un double intérêt :

– la réduction de la dimensionnalité ;
– l'inspection visuelle des données, des échantillons d'apprentissage et de la structure de ces données.

Concernant la classification, on peut mentionner que, selon Zhang & Zhao (2004), les résultats que l'on obtient avec des SVM sont légèrement meilleurs que pour les arbres de décision et les LVQ. Cependant, selon Gao *et al.* (2008), si on prend en ligne de compte les temps de calcul, le k-NN peut même s'avérer être plus intéressant que les SVM. Mais pour ces systèmes de classification supervisée, l'important n'est pas tant l'algorithme utilisé que la qualité de l'échantillon d'apprentissage. L'idéal serait de pouvoir se passer d'échantillons d'apprentissage. Malheureusement nos données ne sont pas suffisamment séparées pour que nous utilisions l'algorithme de *k-mean* ou le *mean shift*. Globalement, parce qu'il est le seul à fournir des probabilités d'identification, nous avons préféré dans la plupart des cas la *kernel density classification*. le tableau Tab. 8 récapitule les différentes techniques abordées dans ce chapitre et quelque-unes de leurs caractéristiques communes.

Méthode \ Propriétés	ana. visu.	classif.	robuste	non supervisé	proba	kernel trick
ACP	X			X		
ACP modifiée	X		/	X		
Kernel PCA	X			X		X
Kohonen SOM	X	X		X		
Kohonen LVQ	X	X	X		/	
k-means		X	/	X	/	
k-NN		X	X		/	
Kernel density classif		X	X		X	
Mean Shift		X	X	X	/	
SVM		X	X		/	X
Arbres de décision		X	X		/	



Références

BAZELL, D. & PENG, Y., 1998 ; *A Comparison of Neural Network Algorithms and Preprocessing Methods for Star-Galaxy Discrimination*. ApJS, **116**, 47

BRANHAM, R. L., 2001 ; *Astronomical data reduction with total least squares*. New Astronomy Review, **45**, 649

BRANHAM, JR., R. L., 1999 ; *A Covariance Matrix for Total Least Squares with Heteroscedastic Data*. AJ, **117**, 1942

CHEN, H., WANG, S. & JI, Y., 2006 ; *Robust PCA by Projection Pursuit and Mean Shift Analysis*. Dans *ISDA '06 : Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications*. IEEE Computer Society, Washington, DC, USA

COMANICIU, D. & MEER, P., 2002 ; *Mean Shift : A Robust Approach Toward Feature Space Analysis*. IEEE Trans. Pattern Anal. Mach. Intell., **24** (5), 603

CORTIGLIONI, F., MÄHÖNEN, P., HAKALA, P. & FRANTTI, T., 2001 ; *Automated Star-Galaxy Discrimination for Large Surveys*. ApJ, **556**, 937

CROUX, C., FILZMOSER, P. & OLIVEIRA, M. R., 2006 ; *Algorithms for Projection-Pursuit Robust Principal Component Analysis*. SSRN eLibrary

DHILLON, I. S., 1998 ; *A new O (N(2)) algorithm for the symmetric tridiagonal eigenvalue/eigenvector problem*. Thèse de doctorat, Berkeley, CA, USA

DOOLIN, D. M., DONGARRA, J. & SEYMOUR, K., 1999 ; *JLAPACK - compiling LAPACK Fortran to Java*. Sci. Program., **7** (2), 111

GAO, D., ZHANG, Y.-X. & ZHAO, Y.-H., 2008 ; *Support vector machines and kd-tree for separating quasars from large survey data bases*. MNRAS, **386**, 1417

GENTIANE, C. C., CROUX, C. & HAESBROECK, G., 2000 ; *Principal Component Analysis based on Robust Estimators of the Covariance or Correlation Matrix : Influence Functions and Efficiencies*. Biometrika, **87**, 603

HOPFIELD, J. J., 1982 ; *Neural Networks and Physical Systems with Emergent Collective Computational Abilities*. Proceedings of the National Academy of Science, **79**, 2554

HÖSSJER, O. & CROUX, C., 1995 ; *Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter*. J. Nonparametr. Statist., **4** (3), 293

KOHONEN, T., 1988 ; *Self-Organization and Associative Memory*. Self-Organization and Associative Memory, 100 figs. XV, 312 pages.. Springer-Verlag Berlin Heidelberg New York. Also Springer Series in Information Sciences, volume 8

KOHONEN, T., 1988 ; *Self-organized formation of topologically correct feature maps*, 509

KOHONEN, T. (réd.), 1997 ; *Self-organizing maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA

LETTVIN, J. Y., MATURANA, H. R., MCCULLOCH, W. S. & PITTS, W. H., 1959 ; *What the frog's eye tells the frog's brain*

MALYSZKO, D. & WIERZCHON, S. T., 2007 ; *Standard and Genetic k-means Clustering Techniques in Image Segmentation*. Dans *CISIM '07 : Proceedings of the 6th International Conference on Computer Information Systems and Industrial Management Applications*. IEEE Computer Society, Washington, DC, USA

MCCULLOCH, W. & PITTS, W., 1943 ; *A Logical Calculus of the Ideas Immanent in Nervous Activity*. Bulletin of Mathematical Biophysics, **5/115**

MCGLYNN, T. A., SUCHKOV, A. A., WINTER, E. L., HANISCH, R. J., WHITE, R. L., OCHSENBEIN, F., DERRIERE, S., VOGES, W., CORCORAN, M. F., DRAKE, S. A. & DONAHUE, M., 2004 ; *Automated Classification of ROSAT Sources Using Heterogeneous Multiwavelength Source Catalogs*. ApJ, **616**, 1284

MURTHY, S. K., KASIF, S. & SALZBERG, S., 1994 ; *A System for Induction of Oblique Decision Trees*. ArXiv Computer Science e-prints

PELAT, D., 1998 ; *Bases et Méthodes pour le Traitement des Données (Bruits et Signaux)*. Obspm, Paris

POINÇOT, P., 1999 ; *Classification et recherche d'information bibliographique par l'utilisation des cartes auto-organisatrices, applications en astronomie*. Thèse de doctorat, Strasbourg, France

POINCOT, P., LESTEVEN, S. & MURTAGH, F., 1998 ; *A spatial user interface to the astronomical literature*. A&AS, **130**, 183

PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T. & FLANNERY, B. P., 2007 ; *Numerical Recipes 3rd Edition : The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA

RIFKIN, R. & KLAUTAU, A., 2004 ; *In Defense of One-Vs-All Classification*. J. Mach. Learn. Res., **5**, 101

ROJAS, R., 1996 ; *Neural Networks — A Systematic Introduction*. Springer-Verlag, Berlin-New York. URL http://www.inf.fu-berlin.de/rojas/neural

SAPORTA, G., 2006 ; *Probabilités, analyse des données et statistique*. Technip, Paris

SCHÖLKOPF, B., SMOLA, A. & MÜLLER, K.-R., 1998 ; *Nonlinear component analysis as a kernel eigenvalue problem*. Neural Comput., **10** (5), 1299

SCHÖLKOPF, B., SMOLA, A. J. & MÜLLER, K.-R., 1999 ; *Kernel principal component analysis*, 327

SCHUERMANS, M., MARKOVSKY, I., WENTZELL, P. & HUFFEL, S. V., 2005 ; *On the equivalence between Total Least Squares and Maximum Likelihood PCA*. Analytica Chimica Acta, **544** (1-2), 254. URL http://eprints.ecs.soton.ac.uk/13302/

SKINNER, S. L., ZHEKOV, S. A., GÜDEL, M. & SCHMUTZ, W., 2007 ; *XMM-Newton X-ray observations of the Wolf-Rayet binary system WR 147*. MNRAS, **378**, 1491

TUFFÉRY, S., 2007 ; *Data Mining et statistique décisionnelle*. Technip, Paris

VARDI, Y. & ZHANG, C.-H., 2000 ; *The multivariate L1-median and associated data depth*. Proceedings of the National Academy of Science, **97**, 1423

VESANTO, J., HIMBERG, J., ALHONIEMI, E. & PARHANKANGAS, J., 2000 ; *Self-Organizing Map in Matlab : the SOM Toolbox*. Dans *In Proceedings of the Matlab DSP Conference*

WENTZELL, P. D., ANDREWS, D. T., HAMILTON, D. C., FABER, K. & KOWALSKI, B. R., 1997 ; *Maximum likelihood principal component analysis*. J. Chemometrics, **11**, 339

WHITE, R. L., 2008 ; *Astronomical Applications of Oblique Decision Trees*. Dans *American Institute of Physics Conference Series*, tm. 1082 de *American Institute of Physics Conference Series*

ZHANG, Y. & ZHAO, Y., 2004 ; *Automated clustering algorithms for classification of astronomical objects*. A&A, **422**, 1113

ZHAO, Y. & ZHANG, Y., 2008 ; *Comparison of decision tree methods for finding active objects*. Advances in Space Research, **41**, 1955

Chapitre IV

Résultats

Ce chapitre traite des résultats que nous avons obtenus grâce aux algorithmes présentés dans les chapitres précédents. Nous aborderons d'abord les résultats de la corrélation croisée entre les catalogues 2XMMi et le SDSS DR7. Ces résultats sont notamment utilisé pour la création d'un échantillon d'identification qui est à l'origine de tous les échantillons d'apprentissage que nous utilisons. Cet échantillon d'identification est en effet basé sur l'ACDS, les résultats de nos corrélations et une *kernel density classification* sur les données de la *data release 7* (DR7) du *Sloan Digital Sky Survey* (SDSS). Nous proposons dans le même section une discussion illustrant les intérêts scientifiques de cette corrélation. Nous verrons ensuite les résultats de la classification d'un jeu de données issu de la corrélation croisée du 2XMMi avec le 2MASS et le GSC2, puis la classification des sources du 2XMMi uniquement à partir de leurs rapports de dureté. Enfin, nous comparerons différents algorithmes de classification en les appliquant à plusieurs échantillons. Une grande partie de la sections 1 fait l'objet d'un papier en préparation qui est fourni dans l'annexe H.

1 Corrélation du 2XMMi avec le SDSS DR7

Nous avons réalisé la corrélation croisée du catalogue 2XMMi avec la *Data Release 7* (DR7) du *Sloan Digital Sky Survey* (SDSS) en suivant l'algorithme présenté au chapitre II. Les résultats ont été inclus dans la XCAT-DB (voir chapitre V) et sont par ce biais disponibles pour la communauté. Cette partie traite dans un premier temps des statistiques de cette corrélation avant d'aborder la réalisation de l'échantillon d'identifications pour les sources XMM. Nous verrons ensuite une analyse scientifique de ces résultats sous forme d'une discussion. Enfin, on verra la classification des sources issues de ces corrélations.

1.1 Résultats de la corrélation

1.1.1 Quelques chiffres

Un total de 1 337 champs de vue (FOV) XMM contiennent au moins une source ayant un candidat contrepartie dans le SDSS. Ces 1 337 FOV contiennent 73 636 sources 2XMMi uniques et un total de 95 452 détections. La corrélation croisée du catalogue 2XMMi avec le SDSS DR7, suivant l'algorithme décrit chapitre II, mène à 72 169 associations impliquant respectivement 45 727 et 55 726 sources XMM et SDSS uniques. Ce nombre total d'associations représente respectivement 20% et 62% des sources uniques disponibles dans tout le catalogue 2XMMi et dans les 1 337 FOV. La distribution du nombre de candidats SDSS par source unique XMM est donnée dans la table IV.1. La figure IV.1 présente la distribution des probabilités d'identification et les valeurs cumulées correspondantes sont fournies dans la table IV.2.

Environ 40% de l'ensemble des sources uniques disponibles dans le catalogue 2XMMi, et dans les 1337 FOV, sont identifiées avec une source DR7 et une probabilité d'au moins 90% . Parmi

toutes les sources XMM uniques ayant au moins un candidat, 66% ont une probabilité d'identification supérieure à 90%. Cette valeur atteint respectivement 72% et 90% si on considère les sources optiques situées à une distance de moins de 5″de la source X et plus brillantes que la magnitude 22.2.

TAB. IV.1 – Distribution du nombre total de sources SDSS DR7 candidates uniques par source 2XMMi unique.

Nbr Cand	1	2	3	4	5	6	7	8	>8
XMM Src	37 988	6059	1196	317	96	47	12	4	8

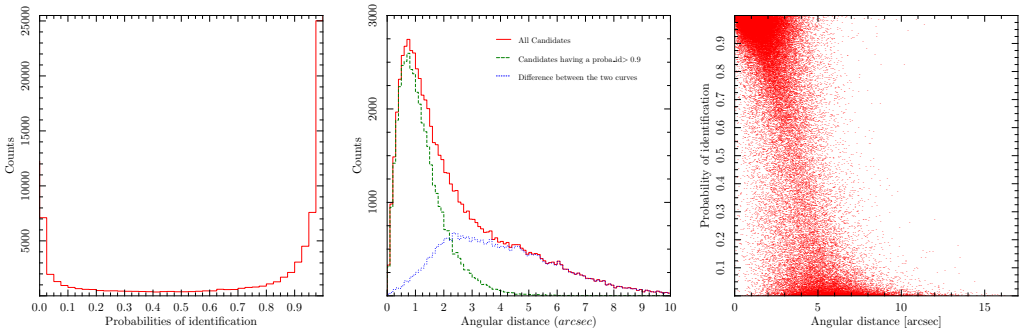


FIG. IV.1 – Gauche : distribution des probabilités d'identification. Centre : distribution des distances des sources XMM aux candidats SDSS. Droite : probabilités d'identifications versus distances des candidats.

TAB. IV.2 – Nombre d'associations avec les détections XMM (*det_id*) et les sources uniques (*src_id*) ayant une probabilité d'identification supérieure à un seuil donné. Si une source X a plusieurs candidats, on garde celui ayant la plus grande probabilité d'identification. Pour la fiabilité (R) et la complétude (C), on ne considère que le meilleur candidat pour chaque source X unique ayant ou moins un candidat.

	Seuil en probabilité d'identification					
id	0.0	0.5	0.7	0.8	0.9	0.95
# det_id	60 567	53 347	49 527	46 387	40 193	32 610
# src_id	45 727	39 839	36 943	34 605	30 055	24 327
R	79.4%	91.7%	94.7%	96.2%	98.0%	99.0%
C	100.0%	96.8%	92.1%	87.5%	77.2%	63.2%
Frac X	62.1%	54.1%	50.2%	47.0%	40.8%	33.0%

1.1.2 Ajustement d'une fonction de Rayleigh

La distance en sigma entre une source X et sa contrepartie correspond à la distance normalisée par rapport aux erreurs de position (Eq. II.5). La distribution de cette distance sur un grand nombre de paires source X–contrepartie suit normalement une distribution de Rayleigh dont la forme est donné §2.1.3, Eq. IV.2. La distribution de Rayleigh est l'expression de la composante radiale d'une Gaussienne en coordonnées polaires. Or la distance entre une source et sa contrepartie dépend uniquement des incertitudes gaussiennes sur les positions dont le produit de convolution est lui-même une gaussienne (voir chapitre II). En utilisant la distance en sigma, le changement de variable transforme ces gaussiennes en gaussienne unité. La distribution des distances en sigma de toutes les associations doit donc suivre une distribution de Rayleigh de paramètre $\sigma = 1$. Dans l'ensemble des associations, certaines sources optiques ne sont pas les contreparties des sources X.

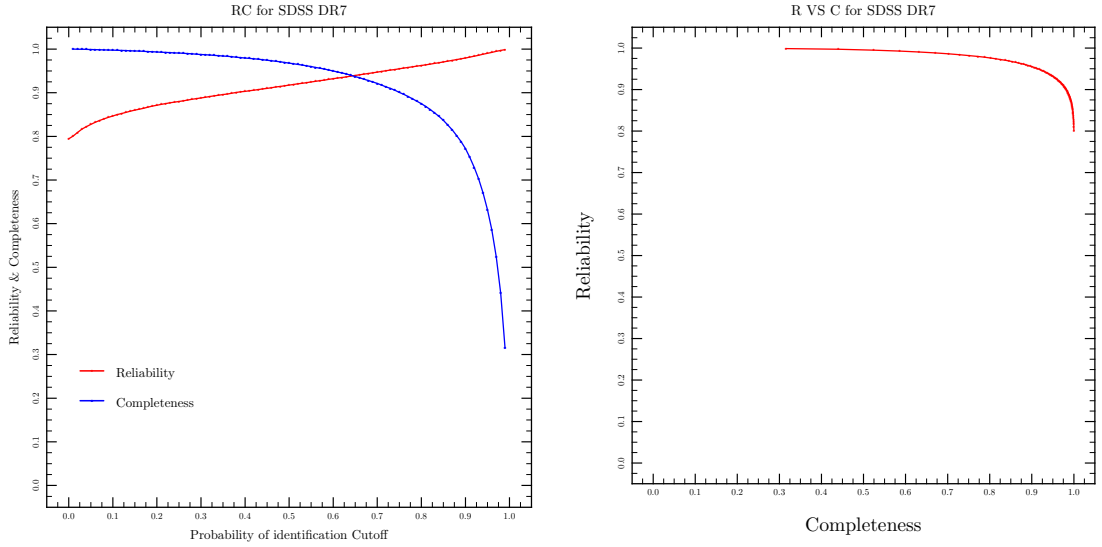


FIG. IV.2 – Gauche : pourcentage de fiabilité et de complétude en fonction du seuil en probabitité d'identification. Droite : diagramme pourcentage de fiabilité versus complétude.

Ces sources apportent une composante linéaire à la distribution de distance en sigma entre sources X et sources optiques.

Si on ajuste l'histogramme des distances en sigma entre toutes les sources XMM et leurs candidats associés par une distribution de Rayleigh plus une composante linéaire, on obtient $\sigma_r = 0.865$ pour le paramètre de la courbe de Rayleigh, 0.178 pour la pente de la droite et $R = 0.675$ pour le rapport entre le nombre de vraies associations et celui des fausses associations (voir Fig. IV.3). Ce résultat semble indiquer que l'erreur de positionnement des sources X tend à être légèrement surestimé. Si on sépare les FOV ayant une erreur systématique de 1.0″ et 0.35″ (voir §2.3.4, chapitre I), on obtient respectivement $\sigma_r = 1.013$, $R = 0.50$ et $\sigma_r = 0.856$, $R = 0.698$. Les erreurs sur σ_r, sur la pente de la droite et sur R sont toutes de l'ordre de 0.003. Ces valeurs semblent indiquer une surestimation des erreurs convoluées entre XMM et SDSS pour les champs de vue dont les positions sont corrigées par *eposcorr*. On ne connaît pas l'origine de cet effet, surtout que le système de calibration des positions SDSS n'est pas directement lié à l'USNOB1 utilisé dans *eposcorr* (pour *eposcorr*, voir le paragraphe 2.3.4 du chapitre I).
Notons enfin que si on choisit de ne garder pour chaque source X que le meilleur candidat, on obtient R=0.81, ce qui est proche de la valeur de 79.4% donné Tab. IV.2.

1.1.3 Calcul de la luminosité X

Enfin, pour tous les AGN ayant un redshift spectroscopique, nous avons calculé la luminosité X à l'aide de la formule IV.1 suivante :

$$L_{abs} = 4\pi F \frac{(1+z)^2}{(1+z)^{2-\gamma}} \left(\frac{c}{H_0} \int_0^Z \frac{1}{\sqrt{\Omega_m (1+z)^3 + \Omega_\Lambda}} dz \right)^2 \qquad \text{(IV.1)}$$

où F est le flux X reçu en erg.s^{-1}cm^{-2}, $H_0 = 73$ km.s^{-1}.Mpc^{-1}, $\Omega_m = 0.3$ et $\Omega_\Lambda = 0.7$. L'indice de photon γ a été choisi égal à 1.9.

1.2 CONSTITUTION D'UN ÉCHANTILLON D'APPRENTISSAGE

La constitution d'un échantillon d'apprentissage est indispensable pour plusieurs raisons. Il permet d'abord de faire ressortir dans l'ensemble des diagrammes les zones couvertes par les différents types d'objets et facilite donc l'analyse scientifique des résultats des corrélations. Mais,

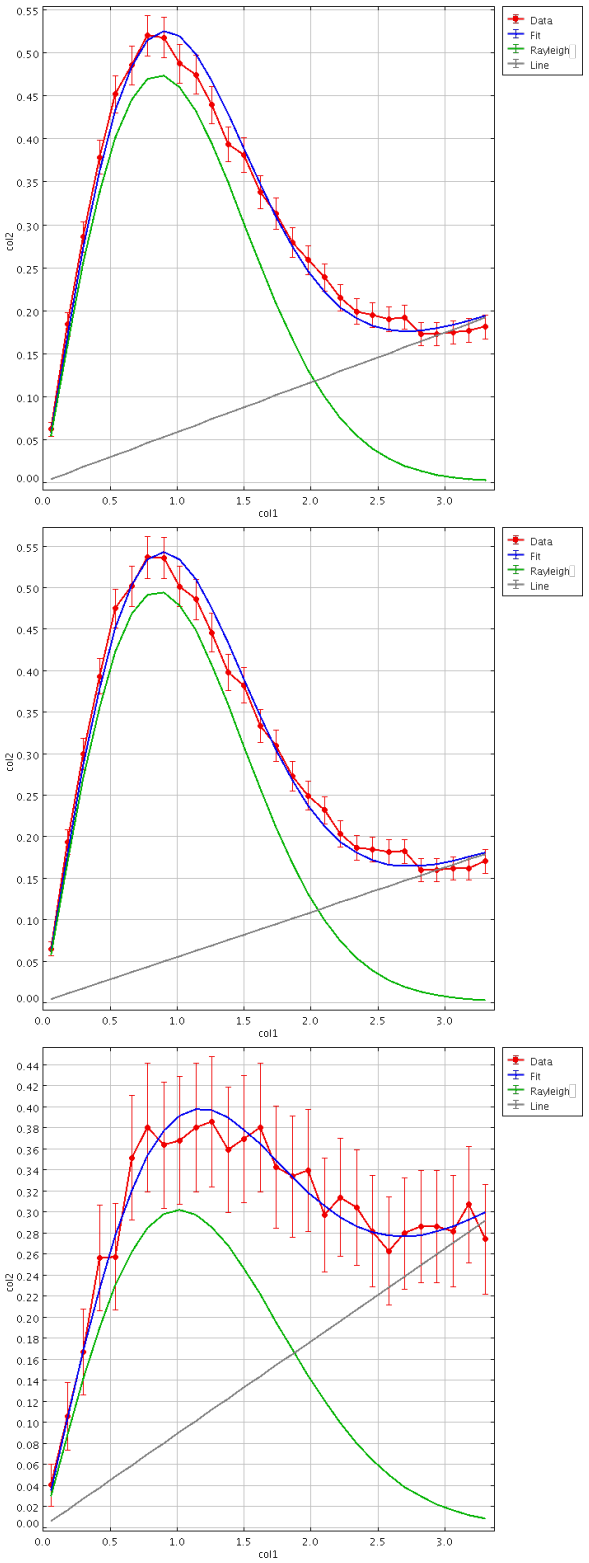


FIG. IV.3 – Ajustement d'une courbe de Rayleigh et d'une composante poissonienne sur l'histogramme des distance en sigma entre sources du 2XMMi et du SDSS DR7. Les barres d'erreur représentées correspondent à la valeur moyenne $\pm 3\sigma$. Haut : toutes les associations. Centre : erreurs systématiques valant 0.35. Bas : erreurs systématiques valant 1.0.

plus important, sans échantillon d'apprentissage, pas de classification supervisée. En effet, la particularité des algorithmes de classification supervisée est précisément la nécessité d'avoir en entrée un échantillon d'apprentissage.

En pratique, nous créons un échantillon d'identifications qui ne contient que deux colonnes : la première contient un identifiant XMM (*src_id*) ; la seconde un type d'objet. La jointure entre cet échantillon et un jeu de données crée un échantillon d'apprentissage. Par abus de langage, nous appellerons également échantillon d'apprentissage cet échantillon d'identifications.

L'idéal est de créer un échantillon d'apprentissage homogène (qui provient par exemple d'un même instrument) et qui couvre l'ensemble de la zone de l'espace des paramètres peuplée par les données sur lesquelles on travaille. Ce sont là deux des conditions indispensables, homogénéité et représentativité, qui, avec le non recouvrement des différents types d'objets, garantissent une classification performante. En pratique on ne peut cependant que rarement les réunir, les données disponibles comportant presque toujours des biais. Par exemple, on a souvent affaire à des types d'objets hétérogènes, les contraintes d'attribution d'une classe variant en fonction des données disponibles et des auteurs. En ce qui concerne la couverture de l'espace des paramètres, on peut prendre l'exemple des objets de faible luminosité. En effet, les objets les moins brillants en optique n'ont souvent pas de classe attribuée sur une base spectroscopique, ce qui biaise la représentation des différentes populations. Les données homogènes ne sont généralement disponibles que pour de petits ensembles de données. Nous avons donc choisi de sacrifier l'homogénéité des données pour obtenir un échantillon de taille conséquente recouvrant le plus possible nos différents jeux de données. Nous avons cependant rendu notre système de création d'échantillon d'apprentissage suffisamment souple pour nous permettre d'ajouter ou de retirer facilement différentes composantes.

Cette section traite de la réalisation de l'échantillon d'apprentissage global qui est utilisé dans l'analyse des corrélations et la classification du 2XMMi.

1.2.1 Constitution globlale

Nous avons essayé de baser notre échantillon d'apprentissage sur un grand ensemble cohérent d'identifications spectroscopiques. Seul le SDSS offre une telle opportunité. Nous nous sommes donc appuyés sur les classes d'objets qu'il propose, à savoir les étoiles, les galaxies et les QSO auxquelles nous avons ajouté les AGN et les binaires X. Les CV ont été incluses dans les binaires X. Ces cinq groupes forment les cinq grandes classes de base. Celles-ci sont subdivisées en sous-classes et , pour certaines, en sous-sous-classes quand on dispose d'informations supplémentaires. De nombreuses sources ont été ajoutées à l'échantillon d'apprentissage grâce à l'utilisation des corrélations de l'ACDS qui lient le catalogue XMM avec de nombreux catalogues spécifiques et avec Simbad. Nous avons également intégré quelques données de la littérature comme les quasars de type II (Padovani *et al.* 2004) dont une liste de candidats est fournie par Zakamska *et al.* (2003) et les CV de Dillon *et al.* (2008).

Sauf mention contraire, tous les objets des catalogues cités ci-après situés à moins de 3 secondes d'arc d'une source XMM ont été insérés dans l'échantillon d'apprentissage. Le choix de 3 secondes d'arc constitue un bon compromis entre la complétude et le taux de fausses associations (voir Fig. IV.1). Voici la liste des catalogues utilisés en fonction du type d'objet :

1. pour les étoiles, nous avons utilisé :
   - les entrées du SDSS DR7 dont la probabilité d'identification est $> 80\%$ et dont l'attribut specClass vaut 1 (STAR) ou 6 (STAR_LATE), ou dont la classe étoile a été attribuée après la *kernel density classification* (voir section suivante) ;
   - les entrées Simbad pour lesquelles le type d'objet vaut 'TTau*', 'pMS*', 'Pulsar', 'WR*', 'Em*', 'Be*' ou 'WD*' ;
   - les sources du catalogue Tycho 2 (Høg *et al.* 2000) ;
   - les sources du LSS, catalogue d'étoiles lumineuses de la partie sud de la Voie Lactée ;
   - le catalogue de pulsars de Taylor *et al.* (1993) ;
   - le catalogue d'étoiles de type Wolf Rayet (WR) de van der Hucht (2001) ;
   - le catalogue d'étoiles de type Be de Jaschek & Egret (1982) ;
   - le catalogue d'étoiles de type T Tauri de Krautter *et al.* (1997), constitué à partir des sources Rosat ;

- le catalogue de naines blanches (WD) de McCook & Sion (1999) ;
- le catalogue de pulsars de Fleming *et al.* (1996), constitué à partir des sources Rosat ;

2. pour les binaires X et les CV :
   - le catalogue de CV de Downes (Downes *et al.* 2001) ;
   - le catalogue de CV de Ritter (Ritter & Kolb 2003) ;
   - le catalogue de LMXB de Ritter (Ritter & Kolb 2003) ;
   - le catalogue de LMXB de Liu (Liu *et al.* 2001) ;
   - le catalogue de HMXB de Liu (Liu *et al.* 2000, 2006) ;
3. pour les Galaxies :
   - les sources du SDSS DR7 dont la probabilité d'identification est $> 80\%$ et dont l'attribut specClass vaut 2 (GALAXY) ;
   - le catalogue Uppsala (Lauberts 1982) corrélé à partir de l'extension optique des galaxies. Cependant, afin d'éviter de nombreuses fausses associations, nous avons uniquement sélectionné les galaxies dont l'axe majeur est inférieur à 5 minutes d'arc. Au delà de cette valeur, de nombreuses étoiles d'avant plan détectées en X se trouvent associées à des galaxies proches très étendues ;
   - les sources venant du catalogue de Véron (Véron-Cetty & Véron 2006) ayant un type spectral sp='H2'.
4. pour les AGN : les sources venant du catalogue de Véron (Véron-Cetty & Véron 2006) ayant un type spectral sp='S' , sp='Sx', sp='Sxx' ou vide mais ayant une classe associée cl='A'. Par conséquent, cet échantillon contient également des sources de type 'S3' qui correspondent aux LINER. Nous ferons également remarquer que le catalogue de Véron est une compilation issue de sources diverses. Il est donc par nature assez hétérogène.
5. pour les QSO :
   - les sources du SDSS DR7 dont la probabilité d'identification est $> 80\%$ et dont l'attribut specClass vaut 3 (QSO) ou 4 (HIZ_QSO) ;
   - les sources venant du catalogue de Véron (Véron-Cetty & Véron 2006) ayant un type spectral sp='HP' ou vide et une classe cl='Q'

Pour chacun de ces types d'objets, nous avons également ajouté les sources identifiées dans le cadre du XMM-Newton Medium Sensitivity Survey (XMS, Barcons *et al.* 2002, Carrera *et al.* 2007, Barcons *et al.* 2007) et du Bright Sources Survey (XBS ou BSS, Della Ceca *et al.* 2004), deux relevés réalisés par le SSC d'XMM. Pour le XMS, on considère les 'NELG' comme des AGN et les 'BLAGN' et 'BLLac' comme des QSO. Pour le XBS, les types 'AGN2' et 'elusiveAGN' ont été assimilés à la classe AGN et les 'AGN1' et 'BLLac' à la classe QSO.

TAB. IV.3 – Composition des étoiles de l'échantillon d'apprentissage en nombre de sources par catalogue d'origine.

Catalogue	Tycho2	DR7 KDC	Simbad	LSS	XBS	DR7	Taylor Puls.	VanDerHucht WR
Nombre	3229	531	491	88	53	43	22	19

XMS	Jaschek Be	Krautter T Tauri	McCook WD	Rosat Puls.
15	9	7	2	2

TAB. IV.4 – Composition des binaires X de l'échantillon d'apprentissage en nombre de sources par catalogue d'origine.

Catalogue	Downes CV	Ritter CV	Ritter LMXB	Liu LMXB	Liu HMXB
Nombre	152	110	39	28	21

Enfin, une source X associée à la fois avec une étoile et avec une binaire X est marquée comme binaire X. Nous avons appliqué cette même règle pour les couples étoile–AGN, étoile–QSO, binaire X–AGN, binaire X–QSO et QSO–AGN.

TAB. IV.5 – Composition des galaxies de l'échantillon d'apprentissage en nombre de sources par catalogue d'origine.

Catalogue	DR7	Uppsala	Véron	XMS	XBS
Nombre	965	667	11	10	3

TAB. IV.6 – Composition des AGN et des QSO de l'échantillon d'apprentissage en nombre de sources par catalogue d'origine.

Catalogue	Véron AGN	XBS AGN	XMS AGN	DR7 QSO	Véron QSO	XBS QSO	XMS QSO
Nombre	1369	50	47	1455	1661	220	194

La constitution finale de l'échantillon d'apprentissage, que nous nommons LS_X (pour *Learning Sample of XMM*) est présentée Tab. IV.7.

TAB. IV.7 – Composition de l'échantillon d'apprentissage en nombre de source de chaque type.

Type	étoile	binaires X	galaxies	AGN	QSO
Nombre	4085	214	1506	1330	2399

1.2.2 *Kernel density classification* pour les étoiles du SDSS

a Description du système mis en place

À partir des résultats de la corrélation croisée entre les catalogues 2XMMi et SDSS DR7 (voir section 1.1), nous avons créé un échantillon fiable de sources dont on connaît à la fois les propriétés X et optiques. On nomme cet échantillon XS (pour XMM/SDSS) ou XMM/DR7. Nous avons dérivé de l'échantillon d'apprentissage LS_X un sous-échantillon contenant uniquement les sources X incluses dans XS. Cette échantillon constitue l'échantillon d'apprentissage de XS, on le nomme par conséquent LS_XS.

Avant le travail presenté dans cette section, l'échantillon d'apprentissage LS_XS manquait de sources stellaires. Ces sources sont néanmoins indispensables pour la classification supervisée des sources de l'échantillon XMM/DR7 et, plus généralement pour l'analyse de ces corrélations (voir par exemple Fig. IV.11). Il y a au moins deux raisons qui expliquent le peu d'étoiles dans l'échantillon d'apprentissage LS_XS. Premièrement, la majorité des étoiles dans LS_X viennent du catalogue Tycho 2. Or les sources Tycho 2 sont des objets brillants avec des magnitudes optiques qui sont typiquement inférieures à 11 mag. Ces mêmes sources sont toutes saturées dans le SDSS et sont par conséquent exclues de l'échantillon XS. Deuxièmement, parmi les étoiles spectroscopiquement identifiées dans le SDSS, seulement 26 figurent dans XS.

Afin de constituer un échantillon d'étoiles peu brillantes en optique (magnitude supérieure à 15-16 mag), couvrant l'échantillon XS, nous avons dû réaliser une classification des sources SDSS sur la base de leurs couleurs optiques. Nous avons pour cela utilisé une *kernel density classification* (KDC, Richards *et al.* 2002, voir aussi chapitre III, §7) sur l'ensemble des candidats contreparties SDSS optiquement non résolues (type=6), de magnitudes plus brillantes que 22.2 dans toutes les bandes, et n'étant associées à aucun des *flags* BLENDED, EDGE, DEBLENDED_AS_MOVING ou PSF_FLUX_INTERP. Ce sous-échantillon (XS) représente 10 533 sources SDSS DR7 corrélées avec le 2XMMi. L'espace des paramètres des objets à classer, en entrée de la KDC, est uniquement constitué des quatre couleurs $u-g$, $g-r$, $r-i$ et $i-z$.

Nous avons besoin pour classer les sources du SDSS sur la base de leurs propriétés optiques d'un échantillon d'apprentissage spécifique. Nous le nommons LS_KDC_S. LS_KDC_S n'a besoin de contenir que deux classes : étoiles et sources extragalactiques. Étant donné qu'on ne travaille que sur les sources ponctuelles, cette deuxième classe contient quasi uniquement des quasars. On la nommera par conséquent « QSO ». LS_KDC_S a été construit indépendamment des corrélations avec le 2XMMi à partir de tous les objets de bonne qualité (non saturé, etc) spectroscopiquement

TAB. IV.8 – Résultat de l'auto-classification de l'échantillon d'apprentissage par *kernel density classification*

Initiale\Attribuée	Star	QSO
Star	96.89%	3.11%
QSO	1.62%	98.38%

identifiés dans la DR7 et récupérés via CasJob[1]. L'échantillon stellaire de LS_KDC_S contient 67 269 sources dont l'attribut SDSS specClass correspond aux étoiles (STAR ou STAR_LATE). Il contient donc également des naines blanches et des variables cataclysmiques. L'échantillon de sources extragalactiques de LS_KDC_S compte 75 248 sources dont l'attribut specClass correspond aux QSO (QSO) ou aux QSO à haut redshift (HIZ_QSO) ainsi que 253 dont l'attribut specClass correspond aux galaxies (GALAXY). Le faible nombre de galaxies justifie l'utilisation de QSO pour désigner cet échantillon.

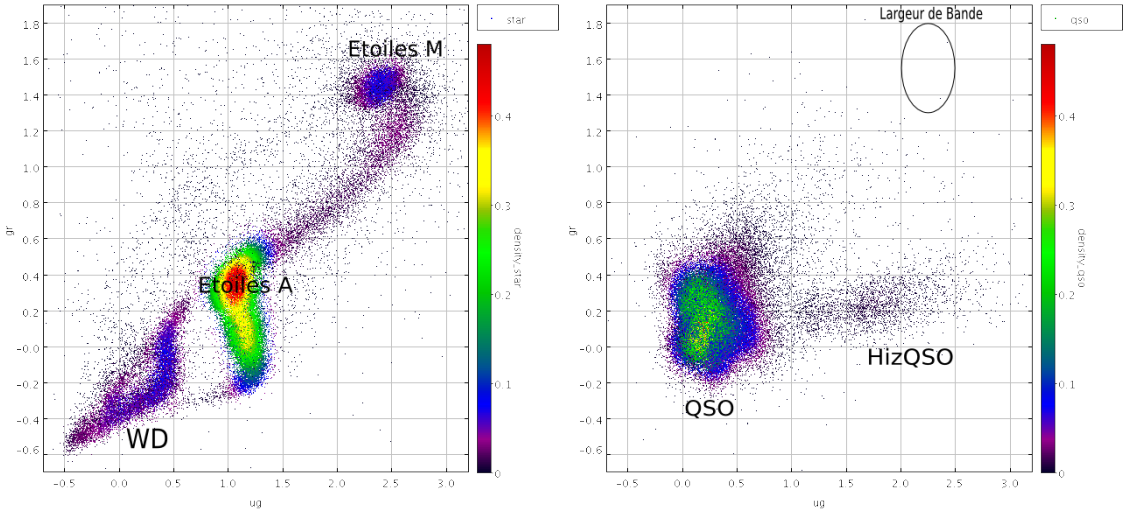


FIG. IV.4 – Étoiles (gauche) et QSO (droite) de l'échantillon d'apprentissage de la classification des sources SDSS dans le diagramme $u-g$ / $g-r$. La couleur code la densité de probabilité associée aux distributions de ces deux échantillons dans l'espace 4D des couleurs $u-g$, $g-r$, $r-i$ et $i-z$.

Les densités de probabilité ont été estimées à partir d'un *kernel smoothing* à largeur de bande fixe (voir chapitre II, §5.5.1, Eq. II.71) utilisant pour fonction noyau un profil d'Epanechnikov (voir Eq. II.80) de largeur valant 0.25 mag (voir Fig. IV.4). Celle valeur est proche de la moyenne des erreurs sur la couleur $u-g$.La table Tab. IV.8 liste les résultats de l'auto-classification de l'échantillon d'apprentissage, à savoir le résultat de la classification appliquée uniquement à l'échantillon d'apprentissage et dont les classes attribuées par l'algorithme sont comparées aux classes connues en entrée. On peut voir Fig. IV.5 que les sources mal classées sont essentiellement les naines blanches, confondues avec les quasars et le début des quasars de la branche des QSO à haut redshift, confondus avec les étoiles de type A notamment.

La probabilité à priori p(star) a été fixée à 0.25 et donc p(qso) à 0.75 après plusieurs itérations aboutissant en sortie de classification à ces valeurs relatives d'étoiles et de QSO dans l'échantillon XMM/DR7. Afin d'obtenir l'échantillon d'étoiles XMM/DR7 le plus propre possible, nous avons uniquement sélectionné les objets ayant une probabilité d'être une étoile en sortie de KDC supérieure à 99%. De même, afin de ne sélectionner que les identifications XMM/DR7 ayant une grande chance d'appartenir aux étoiles normales, nous avons supprimé celles se situant dans des régions correspondant aux 13% des sources de l'échantillon d'apprentissage stellaire situées dans les zones de plus faible densité. Ces zones sont donc éloignées de la branche des étoiles dans l'es-

[1] `http://casjobs.sdss.org/CasJobs/`

pace des couleurs optiques, et les objets les peuplant ont des chances d'être des étoiles binaires, des CV voire même des AGN mal classés. Cette sélection supprime 32% des sources XMM/DR7 classées comme étoiles par la KDC, suggérant que les étoiles SDSS situées en dehors de la branche photométrique des étoiles sont pour bonne partie des étoiles actives en X ou des AGN.

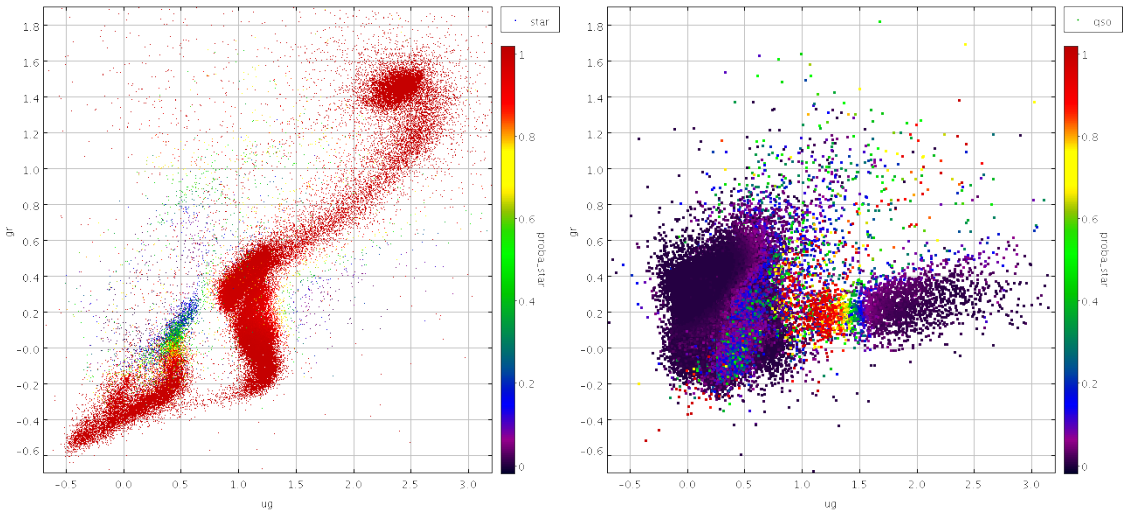


FIG. IV.5 – Étoiles (gauche) et QSO (droite) de l'échantillon d'apprentissage de la classification des sources SDSS dans le diagramme $u-g$ / $g-r$. La couleur code la probabilité qu'a une source d'être une étoile selon la *kernel density classification* dans l'espace 4D des couleurs $u-g$, $g-r$, $r-i$ et $i-z$.

La classification étoiles/QSO a été réalisée sur la base des propriétés optiques des sources identifiées spectroscopiquement par le SDSS. Cependant, la distribution en couleurs des sources de l'échantillon SDSS/2XMMi n'a pas de raison de suivre celle des objets SDSS n'émettant pas en X, ce qui introduit nécessairement des biais dans notre classification. Par exemple, il peut y avoir une importante surdensité d'étoiles n'émettant pas en X dans le diagramme 4D des couleurs. Si cette zone contient quelques QSO, ils seront cependant classés comme étoiles par la KDC. On rencontre ce problème dans la région correspondant aux étoiles de type A qui est également couverte par le début de la branche des QSO à haut redshift. Bien que quelques étoiles de type A semblent émettre en X pour des raisons encore débattues et incertaines, c'est loin d'être le cas pour toutes. Nous avons donc manuellement supprimé des résultats de la classification les sources XMM/DR7 classées comme étoiles ayant une couleur $u-g$ supérieure à 1.2 (pour la valeur, voir Covey *et al.* 2007). Au final, l'échantillon stellaire issu de la KDC contient 842 entrées ayant des probabilités d'être des étoiles supérieures à 99%, des couleurs $u-g>1.2$ et dont les possibles *outliers* ont été supprimés (voir Fig. IV.6).

Nous avons également vérifié que l'échantillon d'étoiles créé suivait bien la branche des étoiles définie dans (Covey *et al.* 2007) dans le diagramme couleur 4D à partir de photométrie synthétique. La correspondance est bonne, sauf pour les étoiles les plus rouges à partir du type spectral M5 environ.

b Proposition d'amélioration

Nous proposons ici une amélioration au système décrit ci-dessus. Les gains possibles sur la qualité de la classification ne nous ont pas paru suffisamment importants pour justifier l'implémentation de cette version plus complexe et dont le temps d'exécution est plus important.

Nous suggérons ici de remplacer la largeur de bande fixe par deux premières matrices de variance-covariance de dimensions 4x4. Chacune de ces matrices est associée à un des échantillons d'apprentissage (« étoiles » ou « QSO ») et est calculée par l'algorithme MCMC présenté paragraphe 5.5.2 du chapitre II. Ces matrices représentent les largeurs de bande optimale pour l'estimation de densité de probabilité dans l'espace des paramètres par *kernel smoothing* pour les

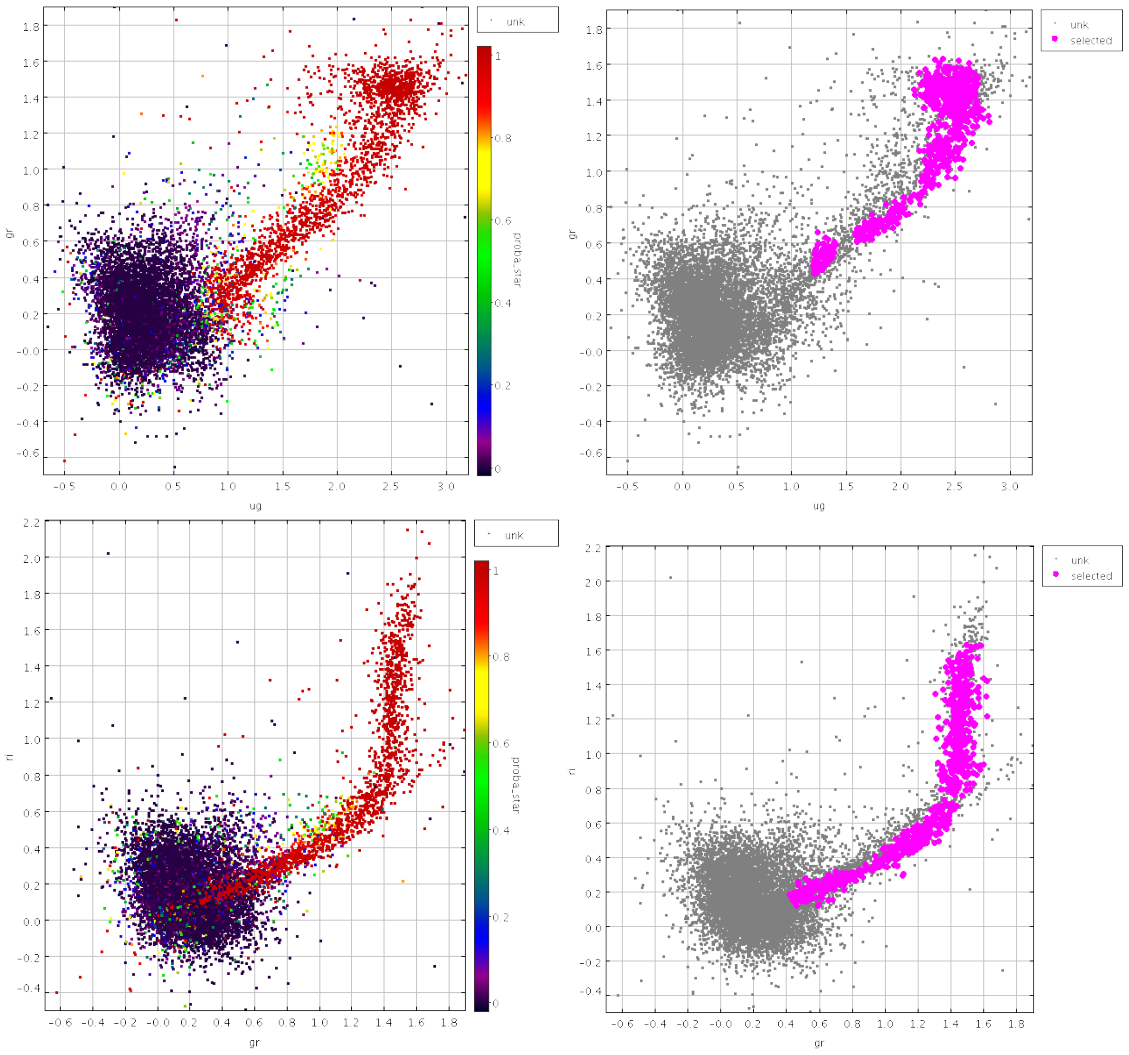


FIG. IV.6 – Diagrammes $u-g$ / $g-r$ et $g-r$ / $r-i$ des sources SDSS corrélées avec des sources du 2XMMi. A gauche, la couleur code la probabilité d'être une étoile. A droite, les sources en magenta sont les sources satisfaisant toutes les conditions pour appartenir à notre échantillon stellaire.

deux échantillons « étoiles » et « QSO ». Pour chacun de ces deux échantillons, on se débarrasse des *outliers* en imposant une limite minimale sur la densité de probabilité en chaque point de la distribution. On recommence alors la première étape avec une nouvelle estimation des deux matrices de variance-covariance. On se débarrasse ensuite des *outliers* dans notre échantillon de sources à classer en rejetant celles qui sont en dessous des deux densités de probabilités minimale définies précédemment pour les étoiles et les QSO. On classe les sources restantes en suivant les règles normales de la *kernel density classification*.

1.2.3 Composition de l'échantillon d'apprentissage pour les données XMM/DR7

Un des intérêts de la corrélation croisée est d'analyser les lieux dans l'espace des paramètres qui sont occupés par les différentes classes d'émetteurs X qui sont associés de façon fiable avec des sources optiques. Pour identifier ces lieux, et pouvoir faire plus tard de la classification automatique, nous avons créé un échantillon d'apprentissage nommé LS_XS (voir §1.2.2). Nous avons pour cela utilisé l'échantillon d'apprentissage réalisé pour l'ensemble des sources du 2XMMi.

Pour rappel, nous avons défini dans cet échantillon les classes issues du SDSS, à savoir : étoiles, galaxies, quasars (QSO), quasars à haut redshift (HizQSO) auxquelles nous avons ajouté les bi-

naires X et les AGN. L'identification spectrale SDSS, accessible dans la base de donnée du SDSS DR7 via CasJob, concerne une partie des sources les plus brillantes (magnitude<19-20). En plus de ces identifications spectroscopiques, nous avons enrichi les différents échantillons par des corrélations croisées « brutes » avec divers catalogues et par l'utilisation d'une *kernel density classification* sur les couleurs des sources SDSS présentes dans cette corrélation 2XMMi/DR7 (voir §1.2.2).

Dans l'étude qui suit, nous ne gardons que les associations 2XMMi/DR7 dont la probabilité d'identification et supérieure à 0.90 et où les magnitudes g, r, et i de la source SDSS sont inférieures à 22.2 mag et où cette même source doit être non saturée. Nous avons ajouté à l'échantillon d'apprentissage des candidats QSO de type II extraits de Zakamska *et al.* (2003). Nous avons également ajouté les CV qui figurent dans (Dillon *et al.* 2008). La composition finale de l'échantillon d'apprentissage pour le sous-ensemble de données constitué des corrélations de bonnes qualité entre 2XMMi et SDSS DR7 est fourni Tab. IV.9.

TAB. IV.9 – Composition de l'échantillon d'apprentissage. Figure ici le nombre de sources XMM uniques (*src_id*) associées avec des sources SDSS non saturées de magnitudes g, r et i $<$ 22.2 mag.

Type d'objet	Etoile	Binaire X	Galaxie	AGN	QSO	QSO II
Nombre	536	26	673	572	1 425	6

1.3 DISCUSSION SCIENTIFIQUE

La corrélation croisée de deux catalogues aussi volumineux que le 2XMMi et la DR7 du SDSS ouvre la voie à de nombreuses investigations. De plus, grâce à l'accès aux raies spectrales de plus d'un million de sources optiques via la base de données du SDSS, on a pour la première fois la possibilité d'étudier la microphysique des régions d'émission optique en fonction des propriétés X des objets. Dans le cas des étoiles actives, il pourrait être possible d'étudier les liens physiques entre les mécanismes de transport de chaleur entre la région d'émission de raie optique dans la chromosphère et la couronne active en X. En ce qui concerne les galaxies actives, on pourrait étudier les liens entre l'émission X et les diagrammes classiques servant à discriminer entre galaxies *starburst*, galaxies *star forming*, LINER et galaxies de Seyfert. Ces études sont cependant en dehors du champ du présent travail.

En revanche, une analyse globale apporte des informations utiles pour la classification. Aussi on s'intéresse essentiellement ici aux parties de l'espace des paramètres occupés par les différentes classes d'objets.

1.3.1 Mise en évidence des principales classes des sources X

La figure Fig. IV.9 montre les positions des objets de l'échantillon d'apprentissage et de toutes les corrélations ayant une probabilité d'identification supérieure à 90% dans le diagramme $g - i$ / $log(f_x/f_r)$, avec d'un coté les objets non-résolus, et de l'autre les objets étendus.

a Séparation entre étoiles et sources extragalactiques

Comme on pouvait s'y attendre, le rapport Fx/Fopt est extrêmement efficace pour isoler la population composée d'étoiles froides, dont l'émission X vient d'une couronne magnétiquement active, des sources dont l'émission X provient de processus d'accrétion, comme c'est le cas dans les binaires X et les AGN. Cependant, on constate une superposition entre cette population et les galaxies ayant une faible luminosité X. L'utilisation des couleurs optiques permet de lever partiellement cette dégénérescence, notamment en ce qui concerne les étoiles M les plus rouges. Restent les étoiles de type G pour lesquelles l'utilisation du paramètre d'extension spatiale fourni par le SDSS permet de lever en grande partie l'ambiguïté avec les galaxies. En grande partie seulement car certaines étoiles, notamment les plus brillantes, sont vues par le SDSS comme des sources étendues. On ne peut pas non plus totalement exclure la possibilité qu'un nombre marginal de galaxies soient vues comme des sources ponctuelles. Les binaires X occupent en partie la même zone de l'espace des paramètres que les quasars. Cependant, leur dispersion est plus importante et on peut l'utiliser pour en identifier certaines. Par exemple, les sources ayant une couleur $g - i < -0.2$

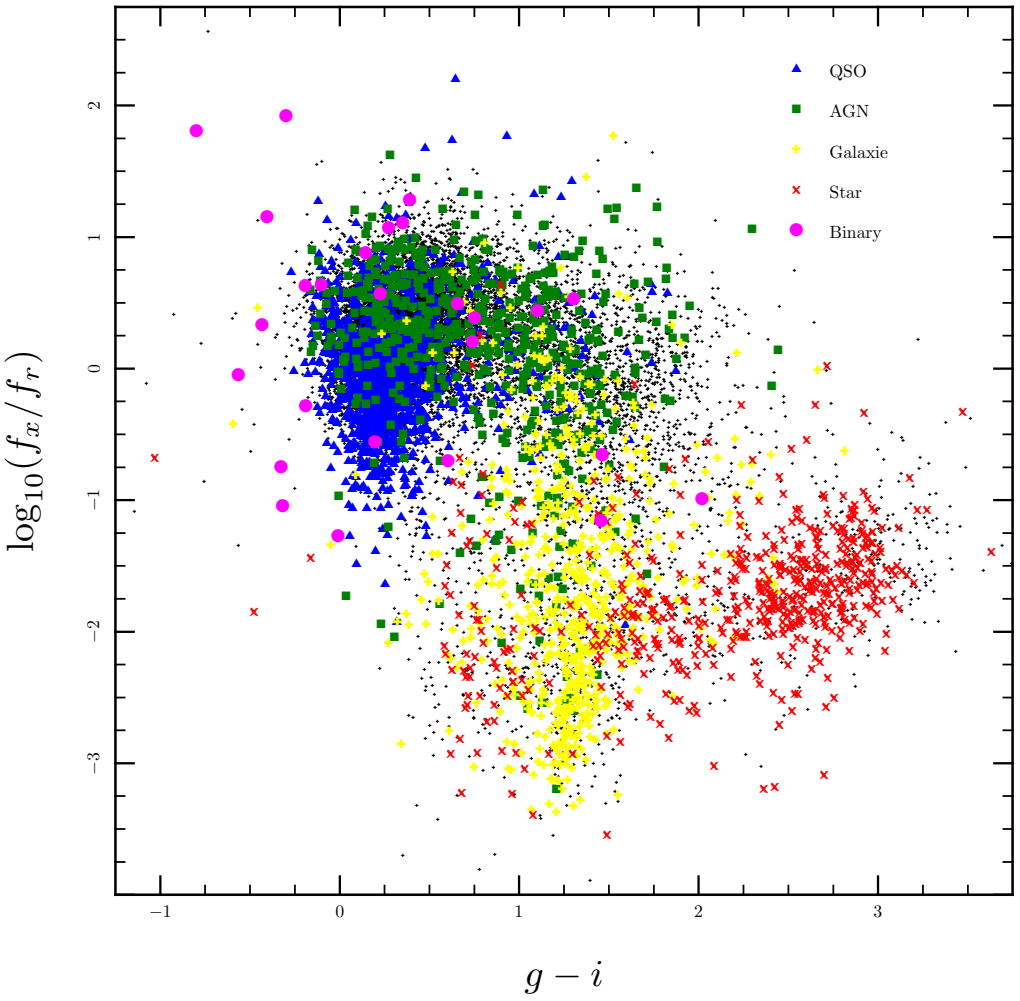


FIG. IV.7 – Diagrammes $log(f_x/f_r)$ versus $g-i$ de l'échantillon XMM/DR7 avec l'ensemble des sources. La couleur code le type d'objet. Points noirs : corrélations qui ne font pas partie de l'échantillon d'apprentissage et qui sont à classer.

ou un rapport Fx/Fopt très important ont de bonnes chances d'être des binaires X. Cette dernière analyse permet d'extraire des candidats binaires X. On a par exemple retrouvé l'ULX Holmberg II X-1 (point noir en haut à gauche Fig. IV.7) qui ne figure pas dans notre échantillon d'apprentissage.

b Séparation en classes de sources extragalactiques

On peut voir Fig. IV.7 qu'un grand nombre de sources 2XMMi ayant une contrepartie SDSS se regroupent dans un intervalle de couleurs bleues allant de $g-i=-0.2$ à $g-i=0.8$. Leurs rapports $\log(f_x/f_r)$ sont ~ 0 et elles sont non-résolues en optique. Cette région du diagramme coïncide avec l'immense majorité des QSO identifiés spectroscopiquement par le SDSS dont la plupart sont des quasars sélectionnés pour leur excès en UV. La moyenne du $log(f_x/f_r)$ des QSO de l'échantillon d'apprentissage apparaît légèrement plus faible que celui de l'échantillon total identifié. Cette tendance peut être logiquement expliquée par le fait que les sources identifiées spectroscopiquement par le SDSS sont plus brillantes que le minimum de détection photométrique. Dans la figure IV.8, on voit que les quasars identifiés spectroscopiquement par le SDSS sont les plus énergétiques avec une luminosité X dépassant les 10^{44}ergs.s^{-1}.

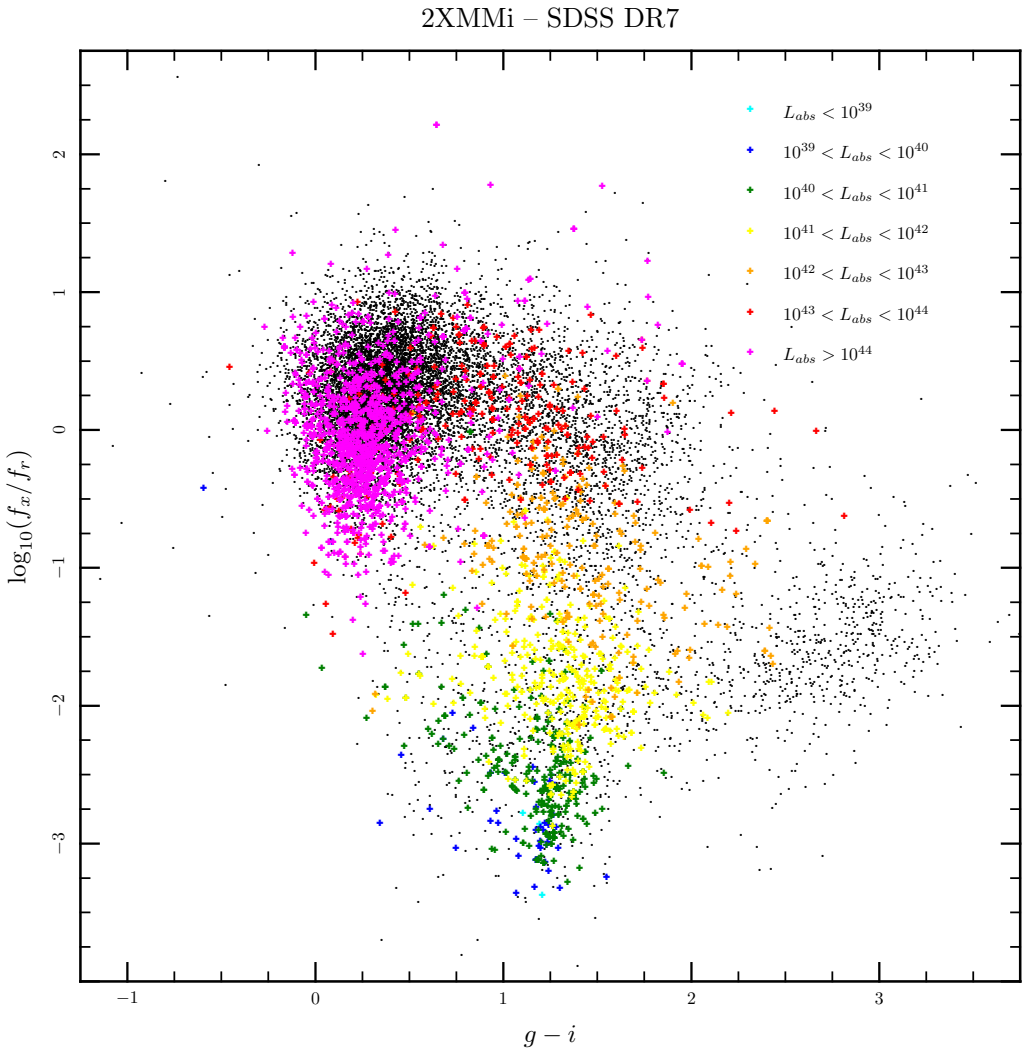


FIG. IV.8 – Diagrammes $log(f_x/f_r)$ versus $g - i$ de l'échantillon XMM/DR7. avec l'ensemble des sources. La couleur code la tranche de luminosité X. Points noirs : corrélations qui ne font pas partie de l'échantillon d'apprentissage et qui sont à classer.

Pour des valeurs de $g - i$ allant de 0.8 à 2.0 apparaît une population de sources, moins importante que la population de quasars précédente, ayant un $\log(f_x/f_r) \sim 0.2$. La plupart des contreparties optiques de ces objets sont étendues et sont pour une partie spectroscopiquement identifiées par les SDSS comme quasars et pour une autre comme galaxies. Leur luminosité X est de l'ordre de 10^{43-44}ergs.s^{-1}. Ces valeurs suggèrent que ces objets, plus rouges que les précédents, soient en grande majorité des galaxies de Seyfert. On remarquera que seules quelques identifications spectroscopiques ont été faites dans la partie la plus rouge du diagramme, pour $g - i \geq 2$, et dans la branche correspondent à des valeurs de $log(f_x/f_r) \sim 0.2$. Nous abordons la nature probable des AGN de cette branche dans la section 1.3.3.c.

Le reste des sources extragalactiques sont étendues en optique et ont une couleur $g - i$ comprise entre 0.8 et 1.8 ainsi qu'un $\log(f_x/f_r)$ compris entre 0 et -3. Leur luminosité X s'étend de 10^{42-43}ergs.s^{-1}, clairement dans le domaine des AGN, à des luminosités inférieures à 10^{40}ergs.s^{-1} qui peuvent s'expliquer par des ULX, des sursauts de formation d'étoiles, ou encore des ensembles de LMXB dans les galaxies elliptiques.

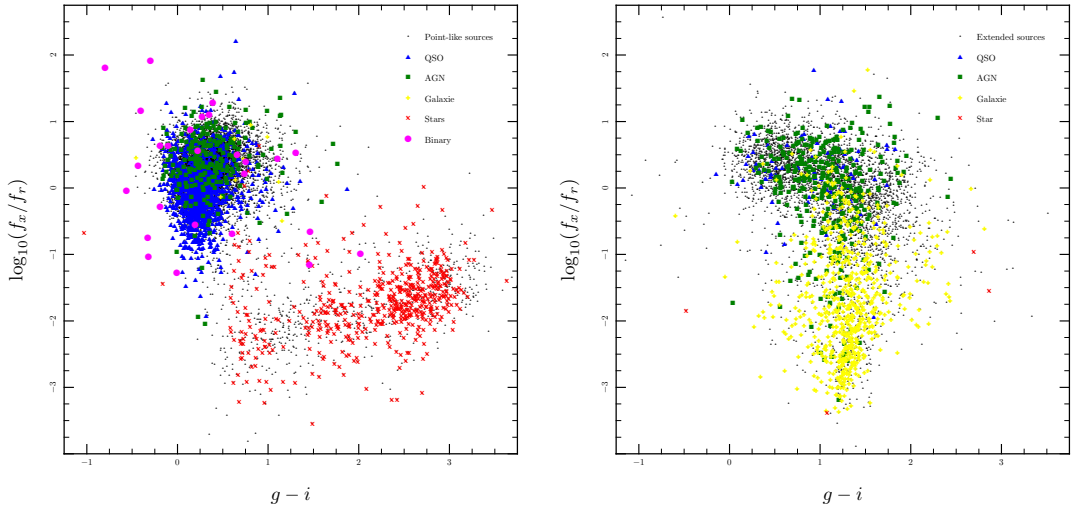


FIG. IV.9 – Diagrammes $log(f_x/f_r)$ versus $g-i$ de l'échantillon XMM/DR7. À gauche : sources optiques ponctuelles uniquement, la couleur code le type d'objet. À droite : sources étendues uniquement, la couleur code le type d'objet. Points noirs : corrélations qui ne font pas partie de l'échantillon d'apprentissage et qui sont à classer.

1.3.2 Étoiles actives en X

Aux hautes latitudes galactiques, qui sont couvertes ici par la corrélation croisée ($|b| \geq 20°$avec une moyenne $E\{|b|\}$ de 58°), la plupart des sources X sont d'origine extragalactique. Par exemple, le *Extended Medium Sensitivity Survey*, qui constitue le plus grand échantillon de sources X dures observées avec le satellite Einstein à hautes latitudes galactiques, ne contient que 25% d'étoiles actives (Stocke *et al.* 1991). Les campagnes d'identifications optiques des sources X du ROSAT *All-Sky Survey* (RASS) ont abouti à des résultats similaires (par exemple, 35% d'étoiles actives en X dans Zickgraf *et al.* (1997)) alors que le ROSAT *deep survey* du champ de Lockman (Schmidt *et al.* 1998), qui atteint une sensibilité ~ 50 fois supérieure à celle de RASS, a recensé moins de 10% de sources stellaires.
Par conséquent, la contribution relative des étoiles actives par rapport au total des sources détectées à hautes latitudes galactiques doit diminuer avec l'augmentation de sensibilité qu'offrent les observations de Chandra et XMM-Newton. Elles offrent en effet un flux limite plus bas et une réponse dans les énergies plus dures que les anciens instruments (voir par exemple la figure 1 de Hérent *et al.* (2006)). Les couronnes actives constituent ainsi moins de 10% des sources identifiées dans le XMM-Newton *serendipitous survey* de Barcons *et al.* (2007). Les relevés Chandra les plus profonds comme le Chandra *Deep Field North* (Alexander *et al.* 2003) ne comptent que $\sim 3\%$ d'étoiles.

Cette situation contraste avec la situation à basses latitudes galactiques. En effet, dans ce dernier cas, la plupart des sources RASS ont été identifiées avec des étoiles (Motch *et al.* 1997). XMM-Newton et Chandra ont étendu aux flux plus faibles cette prédominance des couronnes actives dans la population de sources X molles et de basses latitudes galactiques (Motch *et al.* 2003, Rogel *et al.* 2006). À des énergies plus importantes, une population de sources galactiques dures en X et dépendant de la longitude se superpose au fond de sources extragalactiques qui domine habituellement (Hands *et al.* 2004, Motch 2006*a*).

Récemment, dans Covey *et al.* (2008), ont été publiés les résultats de la corrélation croisée de l'*Extended Chandra Multiwavelength Project* (ChaMP) avec le SDSS DR6. Un total de 384 sources Chandra ont été identifiées avec des étoiles actives en prenant en compte les sources SDSS non résolues, les coïncidences en positions, les couleurs optiques et, pour une partie des sources, des confirmations spectroscopiques.

Comme mentionné plus haut, le nombre d'étoiles émettant en X et spectroscopiquement iden-

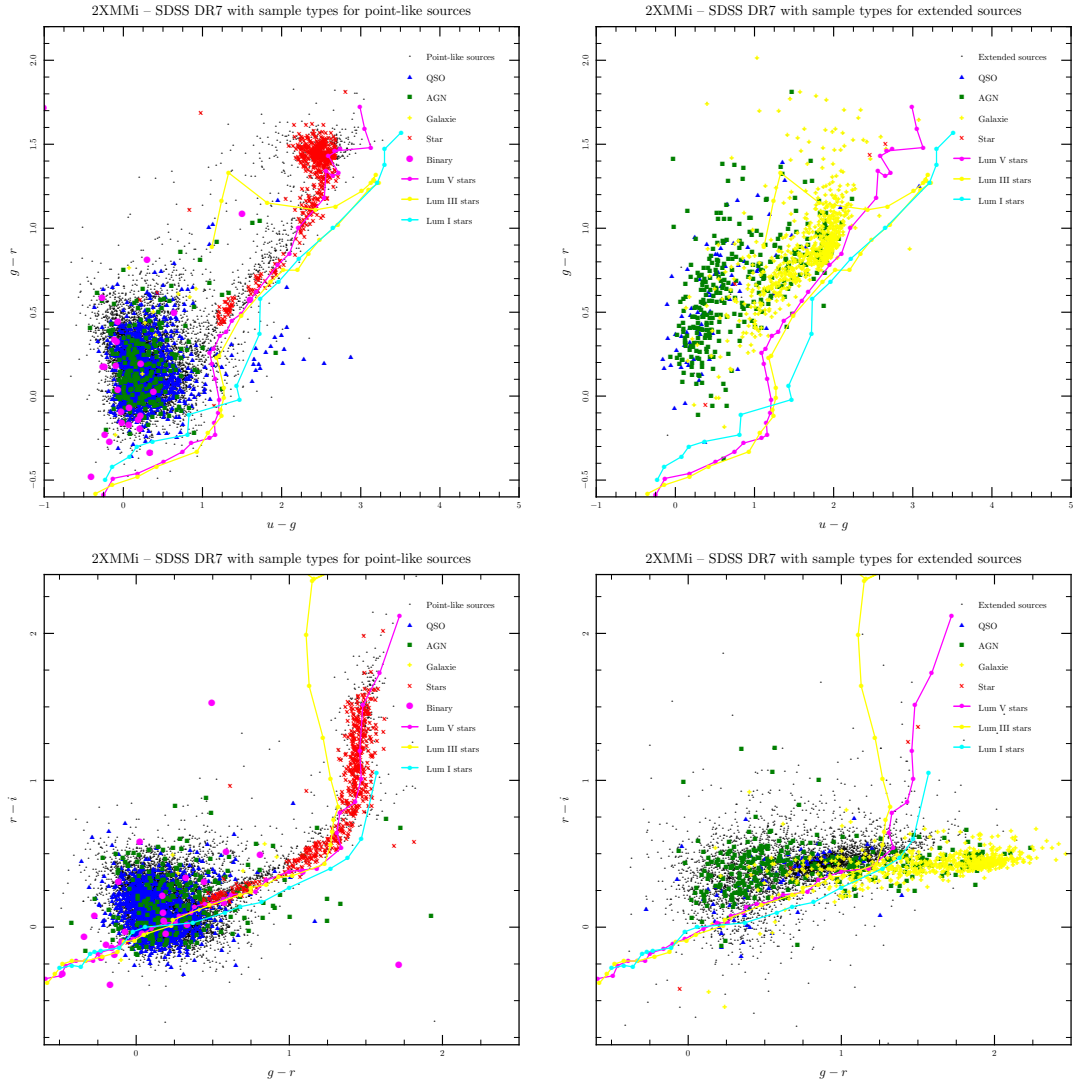


FIG. IV.10 – Diagrammes $u-g$ versus $g-r$ et $g-r$ versus $r-i$ de l'échantillon XMM/DR7. En haut : diagrammes $u-g$ versus $g-r$ pour les sources ponctuelles (gauche) et étendues (droite). En bas : diagrammes $g-r$ versus $r-i$ pour les sources ponctuelles (gauche) et étendues (droite). Points noirs : corrélations qui ne font pas parties de l'échantillon d'apprentissage et qui sont à classer.

tifiées dans le DR7 est relativement faible et se concentre sur les étoiles de type M les plus rouges, ce en raison des objectifs scientifiques lors de la sélection des cibles spectroscopiques. La mise à disposition de nouvelles données SEGUE avec la DR7 a augmenté le nombre d'étoiles identifiées spectroscopiquement mais l'effet sur la statistique de la corrélation croisée est négligeable. Comme mentionné dans la section 1.2, afin d'augmenter le nombre d'étoiles dans l'échantillon d'apprentissage, nous avons utilisé une *kernel density classification* sur les 4 couleurs optiques $u-g$, $g-r$, $r-i$ et $i-z$. Pour des indices de couleur $g-r$ en dessous de ~ 1.5, de nombreuses galaxies actives ont des couleurs se confondant avec celles des étoiles. Cependant, le fait que les étoiles montrent des indices de couleur en $u-g$ plus bleus, et qu'elles ne soient pas étendues, permet de les séparer efficacement des galaxies jusqu'à $g-r \sim 0.2$.

Pour la moyenne de la valeur absolue des latitudes galactiques, qui vaut 58°, et pour des distances typiques de l'ordre de quelques centaines de parsec, l'absorption par le milieu interstellaire reste négligeable en comparaison des autres incertitudes et a peu d'effet sur les couleurs observées des étoiles. Nous avons calculé le rougissement galactique total dans la direction de chaque

étoile en suivant Schlegel *et al.* (1998). La moyenne de E(B-V) vaut 0.029 avec une erreur rms de 0.025. En admettant les coefficients d'absorption calculés par Girardi *et al.* (2004) ($T_{\rm eff}$ = 4500K ; $\log g$ = 4.5), le rougissement maximum applicable en moyenne à notre échantillon d'étoiles serait de 0.028 et 0.018 respectivement pour les couleurs $g-r$ et $r-i$. Seules quelques sources SDSS non résolues associées à des sources du 2XMMi ont des couleurs combinées $u-g$, $g-r$ et $r-i$ compatibles avec celles attendues de géantes de classe de luminosité III, et très peu en ce qui concerne les super géantes. Dans un article récent, Guillout (2009) estiment que leur échantillon de sources RASS à $|b| \lesssim 30°$identifiées avec des sources Tycho brillantes a une contamination moyenne de l'ordre de 35% par des étoiles évoluées, avec un pic à 60% pour les étoiles de type K. Cependant, une fraction plus faible d'étoiles évoluées émettant en X ($\sim$ 10%) était présente dans l'échantillon de Covey *et al.* (2008) qui se trouve à une plus haute latitude galactique et à un flux plus faible, ce qui correspond davantage à notre cas. Dans la suite, nous assumerons que toutes les étoiles appartiennent à la séquence principale. Nous garderons cependant en mémoire qu'une fraction d'étoiles de classes de luminosités III et IV, notamment les systèmes binaires serrés non accrétant tels que les RS CVn, peuvent contribuer à cette population, et le fait de les considérer comme des étoiles naines doit avoir pour effet de sous-estimer les luminosités X calculées pour ces objets.

Notre échantillon comprend $\sim$ 820 candidats étoiles actives ayant des probabilités d'identification supérieures à 90%. Cette limite inférieure sur la probabilité correspond à un taux de fausses associations dans l'échantillon de 2% environ. L'intervalle de couleur $g-r$ de cet échantillon couvre les étoiles de type F jusqu'aux étoiles de type M. On calcule les distances de ces étoiles en négligeant le rougissement galactique et en se servant de la calibration en magnitude absolue de Covey *et al.* (2007). À partir de cette distance, on peut calculer la luminosité X en utilisant le flux reçu dans la bande large 0.2-12 keV fournie dans le catalogue 2XMMi. La distance photométrique moyenne est de 570 pc, et 90% des étoiles ont des distances comprises entre 80 pc et 2300 pc. La distribution globale des Log(Lx) pique à 29.04 et s'étend de $\sim$ 27.7 à $\sim$ 31.9. Cet intervalle de luminosité X couvre à la fois celui des étoiles tardives comme le soleil voire encore moins actives, jusqu'à celui des étoiles les plus actives comme les T Tauri ou les systèmes RS CVn. La luminosité X moyenne ne varie pas avec la latitude galactique dans l'intervalle couvert par notre échantillon.

On peut voir Fig. IV.11 la distribution des étoiles actives en X dans le diagramme $g-r$ / $r-i$ pour différentes gammes de luminosité X. Il apparaît que le *locus* des sources brillantes en X est décalé d'environ 0.1 mag par rapport à celui des sources moins brillantes. Les étoiles actives apparaissent plus bleues en $g-r$ pour une couleur $r-i$ donnée. Ce décalage ne peut pas être dû à une correction du rougissement insuffisante car, au contraire, les étoiles les plus brillantes en X sont supposées être les plus éloignées et les plus absorbées, et devraient donc apparaître plus rouges que les étoiles moins lumineuses en X, qui sont aussi les plus proches. Nous avons également vérifié avec les isochrones de Girardi *et al.* (2004), faites pour les bandes du SDSS, que les effets d'âge et de métallicité ne pouvaient pas expliquer ce décalage entre sources brillantes et moins brillantes en X dans le diagramme couleur / couleur. De façon similaire, l'effet d'une émission H_α renforcée attendue pour les étoiles de type M ($g-r > 1.3$) est exclu puisque qu'elle aurait tendance à augmenter la valeur de $g-r$. Il est intéressant de noter que Covey *et al.* (2008) rapporte des indices de couleur $u-g$ plus bleus de 0.1 mag pour leur échantillon que la moyenne des étoiles de faible masse. Il ne mentionne cependant rien concernant les indices de couleur $g-r$ et la comparaison avec leurs travaux n'est pas évidente. Alors que la courbe suivie dans le diagramme couleur/couleur par nos étoiles actives est compatible avec celle de Covey *et al.* (2007) pour l'ensemble du SDSS, ne contredisant pas les résultats de Covey *et al.* (2008), la différence que nous trouvons se situe entre deux groupes d'étoiles actives. Les *templates* d'étoiles M disponibles dans Bochanski *et al.* (2007) peuvent indiquer une tendance similaire dans l'indice de couleur $u-g$ entre étoiles actives et inactives avec cependant une dispersion intrinsèque importante, alors que cet effet n'apparaissait pas en $g-r$. Malheureusement, nous avons à notre disposition trop peu de mesures précises dans la bande u pour confirmer l'effet observé par Covey *et al.* (2008) sur la couleur $u-g$. Nous sommes d'accord avec ces auteurs pour dire que des *flares* optiques de faible importance peuvent être responsables du décalage vers le bleu des couleurs observées pour les étoiles actives en X.

L'émission des couronnes actives consiste essentiellement en un spectre de plasma optiquement mince dominé par des raies d'émissions étroites se superposant à un continu assez faible (bremstrahlung thermique). Dans la plupart des cas, deux composantes thermiques sont néces-

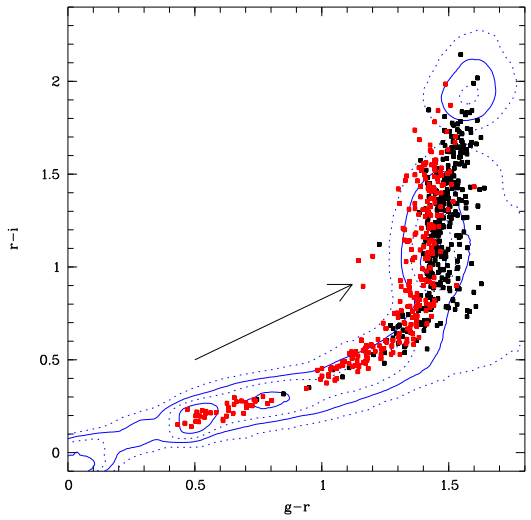


FIG. IV.11 – Distribution des étoiles actives en X dans le diagramme $g-r$ / $r-i$ pour deux intervalles de luminosité X. Les étoiles marquées par des points noirs ont un Log(Lx)<29, et celles marquées par des croix rouges ont un Log(Lx)>29. La flèche rouge montre la direction du rougissement par le milieu interstellaire. Seules les sources ayant une incertitude sur la magnitude g inférieure à 0.1 ont été prises en compte.

saires pour reproduire de façon satisfaisante la distribution en énergie (pour une revue récente, voir Guedel & Naze 2009). Les études en rayons X d'amas ouverts d'étoiles ont permis de dégager une vision d'ensemble relativement cohérente des couronnes actives où vitesses de rotation des étoiles, luminosités X et températures X sont liées. Alors que les étoiles jeunes d'Orion (âge $<$ 1 million d'années) ont des spectres dont les deux composantes thermiques valent typiquement $kT_1 \sim 0.8$ keV et $kT_2 \sim 2.9$ keV, l'analyse des étoiles des Pléiades, plus vieilles (âge $\sim$ 115 Millions d'années), mène à des composantes $kT_1 \sim 0.4$ keV et $kT_2 \sim 1.1$ keV, alors qu'on a $kT1 \sim 0.2$ keV et $kT2 \sim 0.6$ keV (voir Sung *et al.* 2008, et références citées dedans) pour notre Soleil. Grâce aux observations ROSAT et ASCA d'une douzaine d'étoiles sélectionnées avec attention, Guedel *et al.* (1997) a établi que la luminosité X, la température des deux composantes et le rapport entre l'émission des plasmas chaud et froid diminuent avec l'âge et la vitesse de rotation. La gamme de luminosité X observée dans notre échantillon suggère des âges pour nos étoiles de l'ordre de 2 milliards d'années.

On a donc cherché une dépendance entre luminosité et propriétés X de nos étoiles actives. On a pour cela sélectionné 305 couronnes actives se trouvant à des distances comprises entre 80 et 2300 pc et ayant des erreurs inférieures à 0.2 sur le HR2 du 2XMMi. On a ensuite découpé cet échantillon en cinq tranches de luminosité pour lesquelles on a calculé la moyenne de HR2. HR2 peut être considéré comme un indicateur de la forme intrinsèque de la distribution d'énergie en X et est particulièrement adapté pour la gamme de température associée aux couronnes actives. On peut voir Fig. IV.12 la variation du HR2 médian en fonction de la luminosité X. On en déduit que quand la luminosité X augmente, le spectre X devient plus dur. En supposant qu'on ait une seule composante thermique et une colonne densité $N_H \sim 1.7 \times 10^{20}$ cm^{-2}, on obtient des températures d'environ 0.45 keV pour la tranche associée aux luminosités X les plus faibles et de 0.72 keV pour la tranche associée aux plus grandes luminosités X. Cependant, les HR3 présentent une valeur moyenne ~ -0.75, significativement plus dure que celle attendue pour un modèle de températures à une composante basé sur le HR2. On explique naturellement cette différence par le fait que, comme dit plus haut, la distribution d'énergie est mieux modélisée par deux températures de plasmas optiquement minces. Malheureusement, il est impossible d'ajuster kT_1, kT_2 et le rapport entre les émissions de la composante froide et de la composante chaude, soit 3 paramètres, avec les deux seules variables que sont HR2 et HR3. Malgré cela, nos données permettent de confirmer pour la première fois le durcissement spectral qui dépend de la luminosité X des

couronnes actives pour des groupes d'étoiles de champ d'âges hétérogènes.

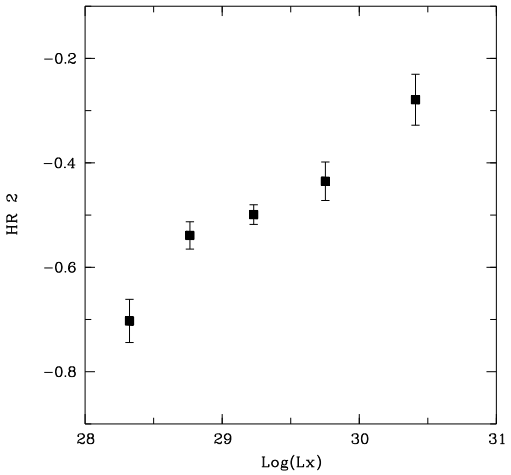


FIG. IV.12 – Variation du HR2 de la caméra EPIC pn en fonction de la luminosité X moyenne de 305 sources ayant des erreurs sur le HR2 < 0.2. Sont tracées ici les médianes et les erreurs associées.

Nous avons sélectionné manuellement les étoiles spectroscopiquement identifiées par le SDSS et ayant des couleurs $r - i > 0.9$, $g - r$>1.2 et $g - r$<1.7, ce qui correspond à la zone couverte par les étoiles de type M plus tardives que M2. De même pour les associations 2XMMi/DR7 ayant une probabilité d'identification supérieure à 0.75 (fiabilité > 0.95) et pour lesquelles la source SDSS est identifiée comme étant une étoile par la *kernel density classification*. Le rapport entre les deux histogrammes par tranche de couleur $r - i$ est montré Fig. IV.13. Pour une couleur $r - i$ dans une plage allant de 1.0 à 1.7 (étoiles M5V), le rapport des étoiles observées en X sur les identifications spectrales SDSS augmente. Cet effet est essentiellement dû à un effet géométrique connu : la luminosité absolue en X des étoiles tardives varie peu en fonction du type spectrale comparée à la luminoqité absolue en optique qui diminue fortement quand on passe des étoiles F aux étoiles M ; les étoiles M5V détectées en optique sont donc plus proches que les étoiles de type M2, M3, et M4 , et donc une plus grande fraction sera détectée en X. Pour $r - i > 1.7$, le rapport entre les deux histogrammes diminue : on pense que c'est dû à un changement des propriétés X des étoiles plus tardives que les étoiles de type M5 qui cessent d'avoir un cœur radiatif et deviennent purement convectives.

1.3.3 Galaxies actives et quasars

Un certain nombre de campagnes d'identification optique de sources X XMM sont en cours et ont déjà apporté plusieurs résultats scientifiques. Une partie de ces campagnes sont conduites au sein du SSC d'XMM-Newton et certaines ont pour but l'étude des sources extragalactiques à hautes latitudes galactiques. On peut citer le relevé BSS, pour *Bright Sources Survey* (Della Ceca *et al.* 2004), qui contient 400 sources dont la luminosité est supérieure à 7×10^{-14}erg cm^{-2} s^{-1}dans les bandes 0.5-4.5 keV et 4.5-7.5 keV. Le papier Caccianiga *et al.* (2007) montre que la fraction d'objets obscurcis varie de façon significative en fonction de la bande d'énergie dans laquelle on sélectionne les objets. Le XMS, pour *XMM-Newton Medium Sensitivity Survey* (XMS, Barcons *et al.* 2002, Carrera *et al.* 2007, Barcons *et al.* 2007), est un relevé effectué pour des flux X intermédiaires de quelques 10^{-14} erg cm^{-2} s^{-1}dans la bande 0.5-4.5 keV. Il contient 318 sources ayant un fort taux d'identification ($\sim$ 90%). La littérature contient plusieurs relevés explorant la nature de sources X dont les flux limites vont de valeurs faibles, de l'ordre de 10^{-15} erg cm^{-2} s^{-1}, jusqu'à des valeurs 100 fois supérieures. Ils ne couvrent cependant qu'un petit nombre de champs de vue et possèdent donc un nombre de sources limité. On peut par exemple citer les 695 sources du champ COSMOS (Brusa *et al.* 2007) et les 328 sources du champ Marano (Krumpe *et al.* 2007). Enfin, le relevé HELLAS2XMM (Cocchia *et al.* 2007), sensiblement plus brillant, contient 220 sources XMM.

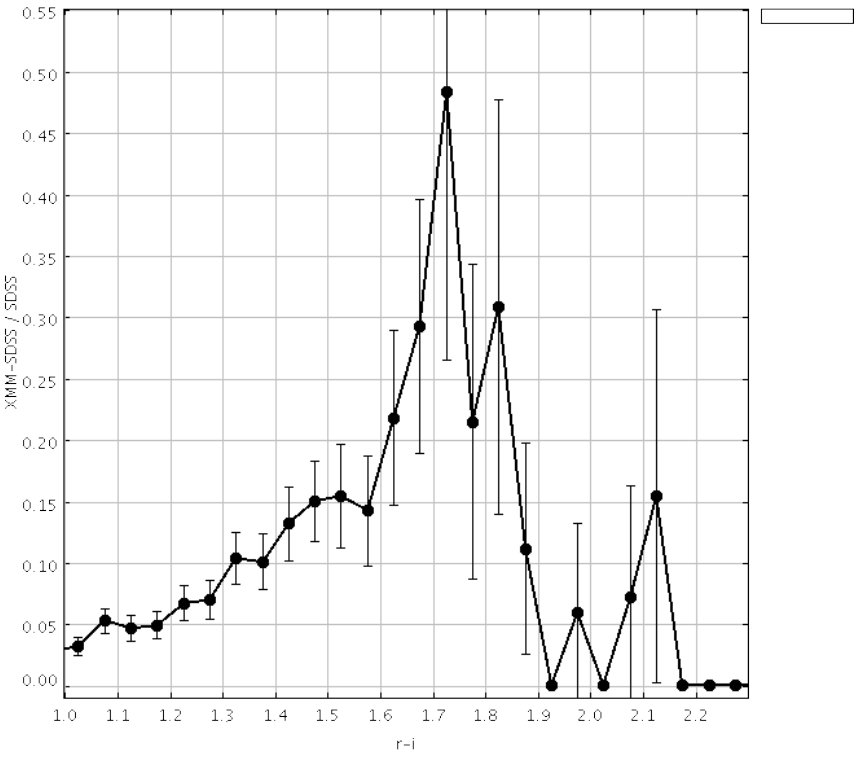


FIG. IV.13 – Histogramme en $r - i$ des étoiles de l'échantillon XMM/DR7 pour les étoiles plus rouges que $r - i = 1.0$

Le principal intérêt de ces études réside dans l'établissement de lois logN-logS pour des flux faibles et des bandes X dures, ce qui correspond à la zone où on s'attend à trouver les AGN responsables du fond diffus X. Un des objectifs est la recherche de QSO de type II, à la fois en X et en optique. Une population importante de sources X brillantes associées à des galaxies normales, des XBONG pour *X-ray Bright Optically Normal Galaxies*, commence à être mise en évidence alors que les raisons de l'absence en optique de signature de l'activité nucléaire visible en X suscite de nombreuses questions.

La plupart des sources du 2XMMi sont associées à des entrées du SDSS DR7 dont la magnitude est proche de la magnitude limite du relevé, à savoir $\sim$ 22-23 mag, et seule une petite partie de ces sources optiques sont assez brillantes pour avoir été sélectionnées comme cibles spectroscopiques. En effet, sur un total de 60 567 détections XMM associées à des sources DR7, 87% ont une erreur sur leur magnitude r inférieure à 0.2 mag, mais seulement 12% ont un spectre. La courbe des erreurs sur les magnitudes en fonction des magnitudes est représentée figure IV.14. De plus, la répartition des cibles d'observations spectroscopiques ne recouvre pas de façon uniforme l'espace des paramètres de notre échantillon total. Selon Abazajian *et al.* (2009), les cibles d'observations spectroscopiques du SDSS DR7 viennent du choix fait d'observer des classes particulières d'objets tels que les galaxies brillantes, les galaxies elliptiques rouges, les contreparties des sources X ROSAT, etc. Pour notre besoin, les deux groupes de cibles spectroscopiques les plus importants sont les candidats quasars définis par Richards *et al.* (2002) et les étoiles du programme SEGUE (Yanny *et al.* 2009). Le programme SEGUE a pour objectif l'étude de la structure galactique. Comme mentionné précédemment, l'échantillon de quasars contient essentiellement des objets choisis pour leur excès en UV, plus quelques cibles rouges suspectées d'être des QSO à haut redshift (HizQSO). La distribution des entrées photométriques du SDSS dans le diagramme $g - i$ / f_x/f_r s'étend plus loin dans la partie rouge que les entrées spectroscopiques composées des quasars UVX (sélectionnés sur la base de leur excès en UV). Seules quelques cibles spectroscopiques classées comme quasar ou galaxie se trouvent dans la partie rouge de l'échantillon photométrique (voir Fig. IV.8). De plus, les entrées SDSS ayant un spectre sont nécessairement plus brillantes que la majorité de

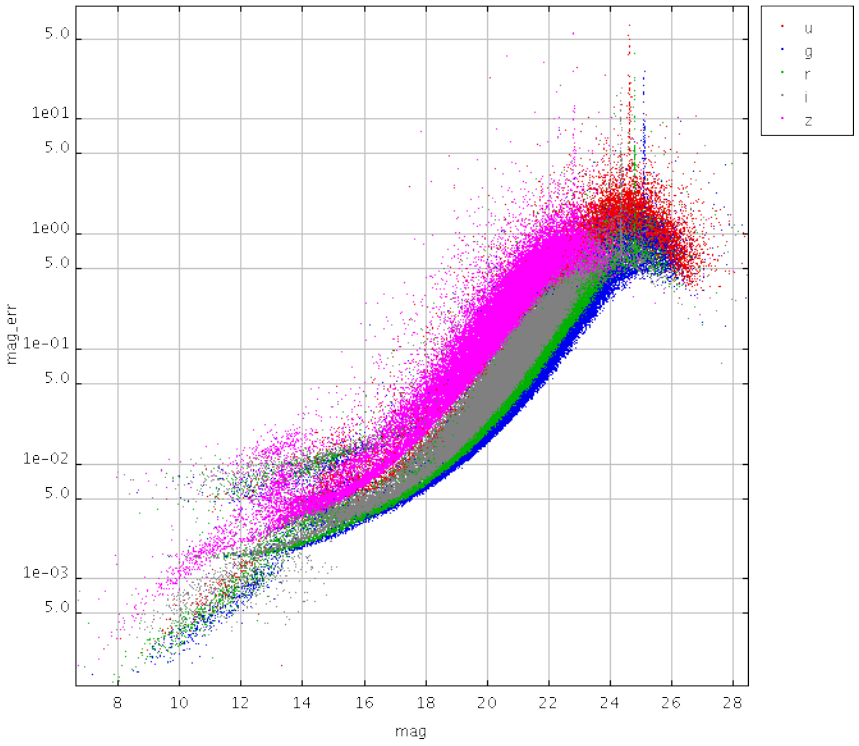


FIG. IV.14 – Erreurs sur les magnitudes optiques u (rouge), g (bleu), r (vert), i (gris) et z (magenta) des sources de l'échantillon XMM/DR7

celles n'en ayant pas et ne permettent donc pas d'apporter des informations sur les sources lointaines ayant d'importants rapports f_x/f_r. Nous soulignons ici le fait que bien que la spectroscopie du SDSS permette d'étudier les émetteurs X brillants en optique les plus fréquents à hautes latitudes galactiques et permette, en raison de la large zone couverte, de découvrir d'éventuels objets rares (des *outliers*), elle ne permet en revanche pas, à elle seule, de déterminer complètement la population d'objets X à hautes latitudes galactiques, ni d'étudier les effets de l'évolution qui affectent la nature des sources X plus faibles que la limite en magnitude du SDSS. Ce dernier point est du ressort des programmes d'identification mentionnés ci-dessus, notamment dans le cadre du SSC, qui ont précisément pour but d'étendre les identifications spectroscopiques aux objets faibles, permettant ainsi d'étudier la population de sources X extragalactiques mal connue, particulièrement en ce qui concerne les objets à haut rapport Fx/Fopt. Les deux stratégies, beaucoup d'objets mais peu profonde d'un coté et peu d'objets mais profonde de l'autre, sont complémentaires et particulièrement appropriées pour caractériser et appréhender le contenu scientifique du catalogue 2XMMi.

Nous faisons cependant remarquer que la qualité de la photométrie du SDSS et la disponibilité des caractéristiques des raies spectrales, tant en absorption qu'en émission, pour les objets observés spectroscopiquement, ouvre des possibilités d'études basées sur les corrélations entre sources X et propriétés optiques.

a Candidats QSO de type II

Les AGN de faible luminosité obscurcis, à savoir les Seyfert de type II, sont détectés en grand nombre dans l'univers local. Selon le modèle unifié standard, l'absorption par le tore de poussière entourant le trou noir centrale bloque, pour des inclinaisons élevées, la vue directe de la machinerie centrale. On n'a donc pas accès à la région d'émission des raies larges en optique et la partie molle de l'émission X est photoélectriquement absorbée. Des photons X mous peuvent cependant s'échapper de l'AGN par de multiples diffusions. L'intensité du fond X dur nécessite l'existence d'une population d'AGN obscurcis conséquente. Les quasars de type II, analogues des Seyfert II mais à plus grandes luminosités X, pourraient contribuer à une grande partie de ce fond X dur.

Se basant sur les identifications spectroscopiques du SDSS, Zakamska *et al.* (2003) a créé un échantillon de 291 candidats quasars de type II se trouvant à des *redshifts* allant de 0.3 à 0.83. Son critère de sélection consiste en l'existence de raies d'émission étroites ayant une largeur équivalente importante, de rapports élevés entre différentes raies d'ionisation, et d'une raie [O III] λ_{5007} forte. Ces deux derniers points sont des indicateurs reliés à la luminosité du noyau actif. Une petite partie de ces candidats ont été observés en X par Chandra ou XMM (Ptak *et al.* 2006) et notre échantillon en contient quelques uns. On recense Tab. IV.10 les principales propriétés des 7 candidats QSO de type II contenus dans nos corrélations. Les luminosités X ont été calculées sur la base d'une courbe moyenne dans la gamme d'énergie 0.2-12 keV et sont donc moins précises que celles données dans Ptak *et al.* (2006), basées sur des analyses spectrales détaillées. Les quatre candidats en commun avec ceux de Ptak *et al.* (2006) ont des luminosités compatibles à un facteur 2 près avec nos valeurs.

TAB. IV.10 – Candidats quasar de type II extraits de Zakamska *et al.* (2003) et correspondant à des entrées dans le 2XMMi.

2XMMi name	id prob	z	u-g	g-i	HR 2 (pn)	Log(L_X)
2XMM J005621.6+003235	+0.964	+0.484	1.891 ±0.643	+1.800 ±0.061	-0.56 ±0.28	+42.78
2XMMi J011522.2+001518	+0.997	+0.390	6.023 ±1.239	+2.437 ±0.046	+0.73 ±0.04	+44.26
2XMM J015716.9-005305	+0.988	+0.540	+.565 ±0.167	+1.836 ±0.050	-0.18 ±0.46	+42.78
2XMM J021047.0-100152	+0.999	+0.540	+.348 ±0.136	+1.771 ±0.047	+0.66 ±0.12	+44.40
2XMM J103951.5+643005	+0.991	+0.402	+.490 ±0.075	+1.209 ±0.027	-0.68 ±0.33	+43.02
2XMM J122656.4+013124	+0.998	+0.732	+.970 ±0.256	+1.134 ±0.056	+0.76 ±0.06	+44.83
2XMM J164131.6+385841	+0.995	+0.596	+.550 ±0.112	+1.768 ±0.028	+0.63 ±0.05	+44.96

On peut voir qu'à l'exception des trois objets ayant des luminosités X faibles, et qui sont probablement des galaxies de Seyfert de type II, le HR2 de ces objets de type II de luminosité X importante est dur alors que la distinction sur la base des couleurs est moins évidente.

b Variations des HR avec la luminosité, le type de galaxie et le redshift

On montre Fig. IV.15 le comportement des 3 HR EPIC pn en fonction de la luminosité X pour l'ensemble des identifications spectroscopiques SDSS de notre échantillon. La majorité des QSO ayant des luminosités $> 10^{44}$ erg.s^{-1} dans la bande 0.2-12 keV se regroupent autour de HR de valeurs <HR2> = -0.12, <HR3> = -0.38 et <HR4> = -0.28, en accord avec les valeurs attendues pour des spectres X en loi de puissance d'indice de photon Γ = 1.9 avec une absorption locale négligeable et une absorption galactique moyenne de 1.16×10^{21}cm^{-2} (moyenne de cette valeur dans l'ensemble des directions des galaxies et des QSO de l'échantillon). Pour cet ensemble de QSO, il n'y a pas de signes évidents d'évolution de l'indice de photon en fonction de la luminosité X. Cependant, certains de ces QSO montrent des spectres X plus durs que ceux des autres objets. Autre point intéressant, quelques sources très lumineuses en X se trouvent associées à des galaxies normales en optique, ou montrant tout au plus des raies étroites. Il s'agit là d'excellents candidats QSO de type II, BL Lacs ou XBONG extrêmes.

Quand on rentre dans le régime des AGN, pour des luminosités X inférieures à 10^{44} erg.s^{-1}, le nombre de sources étendues ayant des spectres de galaxies normales augmente considérablement et les formes que prend la distribution d'énergie en X deviennent plus dispersées. Cependant, la figure impliquant le HR3 révèle l'existence d'un groupe de galaxies actives assez bien défini, ayant une luminosité X de l'ordre de 10^{41} erg.s^{-1}, et se regroupant autour d'une même valeur de HR3 que les quasars plus lumineux de quelques ordres de magnitude. Il faut garder en mémoire le fait que les dispersions plus importantes observées pour les AGN de faible luminosité X sont probablement dues au fort biais introduit par le mode de sélection de l'échantillon spectroscopique du SDSS, basé sur l'excès UV (Richards *et al.* 2002). Par conséquent, les QSO spectroscopiquement identifiés ne sont probablement pas représentatifs de l'ensemble de la population des AGN détectés en X. Comme le montre la figure Fig. IV.9, un nombre important de sources non résolues en optique et dont l'indice de couleur en $g - i$ est plus rouge que ~ 1 possèdent un rapport f_x/f_r similaire à celui des QSO spectroscopiques bleus et sont donc probablement des quasars rouges.

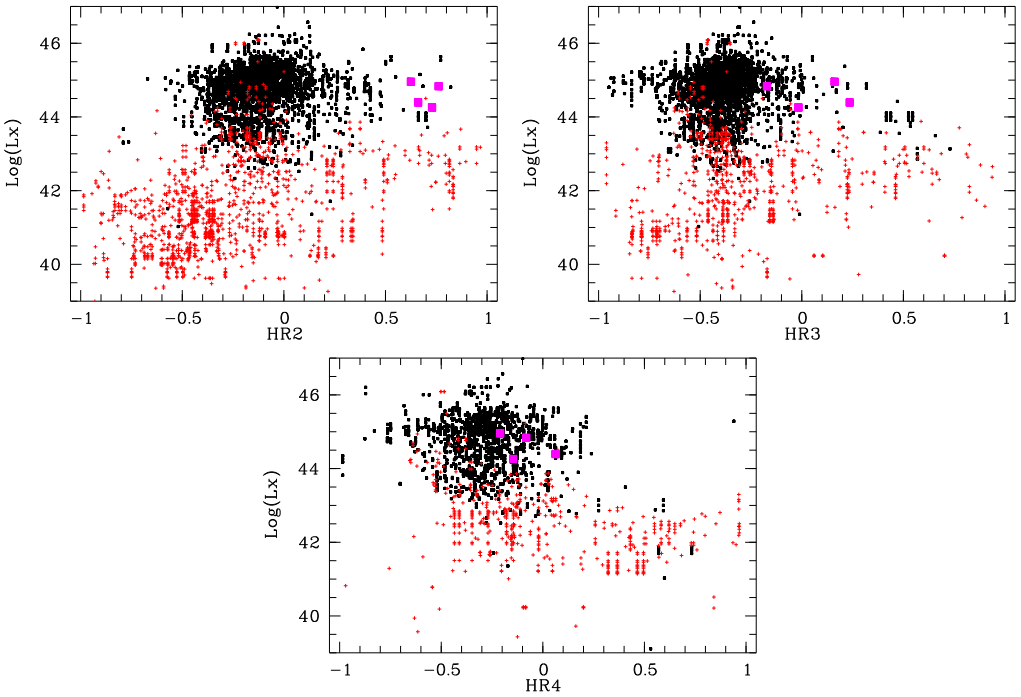


FIG. IV.15 – Variation de la luminosité X en fonction des rapports de dureté EPIC pn pour les sources spectroscopiques du SDSS classées comme quasars (points noirs) et galaxies (croix rouges). Les carrés magenta marquent les candidats QSO de type II de Zakamska et al. On n'a gardé que les sources ayant des erreurs sur les HR inférieures à 0.2.

Une transition spectrale marquée apparaît pour des luminosités X inférieures à $\sim 10^{41-42}$ erg.s^{-1}, à savoir quand l'émission X du trou noir central a moins de chances de dominer l'émission X totale de la galaxie. L'émission combinée des régions de formation d'étoiles, des SNR et des XRB apparaît souvent significativement plus molle que celle des AGN classiques. À cause de leurs spectres mous et de leurs faibles taux de comptage (nombre de photons X reçus par unité de temps), peu de HR4 sont mesurés pour les galaxies à faible luminosité X. Celles-ci manquent notamment dans la partie basse de la figure Fig. IV.15.

Les quatre quasars de forte luminosité extraits de l'échantillon de Zakamska *et al.* (2003) ont tous un HR2 extrêmement dur alors que leur distribution en HR4 ressemble à celle des QSO de type I. Ces valeurs en HR2 reflètent probablement l'importante absorption photoélectrique présente dans de nombreux QSO de type II. Cette absorption explique que l'on ait essentiellement découvert ces objets dans les relevés en X durs (voir par exemple Caccianiga *et al.* 2004). Les ajustements de spectres X révèlent également une colonne densité N_H très importante, de l'ordre de 10^{22}cm^{-2} ou plus, et un indice de photon, et donc un HR4, similaire à celui des quasars de type I.

Nous avons également cherché à déterminer si la forme moyenne des spectres, mesurée via les HR, présente des traces d'évolution en fonction du redshift. La Fig. IV.16 ne montre pas de trace d'une éventuelle évolution de la forme du spectre avec le redshift, du moins pas jusqu'à z=3.

c Caractérisation de populations d'AGN particuliers

Candidats quasars de type II : un certain nombre d'identifications spectroscopiques SDSS occupent la même région dans le diagramme L_X / HR2 que les candidats de Zakamska (voir Fig. IV.15). Nous avons regardé la nature de ces objets en sélectionnant ceux ayant un $\log(L_X) > 44$ et un $HR2 > 0.5$ avec une erreur sur le HR2 inférieure à 0.2. Douze objets remplissent ces conditions, parmi lesquels seulement un (2XMM J134507.9-001901 / SDSS J134507.93-001900.9 à z = 0.419, L_X $\sim 3 \times 10^{44}$ erg.s^{-1}) est un nouveau candidat quasar de type II. Les onze autres sources sélectionnées sont décrites dans la littérature comme étant des QSO de type I à haut redshift, ou montrant

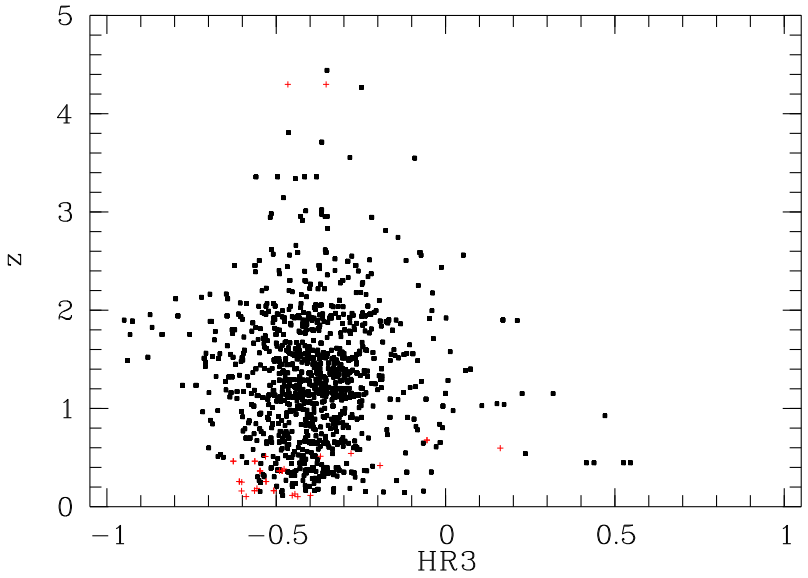


FIG. IV.16 – Variation du redshift en fonction du HR3 de la caméra EPIC pn, à la fois pour les QSO identifiés spectroscopiquement dans le SDSS (carrés noirs) et pour les galaxies (croix rouges). Nous avons uniquement gardé les sources ayant une erreur sur le HR3 inférieure à 0.2 et, pour les QSO, ceux ayant une luminosité X comprise entre 10^{44} to 10^{46} erg.s^{-1}.

de raies d'absorption, ou un rougissement dû à de la poussière, voire sans commentaires particuliers. Il apparaît cependant clairement que bien que les QSO de type II sont bien séparés dans le diagramme L_X / HR, d'autres types d'objets rares se trouvent dans la même zone.

Objets rouges à haut rapport f_x/f_r : comme on peut le voir Fig.IV.9, un grand nombre d'associations XMM/DR7 remplissent une région où le rapport $Log(f_x/f_r)$ est relativement élevé (autour de 0) et la couleur $g-i$ est relativement rouge. Bien qu'ayant un rapport Fx/Fopt similaire à la majorité des quasars bleus, seules quelques-unes de ces sources ont été sélectionnées pour le suivi spectroscopique du Sloan. Si on sélectionne les identifications spectroscopiques ayant un $Log(f_x/f_r) >$ -0.2 et une couleur $g-i > 1.5$ dont l'erreur associée est inférieure à 0.5, on trouve 30 candidats dont 6 seulement sont ponctuels. La moitié sont des galaxies de Seyfert de type I, 10% apparaissent comme des galaxies de Seyfert de type II, et le reste est composé de BL Lac, de QSO à hauts redshifts et d'objets non classés. Dans plusieurs cas, le spectre rouge de la galaxie est clairement visible dans le spectre optique. Un total de 3 817 associations XMM/DR7 occupent cette partie de l'espace des paramètres, ce qui représente 10% des corrélations totales ayant une probabilité d'identification supérieure à 90%. Parmi ces associations, 85% des contreparties optiques sont étendues. Il y a donc de fortes chances pour que la majorité de ces sources à haut rapport Fx/Fopt doivent leur couleur rouge à une contamination de la lumière du cœur, de luminosité modérée, par l'émission plus rouge du reste de la galaxie. Une partie de ces sources, notamment celles qui sont ponctuelles, sont probablement des quasars à haut redshift ou des BL Lac.

AGN de forte luminosité X : nous avons essayé d'explorer la nature des AGN de forte luminosité X en suivant deux méthodes d'investigation.

Dans un premier temps, nous avons uniquement considéré les entrées spectroscopiquement identifiées comme galaxies et associées à des sources X de luminosité supérieure à 10^{44} erg.s^{-1}. Au total, 25 entrées SDSS remplissent ces conditions. Dans au moins deux cas très brillants, un redshift spectroscopique erroné est à l'origine de la luminosité X particulièrement importante. Cependant, la majorité de ces sources sont des BL Lac connus ou de bons candidats BL Lac en raison de leur spectre optique ne montrant aucun signe d'activité et de leur association avec une source radio brillante. Nous retrouvons également dans cet échantillon le candidat quasar de type II 2XMM J134507.9-001901 trouvé par sélection dans le diagramme Lx / HR2.

La seconde étape consiste en l'analyse de la nature des AGN rouges ($g-i > 1$; err$(g-i) <$ 0.5) et très lumineux en X (Lx $> 10^{45}$ erg.s^{-1}), indépendamment de leur classification spectrale

SDSS (QSO ou Galaxy). Parmi les 27 sources retournées par cette sélection, 20 ont un redshift supérieur à 3, alors que les sept autres se trouvent à moins de z=2. Cette surprenante dichotomie reflète probablement la stratégie employée pour sélectionner les cibles spectroscopique du SDSS. Au moins trois de ces 7 objets ont été étudié par Young *et al.* (2008) qui conclut à un rougissement dû à l'absorption par de la poussière.

Candidat BAL QSO : la figure Fig. IV.17 montre la distribution dans le digramme $u-g$ / $g-i$ des identifications spectroscopiques des sources XMM pour différentes tranches de luminosité X auxquelles on a ajouté les sources photométriques. Une fois encore, on peut voir sur ce diagramme l'existence d'une population rouge d'AGN de luminosité intermédiaire qui est vraisemblablement constituée de galaxies de Seyfert pour lesquelles la couleur $g-i$ est fortement affectée par la contribution de la galaxie hôte. En revanche, la couleur $u-g$ est davantage compatible avec celle des quasars classiques sélectionnés sur la base de leur excès en UV. Les galaxies dont la luminosité X est inférieure à 10^{42} erg.s^{-1} suivent la relation couleur/couleur des galaxies normales (Strateva *et al.* 2001) alors que celles qui sont plus brillantes sont légèrement décalées de cette même relation. Les entrées photométriques situées dans la partie la plus rouge du diagramme sont principalement des étoiles.

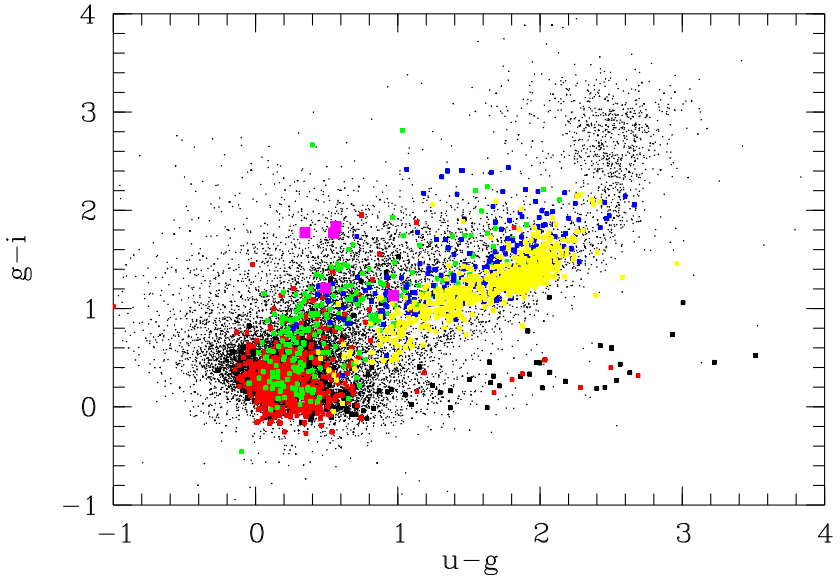


FIG. IV.17 – Distribution des corrélations 2XMMi/DR7 dans le diagramme $u-g/g-i$. Seules les sources ayant des erreurs sur les couleurs inférieures à 0.5 sont représentées. Pour les entrées spectroscopiques du SDSS (carrés), la couleur code l'intervalle de luminosité X. Noir $> 10^{45}$ erg.s^{-1}, rouge 10^{44} erg.s^{-1} à 10^{45} erg.s^{-1}, vert 10^{43} erg.s^{-1} à 10^{44} erg.s^{-1}, bleu 10^{42} erg.s^{-1} à 10^{43} erg.s^{-1}, jaune 10^{40} erg.s^{-1} à 10^{42} erg.s^{-1}. Les candidats QSO de type II issus de Zakamska *et al.* (2003) figurent en magenta. Les points noirs marquent les positions des corrélations entre sources du 2XMMi et entrées photométriques du SDSS.

Une branche distincte de cibles spectroscopiques de forte luminosité X apparaît pour des couleurs $u-g$ comprises entre 1 et 4 et des couleurs $g-i$ comprises entre 0 et 1. Ces quasars partagent les mêmes couleurs $g-i$ que les quasars bleus classiques mais sont beaucoup plus rouges en $u-g$. La plupart de ces objets sont des BAL QSO connus ou nouveaux. Un certain nombre d'entrées photométriques se situent dans la même zone du diagramme et sont donc de bons candidats quasars BAL fortement absorbés.

Malheureusement, il n'y a aucun signe tangible suggérant que les QSO ayant un spectre X dur aient tendance à avoir des couleurs plus rouges. On montre Fig. IV.18 la meilleure preuve en faveur d'une possible corrélation entre couleur $u-g$ et un HR3 dur. En effet les QSO les plus durs en X ont tendance à être très rouges mais la partie du diagramme concernée contient peu de sources. Cependant, plusieurs QSO spectroscopiques et candidats QSO ont une couleur $u-g$ rouge alors que leur spectre X paraît normal. D'autres combinaisons de couleurs et de HR ap-

portent des preuves encore moins convaincantes.

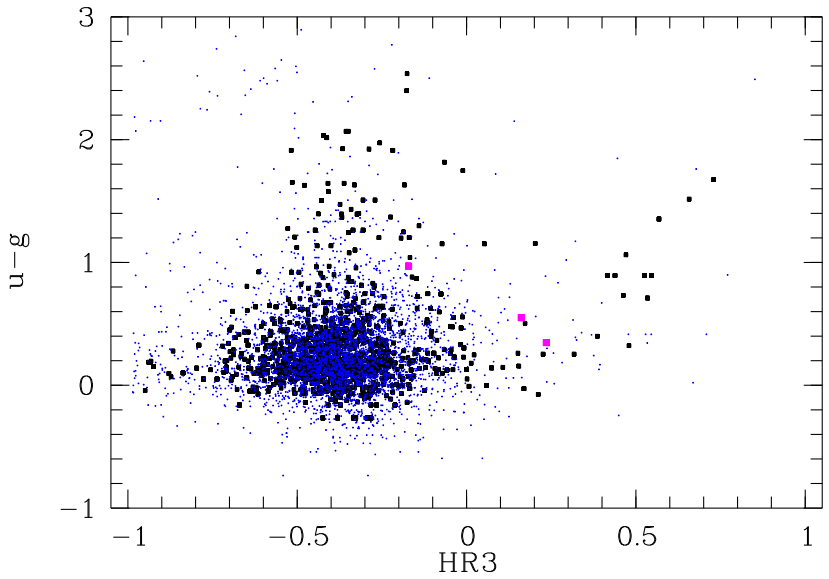


FIG. IV.18 – Variation du HR3 en fonction de la couleur $u-g$ pour les quasars identifiés spectroscopiquement par le SDSS (carrés noirs), les candidats QSO de type II de Zakamska et al. (carrés magenta) et les objets ponctuels issus du catalogue photométrique. Seules les sources présentant une erreur sur le HR3 inférieure à 0.2 et une erreur sur $u-g$ inférieure à 0.3 ont été tracées. Nous avons également ignoré les sources ayant un $\log(f_x/f_r)$ < -1 pour exclure les étoiles.

d Utilisation des raies optiques

Nous avons récupéré via CasJobs les raies des sources optiques observées spectroscopiquement par le SDSS et corrélées aux sources du 2XMMi. Ces sources sont presque exclusivement des sources extragalactiques. Les diagrammes basés sur des rapports de raies permettent de faire des diagnostics fins sur la nature des différents objets. Le diagramme [O III]/H_β versus [N II]/H_α est notamment utilisé pour séparer les galaxies à formation d'étoiles (*star forming galaxies*) des galaxies actives (Stasińska *et al.* 2006, 2008). Ces diagrammes présentent pour la plupart l'avantage d'être peu sensibles au rougissement. Comme dans Stasińska *et al.* (2008), nous avons sélectionné les sources spectroscopiquement identifiées comme des galaxies par le SDSS, ayant une magnitude r comprise entre 14.5 et 17.77 et un *redshift* inférieur à 0.002. Le diagramme [O III]/H_β versus [N II]/H_α de cette échantillon est présenté Fig. IV.19. Comme on pouvait s'y attendre, la branche des galaxies de Seyfert apparaît très nettement. La branche des galaxies à formation d'étoiles (SF) est assez bien définie alors que seuls quelques objets peuplent la branche des LINER. Les sources en dessous de la branche des galaxies SF et situées entre les deux branches principales (SF et Seyfert) sont probablement des *retired*[2] galaxies pauvres en métaux.

Les raies émises par les galaxies SF viennent des régions H II. Ces régions H II sont photoionisées par des étoiles massives alors que les zones centrales des AGN sont photoionisées par des photons plus énergétiques venant du noyau actif. On observe des raies collisionnelles ([O III], [O II], [N II], [S II]) venant directement de la photoionisation et des raies de recombinaison ($H\alpha$, H_β) quand les atomes photoionisés captent des électrons libres du milieu. Pour un rapport [O III]/H_β ou [O III]/[O II] donné, les AGN auront un rapport [O II]/H_β, [N II]/H_α, [S II]/H_α ou [O I]/H_α plus important (Stasińska *et al.* 2006).

Ce type de diagramme de diagnostic basé sur l'information spectrale pourrait à l'avenir permettre une meilleure séparation entre AGN et galaxies dans l'échantillon d'apprentissage. On pourrait également ajouter une classe galaxies SF. Cela permettrait une classification plus fine des sources XMM/DR7.

[2] Stasińska *et al.* (2008) nomme galaxies à la retraite, ou *retired galaxies*, des galaxies qui ne forment plus d'étoiles et possèdent une population stellaire âgée

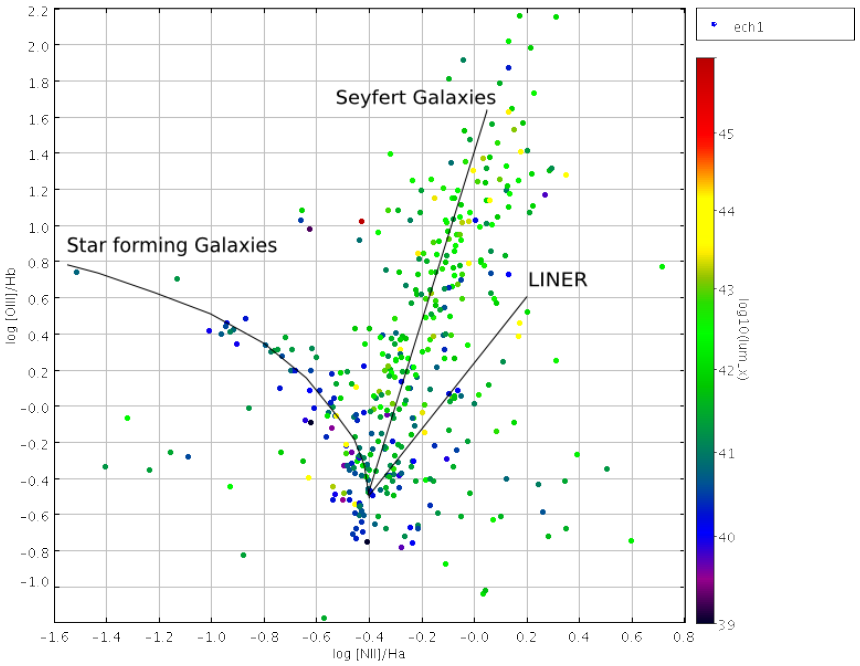


FIG. IV.19 – Diagramme [O III]/H_β vs [N II]/H_α des sources spectroscopiques du SDSS.

1.4 ACP SUR L'ÉCHANTILLON XMM/SDSS

Nous avons appliqué l'ACP améliorée (voir chapitre III, §1.2) sur 24 paramètres de l'échantillon 2XMMi/SDSS DR6. La DR6 est la version du SDSS antérieure à la DR7. Ces 24 paramètres sont composés de 5 rapports Fx/Fopt (pour les 5 bandes de magnitudes optiques), des 4 HR d'XMM, de 10 couleurs SDSS (toutes les combinaisons possibles entre les différentes bandes de magnitudes), des 5 magnitudes SDSS et du paramètre d'extension. Le diagramme de gauche sur la figure IV.20 montre les valeurs des coefficients de corrélations entre les paramètres d'entrée et les deux premiers axes principaux. On voit apparaître une tendance forte qui oppose les objets

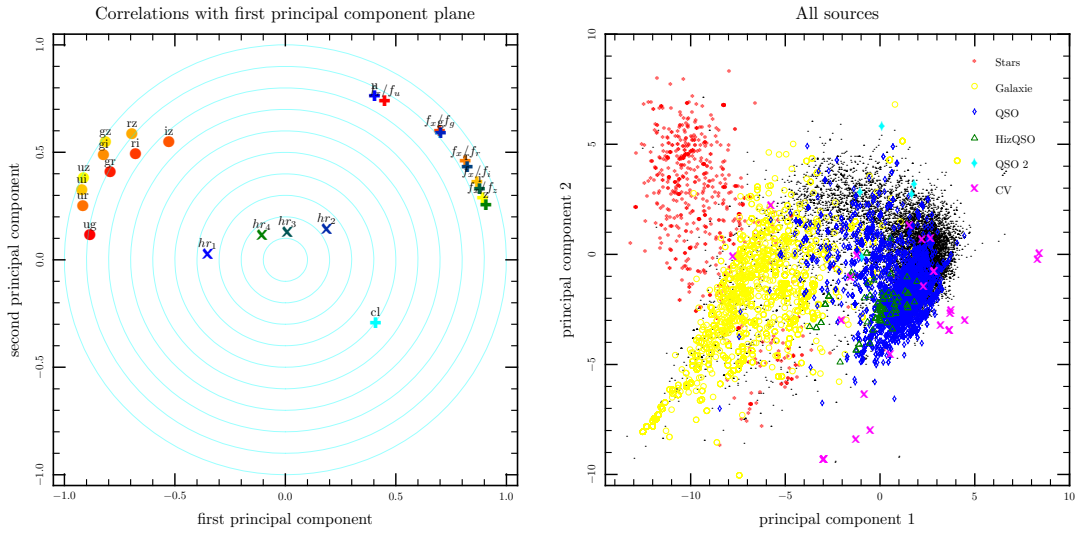


FIG. IV.20 – Résultats de l'ACP sur l'échantillon 2XMMi/DR6. Gauche : correlations entre les paramètres d'entré et les deux premiers axes principaux. Droite : projection des données sur les deux premiers axes principaux.

rouges aux objets qui ont un rapport Fx/Fopt élevé. Si on fait pivoter le diagramme de 45 degrés vers la gauche, l'axe vertical serait complètement dominé par les rapport Fx/Fopt et l'axe horizontale par les couleurs optiques.
Il est également important de noter que les quatre rapports de dureté X ne rentre quasiment pas dans la composition de ces deux premiers axes principaux.
La diagramme de droite de la figure IV.20 montre la projection des sources sur le plan composé des deux premiers axes principaux. On constate que ce diagramme est très proche du diagramme Fx/Fr versus $g-i$ vu précédemment. La population d'objets à forts rapports Fx/Fopt, mais trop faibles en optique pour avoir été observés spectroscopiquement par le SDSS ressort particulièrement bien. Si on regarde ces données en 3D avec pour troisième axe la troisième composante principale, on se rend compte qu'une grande partie des binaires X forme une coquille entourant le *locus* des QSO.

1.5 Classification de l'échantillon XMM/DR7

Nous avons classé les sources issues de la corrélation entre le 2XMMi et le SDSS DR7. L'algorithme utilisé est le *kernel density classification* (KDC) pour lequel le *kernel smoothing* a été réalisé avec une largeur de bande fixe de 0.5. Nous avons pris pour paramètres d'entrée le rapport Fx/Fr, les quatre couleurs optiques $u-g$, $g-r$, $r-i$ et $i-z$ et le paramètre d'extension cl. Dans la DR7, cl vaut 3 si les sources sont étendues et 6 si elles sont non-résolues. De fait, la distance entre un objet étendu et un objet ponctuel sera toujours supérieure (ou égale) à 3, valeur plus grande que la largeur de bande. Les objets entrant dans le calcul de densité des objets respectivement étendus et ponctuels seront donc uniquement étendus dans le premier cas et uniquement ponctuels dans le second.

En sortie de cette classification, nous obtenons 1 187 étoiles, 35 binaires X, 1 537 galaxies, 3 428 AGN et 8 291 QSO. L'échantillon d'apprentissage était constitué de 536 étoiles, 26 binaires X, 673 galaxies, 572 AGN et 1 425 QSO. Le tableau IV.11 présente la matrice de confusion de l'échantillon d'apprentissage total (voir §4.2). La figure IV.21 montre le résultat de la classification dans le diagramme Fx/Fr / $g-i$. On voit que le taux de bonnes identifications des étoiles atteint

TAB. IV.11 – Matrice de confusion pour la classification de l'échantillon 2XMMi/DR7 sur les paramètres Fx/Fr, $u-g$, $g-r$, $r-i$, $i-z$ et cl.

Inital\Attribuée	Étoile	Binaire X	Galaxie	AGN	QSO
Étoile	535	0	0	0	1
Binaire X	1	13	0	0	12
Galaxie	0	0	620	45	8
AGN	0	0	75	312	185
QSO	0	0	7	74	1344

99.81% avec une seule étoile de l'échantillon d'apprentissage mal classée. Le taux d'identification des sources extragalactiques atteint lui 100% mais avec de la confusion entre les différentes classes. En effet, 92.12% de galaxies sont bien classées, mais seulement 54.55% des AGN et 94.32% des QSO. Ces taux peuvent s'expliquer par le fait que les AGN forment une zone de transition entre les galaxies et les QSO et peuvent donc être confondus avec des objets de ces deux classes. On voit enfin que seulement 50% des binaires X de l'échantillon d'apprentissage ont été correctement classées. A l'exception d'une binaires X classée comme étoile, elles ont majoritairement été confondues avec des QSO. Il faudrait pour les binaires X une classification particulière, adaptée aux *outliers*, comme celle proposée plus loin au §4.6.

Nous avons également essayé sur cet échantillon la classification par l'algorithme des *mean-shifts* (voir III, §4) avec une largeur de bande fixe valant 0.5, comme pour la KDC. Si on exclut les points qui correspondent à des *outliers*, on trouve 10 maxima locaux dont la répartition des types d'objet associés est donnée dans le tableau IV.12. La matrice de confusion est cependant légèrement moins bonne que pour la KDC avec par exemple un taux d'étoiles bien identifiées de 98.5% contre 99.81% avec la KDC. Ce taux tombe à 19,2% pour les binaires X contre 50% avec la

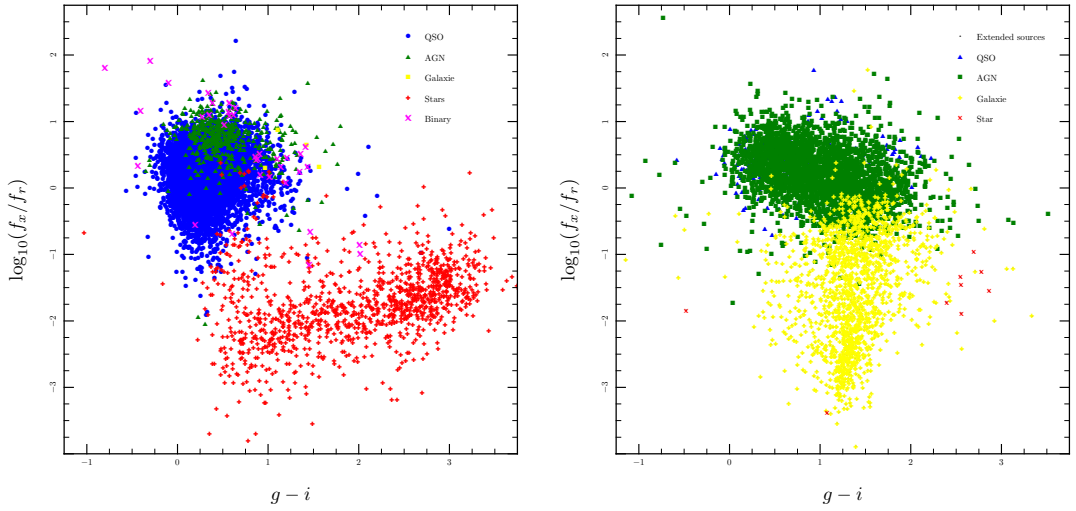


FIG. IV.21 – Diagrammes $log(f_x/f_r)$ versus $g-i$ de l'échantillon XMM/DR7 classé. À gauche : ensemble des sources non résolues, la couleur code le type d'objet. À droite : ensemble des sources étendues, la couleur code le type d'objet.

TAB. IV.12 – Nombre d'objets de chaque type associé aux 10 maxima locaux trouvés par l'algorithme des *meanshifts* sur l'échantillon 2XMMi/DR7. « classe » désigne la classe attribuée au point de *meanshift*. On attribue cette classe à tous les objets qui tombent sur ce même point.

Point\Source	Classe	Unknown	Étoile	Binaire X	Galaxie	AGN	QSO
Point 1	Étoile	24	10	0	0	0	1
Point 2	Étoile	44	3	0	0	0	0
Point 3	Étoile	344	451	0	0	0	0
Point 4	Étoile	175	52	0	0	0	0
Point 5	Étoile	12	2	0	0	1	1
Point 6	Galaxie	101	0	0	50	14	1
Point 7	Galaxie	326	0	0	432	24	2
Point 8	Galaxie	52	0	0	35	4	0
Point 9	AGN	3015	0	0	141	291	77
Point 10	QSO	7024	0	21	9	230	1337

KDC. On a donc choisi de conserver la classification obtenue avec la KDC qui présente également l'avantage de fournir pour chaque objet les probabilités d'appartenance à chaque classe.

2 Corrélations du 2XMMi, du 2MASS et GSC2.2.1

2.1 CRÉATION DE L'ÉCHANTILLON 2XMMI-GSC2-2MASS

2.1.1 Corrélation du 2MASS

Nous avons corrélé le catalogue 2MASS avec le catalogue 2XMMi en suivant la même méthode que pour le SDSS DR7, à savoir celle décrite au chapitre II. On trouve un total de 99 392 associations dont 30 369 ont une probabilité d'identification supérieure à 90%, ce qui représente respectivement 55 461 détections et 21 285 sources XMM uniques, soit respectivement 27.7% et

11% de l'ensemble des sources non étendues et ayant un erreur *radec_err*<0.5" dans le 2XMMi (voir chapitre I, §2.3.1). Les fonctions de complétude et de fiabilité en fonction du seuil en probabilité d'identification sont présentées Fig. IV.22.

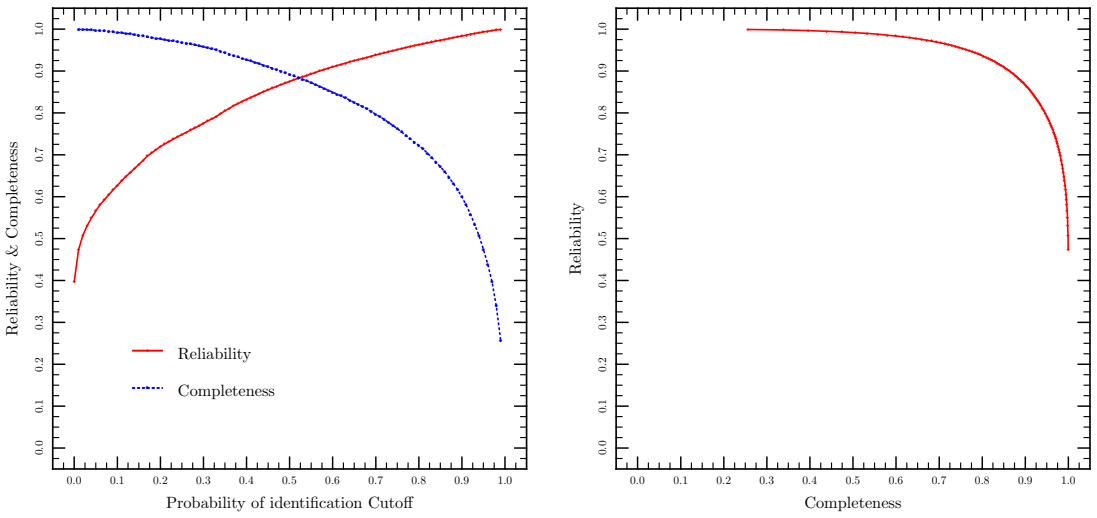


FIG. IV.22 – Pourcentage de fiabilité et complétude en fonction du seuil en probabitité d'identification (gauche). Fiabilité versus complétude (droite).

2.1.2 Corrélation du GSC2

Dans la version que nous présentons ici, le catalogue est le GSC 2.2.1[3] et les corrélations utilisées proviennent de l'ACDS. Pour une première sélection, nous avons fixé une limite de 5 secondes d'arc entre une source XMM et une source GSC2.

2.1.3 Échantillon 2XMMi-GSC2-2MASS

L'erreur *rms* sur les positions du catalogue GSC2, erreurs systématiques comprises, vaut environ 0.4 arcsec pour les sources faibles et 0.28 pour les objets plus brillants que magnitude 18.5 en r (voir Lasker *et al.* 2008, Tab. 14). On nomme cette erreur $\sigma_{\rm GSC2}$. Celle sur les positions des sources du catalogue 2MASS, notée $\sigma_{\rm 2MASS}$, vaut environ 0.1 arcsec (voir Skrutskie *et al.* 2006, Fig. 20). Il ressort que la majorité des sources GSC2 que l'on a sont des sources plus brillantes que magnitude 18.5 en r. Si on corrélait l'ensemble des sources GSC2 et 2MASS, on devrait obtenir une fonction de Rayleigh de paramètre $\sigma = \sqrt{\sigma^2_{\rm GSC2} + \sigma^2_{\rm 2MASS}}$, soit $\sigma = 0.3$ arcsec. La fonction de Rayleigh est définie comme suit :

$$\text{Rayleigh}(r,\sigma) = \frac{r\exp(-\frac{r^2}{2\sigma^2})}{\sigma^2} \tag{IV.2}$$

Soit en cumulé :

$$\int_0^R \text{Rayleigh}(r,\sigma)dr = 1 - \exp(-\frac{r^2}{2\sigma^2})$$

Pour avoir une complétude de x%, le rayon maximum doit donc valoir :

$$r = \sigma\sqrt{-2\ln(1-x/100)}$$

En choisissant le critère classique de 3σ, correspondant à une complétude de 99,7% on obtient un rayon maximum $r = 1$".

[3] Catalogue référencé sous l'identifiant I/271 dans ViZieR.

Au final, nous avons choisi les sources 2MASS ayant une probabilité d'identification calculé par la méthode décrite au chapitre II (voir Eq. II.19) supérieure à 0.6 et une distance maximum à la source XMM de 5". Cette limite supérieure sur la probabilité correspond à une fiabilité de 90% et une complétude de 0.86%.
Nous avons ensuite uniquement retenu les candidats GSC2 situés à moins de une seconde d'arc des sources 2MASS, en accord avec le critère de complétude sur la loi de Rayleigh.
Enfin, nous nous sommes concentrés sur les sources situées à plus de 10°de latitude galactique afin d'éviter les problèmes de rougissement liés à l'absorption galactique.

Notre échantillon, que l'on nomme XGT, contient au final 9 725 sources. L'échantillon d'apprentissage correspondant, dérivé de LX_X (voir §1.2) et nommé LX_XGT, contient 531 étoiles, 46 binaires X, 672 galaxies, 320 AGN et 286 QSO. Parmis les 9 725 sources de XGT, 7 870 sont non identifiées et les autres sont incluses à LX_XGT.

Les magnitudes b vont jusqu'à 19.5 et les magnitudes r jusqu'à 18.5 avec une limite de complétude autour de 17.25 pour ces dernières.

2.2 Analyse sommaire de l'échantillon

Nous avons réalisé une ACP en tenant compte des erreurs de mesure (voir chapitre III, §1.2) sur 19 des paramètres des sources multilongueurs d'onde de l'échantillon 2XMMi-GSC2.2.1-2MASS (XGT). Ces 19 paramètres sont :

- les rapports Fx/Fopt : Fx/Fr et Fx/Fb ;
- les rapports Fx/Fir : Fx/Fj, Fx/Fh et Fx/Fk ;
- les rapports de dureté X : $HR1$, $HR2$, $HR3$ et $HR4$;
- la couleur optique $b-r$;
- les couleurs IR : $j-h$, $h-k$ et $j-k$;
- les couleurs optique–IR : $r-j$, $r-h$, $r-k$, $b-j$, $b-h$ et $b-k$.

Les résultats de l'ACP ont permis de faire ressortir l'importance des rapports Fx/Fopt et des couleurs obtenues par différence entre une magnitude optique et une magnitude infrarouge d'un coté et les couleurs IR de l'autre. Ainsi, les trois paramètres $b-j$, Fx/Fb et $j-k$ dominent chacun dans l'une des trois premières composantes principales. La figure IV.23 présente les diagrammes Fx/Fb versus $b-j$ et $b-j$ versus $j-k$.

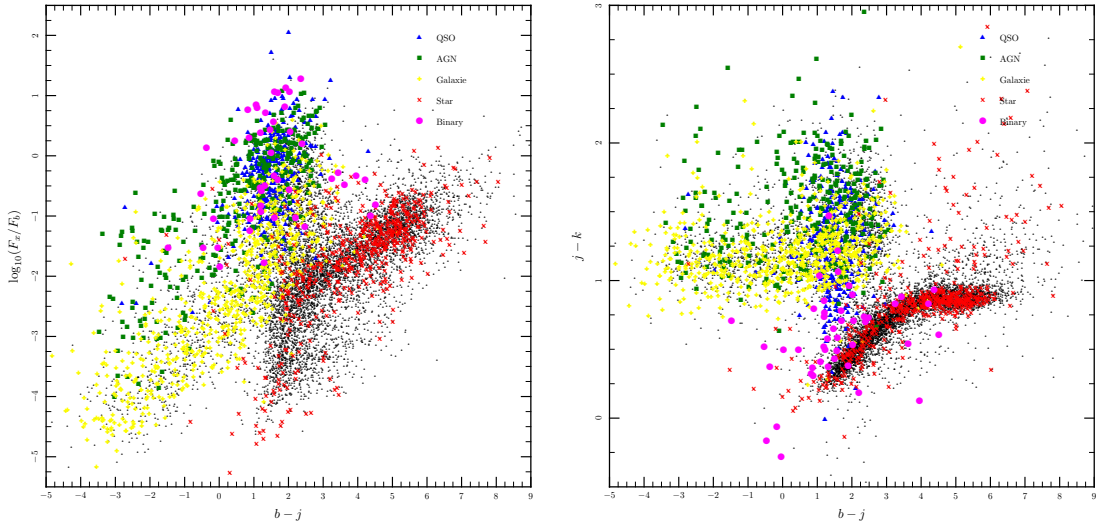


FIG. IV.23 – Diagrammes Fx/Fb — $b-j$ et $b-j$ — $j-k$ de l'échantillon XGT.

Dans le diagramme Fx/Fb versus $b-j$, la distribution se présente sous forme de deux coudes. Dans le premier on trouve en bas à gauche les galaxies, en haut au centre les QSO et les AGN se repartissent à la fois dans les QSO et dans les galaxies mais avec un rapport Fx/Fb typique plus

élevé que ces dernières. Dans le deuxième coude, on trouve uniquement des étoiles. Les binaires X se superposent globalement aux QSO et à la partie la plus active des AGN. Quelques-unes cependant sont plus rouges avec un rapport Fx/Fb supérieur à celui des étoiles.
Il ressort de ce diagramme une séparation assez franche entre étoiles et sources extragalactiques avec un léger recouvrement au milieu des deux coudes. Cependant, dans l'espace des paramètres constitué des trois premières composantes principales issues de l'ACP, ce recouvrement disparaît presque complètement. On peut également le constater dans le diagramme $b-j$ versus $j-k$. Dans ce diagramme, les galaxies et les AGN sont bien séparés des étoiles et en grande partie aussi des binaires X. Les binaires X se séparent également mieux des QSO même s'il reste une zone de superposition entre les deux types d'objets pour $j-k<1.25$. On voit que la gamme des couleurs $j-k$ est assez importante pour les QSO. En traçant le diagramme $j-k$ versus $b-r$ on ne trouve pas de corrélation entre couleurs rouges en optique et en IR. Pour $b-j>4$, on observe une population d'étoiles dont la couleur $j-k$ dépasse l'unité. Sur quatre sources choisies aléatoirement, toutes sont marquées comme étoiles à émission dans Simbad (identifiant : Kiso A-0976 312, V* VY Cha, EM* LkHA 271 et NAME SSS 107) et correspondent à des étoiles de type spectral M0 ou K. On peut donc proposer, à partir de ce diagramme, un certain nombre de candidats stellaires à raies d'émissions de type spectral K ou M0 non recensés dans Simbad et dont le rapport Fx/Fb peut atteindre le même niveau que les valeurs bases des QSO et des binaires X.

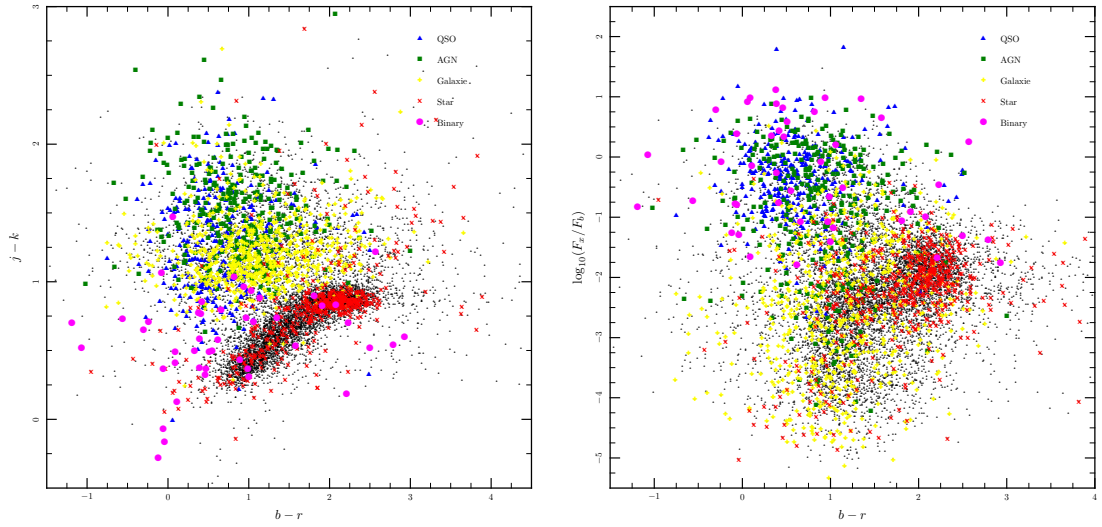


FIG. IV.24 – Diagrammes $j-k$ — $b-r$ et Fx/Fb — $b-r$ de l'échantillon XGT.

Les étoiles les plus rouges en optique ont tendance à avoir les couleurs Optique-IR les plus grandes, les couleurs IR les plus grandes et de même pour les rapports Fx/Fopt. L'évolution de la couleur IR $j-k$ est cependant quasiment plate pour des valeurs de $b-j>3$ et $b-r>1.8$, qui correspond aux étoiles les plus rouges (Fig. IV.24).
On observe également une augmentation du rapport Fx/Fopt avec $b-r$. C'était déjà le cas avec le SDSS et on retrouve donc logiquement ce résultat avec les données GSC2. On l'explique par le fait que plus une étoile est tardive, donc froide, peu massive, et rouge, plus le flux optique qu'elle émet est faible. Or les mécanismes de production de rayons X qui ont lieu dans la couronne ne sont pas directement liés aux processus étant à l'origine du rayonnement optique. Si on considère grossièrement que l'émission X d'une couronne active est constante en fonction de la classe de l'étoile tardive, le flux optique diminuant avec la température, on constate logiquement une augmentation du rapport Fx/Fopt.

2.3 Résultat de la Classification

Nous avons classé les sources de l'échantillon XGT en utilisant une *kernel density classification* dont le *kernel smoothing* a été réalisé avec une largeur de bande fixe de 0.75 dans l'espace des paramètres à trois dimensions : $\log_{10}(Fx/Fb)$, $b-j$ et $j-k$. Les résultats sont présentés Fig. IV.25. Au total, 6 050 sources ont été classées comme étoiles, 23 comme binaires X, 2 391 comme galaxies,

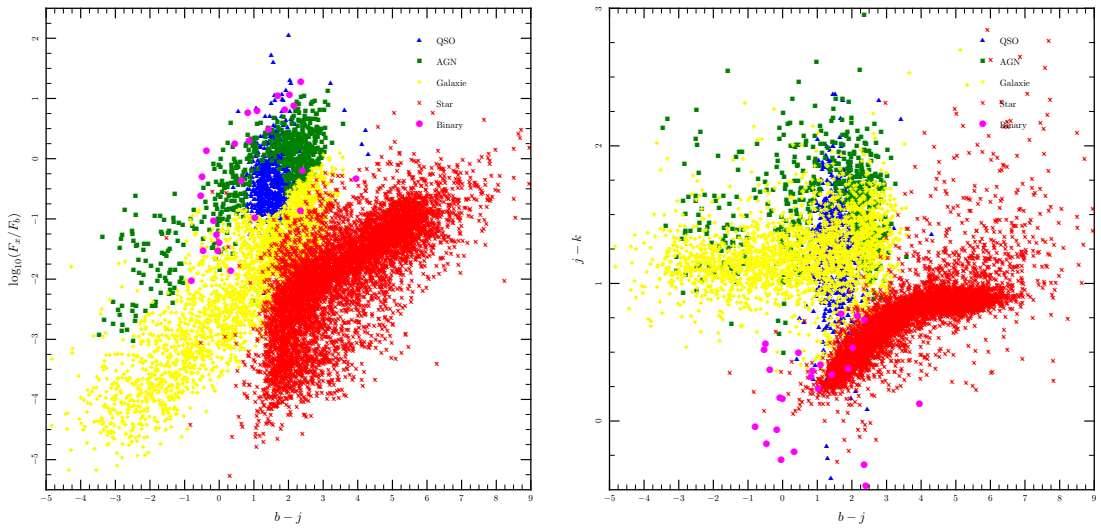


FIG. IV.25 – Diagrammes Fx/Fb versus $b-j$ et $b-j$ versus $j-k$ de l'échantillon 2XMMi-GSC2-2MASS classé. Classification faite sur les paramètres Fx/Fb, $b-j$ et $j-k$.

628 comme AGN et 633 comme QSO.

La classification appliquée uniquement à l'échantillon d'apprentissage (LX_XGT) permet le calcul de la matrice de confusion. Cette matrice permet de contrôler la qualité de l'auto-classification de l'échantillon d'apprentissage (EA) LX_XGT. Elle est donnée Tab. IV.13.

TAB. IV.13 – Matrice de confusion pour la classification de l'échantillon LX_XGT sur les paramètres Fx/Fb, $b-j$ et $j-k$.

Initiale\Attribuée	Étoile	Binaire X	Galaxie	AGN	QSO
Étoile	514	1	12	1	3
Binaire X	7	14	3	2	20
Galaxie	10	0	584	33	45
AGN	1	0	73	173	73
QSO	1	2	47	61	175

Le taux de bonne identification des étoiles atteint 97%, seulement 30% pour les binaires X et 99% pour les sources extragalactiques. Il y a cependant de la confusion entre sources extragalactiques : 87% de galaxies de l'EA sont bien identifiées contre 61% des QSO et 54% des AGN. La confusion se trouve principalement entre QSO et AGN avec 21% de QSO de l'EA classés comme AGN et 23% d'AGN de l'EA classés comme QSO. Ces deux classes se confondent également avec les galaxies : 16% des QSO de l'EA sont classés comme galaxies et 23% pour les AGN. Nous pensons que le nombre d'objets classés comme galaxies est légèrement trop important en raison de la sur-représentation des galaxies dans l'échantillon d'apprentissage. On pourrait diminuer cet effet en jouant sur les probabilités a priori en entrée de la *kernel density classification*. Pour cet exemple, nous avons pris comme valeur par défaut de la probabilité a priori d'avoir un objet de classe donnée le nombre de sources de cette classe dans l'EA sur le nombre total de sources dans l'EA. La probabilité a priori d'avoir une galaxie est surestimée puisque les galaxies sont deux fois plus

nombreuses que les AGN et que les QSO dans l'EA.

Nous avons également classé les sources à partir du résultat de l'ACP. Pour avoir le même nombre de paramètres, nous avons choisi de garder uniquement les 3 premières composantes principales. Celle-ci représentent 76% de la variance totale. Les 5 premières composantes principales (CP) représentent au total plus de 86% de la variance totale des données. En utilisant le critère classique qui consiste à ne conserver que les CP dont la valeur propre associée dépasse l'unité, nous aurions conservé les quatre premières pour une variance de près de 82% de la variance totale. Nous avons appliqué sur les 3 premières composantes principales la *kernel density classification* avec pour largeur de bande fixe la valeur 1.0. Au total, 6 136 sources ont été classées comme étoiles, 62 comme binaires X, 2 172 comme galaxies, 689 comme AGN et 666 comme QSO. Les résultats sont donnés figure IV.26 et la matrice de confusion est présentée table IV.14. Les

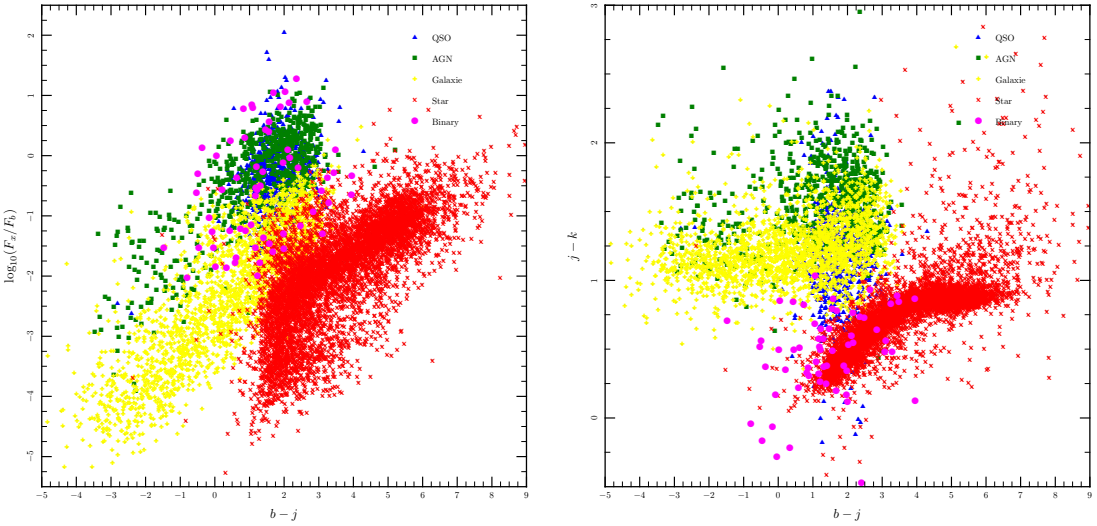


FIG. IV.26 – Diagrammes Fx/Fb versus $b-j$ et $b-j$ versus $j-k$ de l'échantillon XGT classé. Classification faite sur les trois premières composantes principales.

TAB. IV.14 – Matrice de confusion pour la classification de l'echantillon LX_XGT sur les trois premières composantes principales.

Initiale\Attribuée	Étoile	Binaire X	Galaxie	AGN	QSO
Étoile	518	1	8	0	4
Binaire X	5	27	2	1	11
Galaxie	8	0	601	31	32
AGN	0	0	61	207	52
QSO	1	3	30	66	186

résultats sont meilleurs avec 98% des étoiles de l'EA bien identifié, 59% pour les binaires X, 89% pour les galaxies, 65% pour les AGN et les QSO. Remarque : si une source est trop éloignée des sources de l'échantillon d'apprentissage pour mesurer une densité de probabilité par *kernel smoothing*, on utilise un moyennage par k-NN avec k=7. Ce cas s'est présenté pour 30 sources dans la classification à partir des trois premiers axes principaux.
Pour éviter ce genre d'astuce, l'utilisation d'un *kernel smoothing* à largeur de bande variable pourrait être bénéfique. On pense notamment à l'estimateur *sample-point* (voir chapitre II, §5.5.1). Cependant, les temps de calculs sont beaucoup plus long en l'état actuel du code.

Les résultats de la classification sur les trois premiers axes principaux de l'ACP sont stockés dans la XCAT-DB et accessibles par la communauté (voir chapitre V, §3 pour plus de détail).

3 Classification sur les propriétés X uniquement

Si on regarde la taille des échantillons précédents, on remarque que la majorité des sources du catalogue 2XMMi n'en fait pas partie. Pour le SDSS (échantillon XS), la raison principale est que seule une fraction des champs de vue XMM sont couverts par la DR7 du SDSS. Dans les champs couverts, presque 60% des sources X ont une contrepartie optique (voir §1.1).
Pour l'échantillon XGT (voir §2.1.3), on se restreint aux hautes latitudes galactiques pour éviter les effets du rougissement dû à la poussière du milieu interstellaire. Cet échantillon est limité par les complétudes du GSC2 et du 2MASS.
Le nombre de sources dans XS et XGT diminue encore suite aux différents critères de qualité imposés.

Pour classer les sources qui ne font partie d'aucun de ces échantillons, plusieurs stratégies sont possibles et pourraient être mises en place.
Premièrement, on pourrait augmenter la taille des échantillons en mettant des critères moins sévères. La limite inférieure de 90% sur la probabilité d'identification des sources XS pourrait par exemple être abaissée. De même pour les contraintes sur certains *flags* de qualité.
Deuxièmement, on pourrait diminuer le nombre de paramètres des échantillons. Par exemple, dans le cas de XGT, toutes les sources ont à la fois leur magnitude r et b définie alors que ce n'est pas le cas de toutes les sources du catalogue GSC2. On pourrait alors créer deux jeux de données : l'un avec les sources dont la magnitude r est définie ; l'autre avec les sources dont la magnitude b est définie. Dans le cas particulier de l'échantillon XGT, cette stratégie n'est pas pénalisante dans la mesure où la couleur optique ne joue pas un rôle prépondérant dans la classification. Les couleurs optique–IR sont plus importantes. Nous avons eu l'occasion de tester cette stratégie sur les données ROSAT–GSC2–2MASS. Cette solution permet effectivement de classer des sources supplémentaires. La double classification des sources communes aux deux échantillons (les sources ayant leurs deux magnitudes définies) peut alors fournir des renseignements sur la qualité de la classification.
Troisièmement, un peu comme le point précédent, on peut diminuer le nombre de catalogues corrélés. On a par exemple essayé de classer les sources résultant uniquement de la corrélation entre le 2XMMi et le GSC2 d'un coté et de la corréaltion entre le 2XMMi et le 2MASS d'un autre. Le nombre de paramètres est dans ce cas réduit et les différents types de sources se séparent mal dans l'espace des paramètres (voir par exemple panneau de droite de la figure IV.24 pour se faire une idée).

Nous avons classé ici les sources du 2XMMi uniquement à partir de leurs propriétés X. Pour cela, les quatre rapports de dureté ont été utilisés. Nous avons sélectionné toutes les sources du catalogue 2XMMi dont les valeurs de HR1, HR2, HR3 et HR4 sont définies avec une erreur pour chacun de ces rapports inférieure à 0.2. Nous avons posé comme contrainte une valeur minimale de 3 sur le *flag* de qualité *sc_sum_flag*. Cette sélection retourne au total 30 620 des 221 012 sources X uniques que contient le catalogue.
Nous avons classé ces sources via une *kernel density classification* en prenant pour valeur de largeur de bande l'erreur maximale autorisée sur les paramètres, soit 0.2. Nous n'avons pas utilisé l'ACP car le nombre de paramètres était déjà faible. On peut visualiser en 2D l'échantillon d'apprentissage et le résultat de la classification sur la figure IV.27. Pour une meilleur visualisation en 2D de la séparation entre les classes, nous avons fait figuré les grandeurs HR1 et HR2+HR3. La classification a cependant bien été faite sur HR1, HR2, HR3 et HR4.

L'échantillon d'apprentissage contient 751 étoiles, 159 binaires X, 389 galaxies, 984 AGN et 1 304 QSO. Les résultats de la classification donnent en sortie 6 568 étoiles, 630 binaires X, 1 865 galaxies, 6 448 AGN et 15 109 QSO. On trouve donc que environ 20% des sources sont des étoiles et environ 60% sont liées à des trous noir supermassifs (AGN + QSO). La proportion d'étoile trouvée ici est en dessous de la fraction d'étoiles optiquement identifiées dans le plan galactique (entre 27% et 40% Motch 2006*b*), et de celle observée à plus hautes latitudes galactiques dans la bande 0.2-2 keV (environ 30%, Mateos *et al.* 2008). On peut difficilement comparer ces chiffres dans l'état. Il faudrait d'abord étudier l'effet de la sélection sur les HR et leurs erreurs.

La matrice de confusion résultant de cette classification est donnée Tab. IV.15. On constate que 93.21% des étoiles de l'échantillon d'apprentissage ont été correctement classées. Ce pourcentage

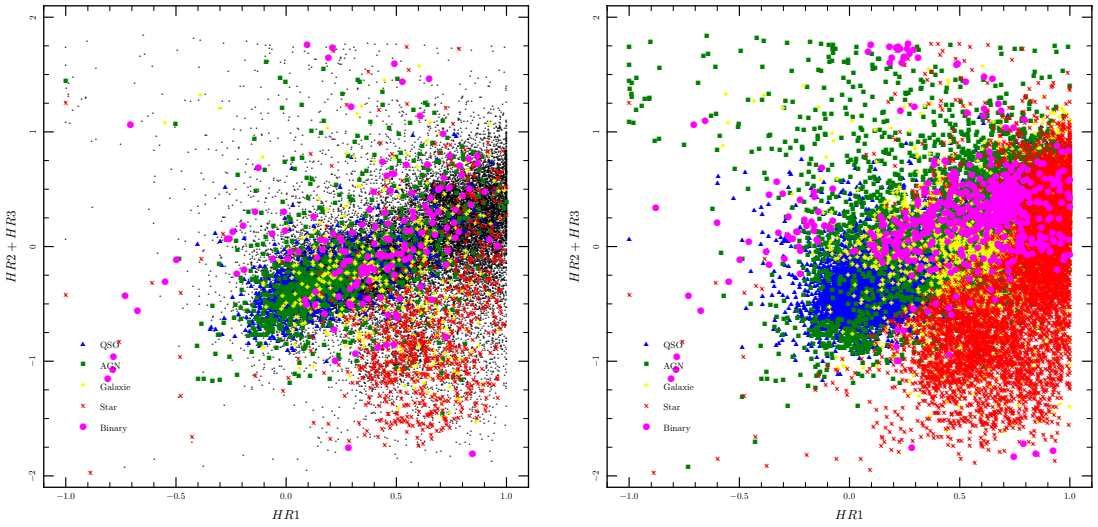


FIG. IV.27 – Diagramme HR1 vs HR2+HR3 des sources du 2XMMi avant (gauche) et après classification (droite).

vaut 93.40% pour les QSO. En revanche, il tombe à 63.81% pour les AGN, 40.87% pour les galaxies et 35.22% pour les binaires X.
À partir de la matrice de confusion, on voit que les AGN sont majoritairement classés comme QSO. Cela est dû au fait qu'ils occupent en grande partie le même locus que les QSO dans l'espace des paramètres composé des quatre rapports de dureté. Or la probabilité a priori des QSO injectée dans la KDC est supérieure à celle des AGN. À valeur de distribution équivalente, une source est donc préférentiellement classée comme QSO.

TAB. IV.15 – Matrice de confusion pour la classification des sources du 2XMMi uniquement sur leurs rapports de dureté.

Initale\Attribuée	Étoile	Binaire X	Galaxie	AGN	QSO
Étoile	700	0	3	18	30
Binaire X	14	56	2	33	54
Galaxie	56	2	159	34	138
AGN	18	0	4	647	345
QSO	11	3	6	66	1 218

Toujours grâce à la matrice de confusion, on voit que 3.45% des objets de l'échantillon d'apprentissage sont classés comme étoiles alors qu'ils correspondent à d'autres types d'objet. Pour certain, il se peut qu'ils ne soient pas de la bonne classe en raison de mauvaises associations lors de la construction de l'échantillon d'apprentissage. Mais on voit également qu'il y a de la confusion dans le diagramme HR 4D. Les différentes classes d'objets ne sont pas bien séparées, en partie à cause de la taille importante des boîtes d'erreurs.

Sur la figure IV.28, on a tracé les histogrammes des distributions des différentes probabilités d'appartenance à chaque classe pour les objets de l'échantillon d'apprentissage (bleu) et pour les objets classés (vert). On voit nettement que l'échantillon d'apprentissage stellaire est relativement bien séparé des autres classes d'objets : la majorité des sources qui le compose ont une probabilité *proba_star* supérieure à 0.6. On constate également qu'une grande partie des binaires X de l'échantillon d'apprentissage ont des probabilités très faibles d'être des binaires X selon la KDC. D'après la matrice de confusion, on voit que ces sources sont classées comme AGN ou QSO. Il existe une population de sources classées comme binaires X avec une forte probabilité. Il doit certainement s'agir de sources proches des *outliers*, loin du locus des QSO, des AGN et des Galaxie.

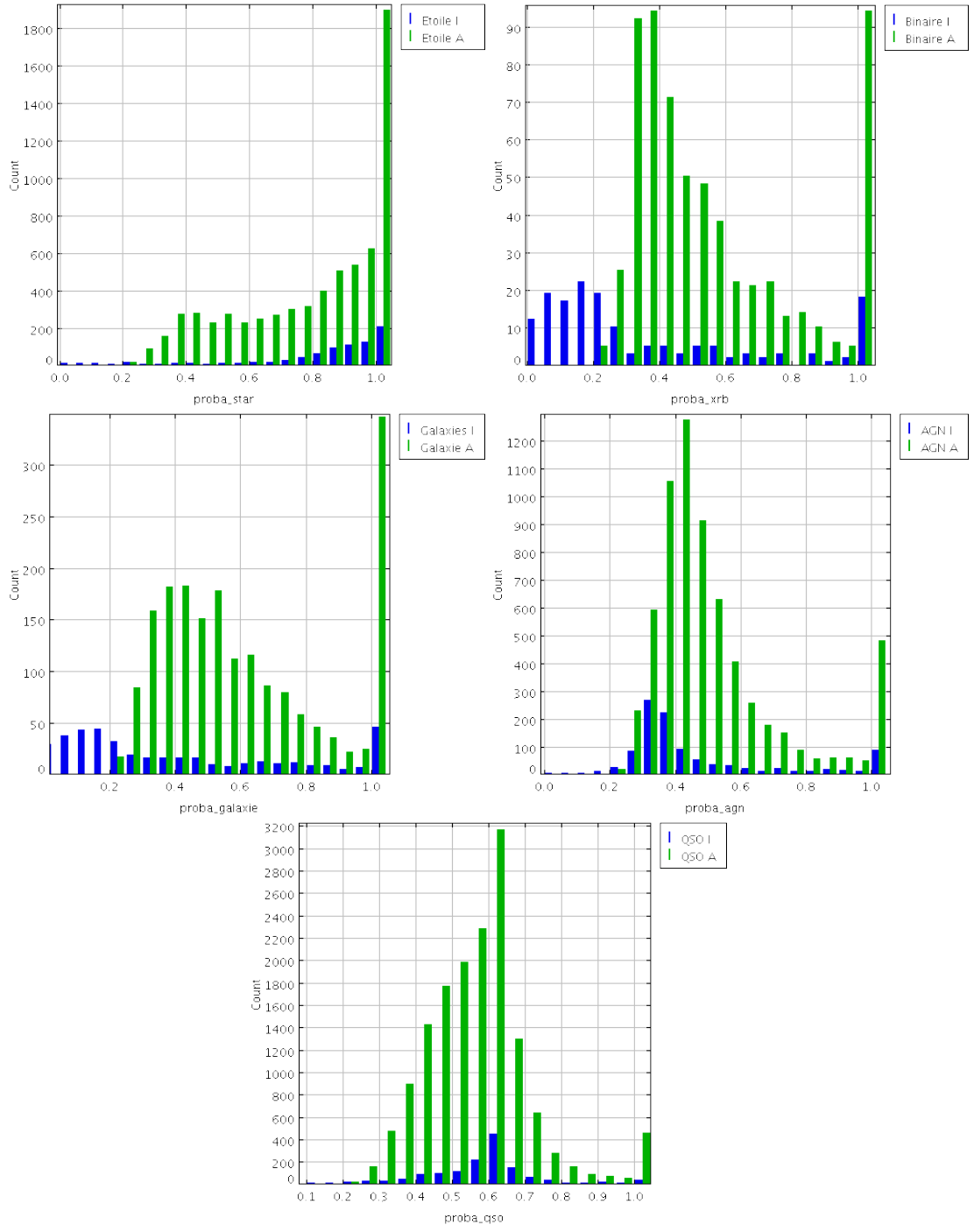


FIG. IV.28 – Histogrammes des probabilités d'appartenance à chaque classe retournées par la KDC. En bleu figurent les sources de l'échantillon d'apprentissage (LS) de chaque classe d'objet (I pour Initiale). En vert figurent les objets classés (A pour Attribuée). En haut à gauche : étoile du LS (bleu) et sources classées comme étoile (vert). En haut au milieu : binaires X du LS (bleu) et sources classées comme binaire X (vert). En haut à droite : galaxies X du LS (bleu) et sources classées comme galaxie (vert). En haut à gauche : AGN du LS (bleu) et sources classées comme AGN (vert). En haut à droite : QSO du LS (bleu) et sources classées comme QSO (vert).

Les remarques faites sur les CV peuvent également être appliquées aux galaxies et aux AGN. Il se peut également que les pics pour les probabilités supérieures à 0.95 correspondent à des parties

de l'espace des paramètres qui ne sont pas couvertes par plusieurs populations d'objets. Grâce à l'histogramme sur les QSO, on voit très nettement que le locus des QSO est partagé avec d'autres objets (AGN, galaxies) mais que la probabilité a priori p(QSO) plus élevée que p(galaxie), p(AGN) et p(XRB), favorise la classification en QSO des objets qui partagent ce même locus.

4 Comparaison de différents algorithmes de classification

Il existe une multitude d'algorithmes de classification. Tous ont leurs avantages et leurs inconvénients. Mais précisons déjà que globalement l'efficacité des algorithmes supervisés réside avant tout dans la qualité de l'échantillon d'apprentissage, et l'efficacité des algorithmes non-supervisés dépend de la séparation des données dans l'espace des paramètres. Dans la plupart des cas, l'utilisateur peut intervenir sur le résultat de la classification en réglant des paramètres d'entrée de l'algorithme : la largeur de bande dans la *kernel density classification* (KDC) et dans le Meanshift (MS), le paramètre λ et la fonction noyau dans les SVM, etc.
Dans cette partie, nous comparons différents algorithmes de classification en les soumettant à plusieurs jeux de données. Nous verrons leurs avantages et leurs inconvénients, leur vitesse d'exécution et leurs résultats.

4.1 Jeux de données en entrée

Nous avons choisi comme point de départ les deux jeux de données présentés dans les sections précédentes de ce chapitre : le premier (XS) est issu des corrélations entre les catalogues 2XMMi et SDSS DR7 ; le second (XGT) provient des corrélations entre les catalogues 2XMMi, GSC2 et 2MASS. Pour chacun de ces deux jeux de données, nous avons utilisé deux échantillons d'apprentissage séparés en deux catégories (étoiles et sources extragalactiques). Le premier contient des CV alors qu'elles ont été retirées dans le deuxième. Nous avons donc quatre échantillons d'apprentissage : deux pour le premier jeu de données et deux pour le second. Les CV sont peu nombreuses et se trouvent parfois près du locus des étoiles, parfois près de celui des QSO. Elles sont donc difficiles à classer et apportent souvent des mauvaises classifications. C'est pour cela que nous avons testé les algorithmes avec et sans les CV.

Nous avons ensuite classé les sources de ces ensembles de données de deux manière différente : une première fois avec l'ensemble de l'échantillon d'apprentissage et la seconde fois avec uniquement les deux tiers de l'échantillon d'apprentissage. Dans le premier cas, la matrice de confusion (*confusion matrix*) a été calculée avec l'ensemble de l'échantillon d'apprentissage. Dans le second cas, elle a été calculée à partir du tiers de l'échantillon d'apprentissage supprimé avant classification. Quand on dit supprimer, cela signifie que les sources sont présentes pour la classification mais sont marquées comme étant de classe non connue. Cette technique qui consiste à séparer l'échantillon d'apprentissage en deux parties dont l'une sert à la classification et l'autre au calcul de la matrice de confusion est appelée dans la littérature *train-test method*. Les sources extraites de l'échantillon d'apprentissage sont choisies de manière aléatoire.

Ces quatre ensembles de données de base ont servi à la création d'autres échantillons dont les différences portent uniquement sur les paramètres qu'ils contiennent (voir Tab. IV.16).

L'échantillon d'apprentissage du jeu de données XS contient 558 étoiles (dont 26 CV) et 1 587 sources extragalactiques pour un total de 7 660 sources non identifiées. Celui du XGT en contient respectivement 544 (dont 43 CV) et 1 311 pour un total de 7 074 sources non identifiées.

4.2 Grandeurs mesurées

Pour ce qui est des grandeurs mesurées, nous avons repris celles utilisées dans Gao *et al.* (2008). Nous redonnons ici leurs définitions.
Pour calculer toutes les grandeurs que nous utilisons, nous avons besoin de calculer la matrice

TAB. IV.16 – Composition des différents échantillons ayant servis à la comparaison des algorithmes de classification.

Noms	Avec CV	LS	Paramétres
XS_3D_1	oui	complet	Fx/Fr, u-g, g-r
XS_3D_2	oui	2/3 1/3	Fx/Fr, u-g, g-r
XS_3D_3	non	complet	Fx/Fr, u-g, g-r
XS_3D_4	non	2/3 1/3	Fx/Fr, u-g, g-r
XS_9D_1	oui	complet	Fx/Fr, HR1, HR2, HR3, HR4, u-g, g-r, r-i, i-z
XS_9D_2	oui	2/3 1/3	Fx/Fr, HR1, HR2, HR3, HR4, u-g, g-r, r-i, i-z
XS_9D_3	non	complet	Fx/Fr, HR1, HR2, HR3, HR4, u-g, g-r, r-i, i-z
XS_9D_4	non	2/3 1/3	Fx/Fr, HR1, HR2, HR3, HR4, u-g, g-r, r-i, i-z
XGT_3D_1	oui	complet	b-r, j-k , r-k
XGT_3D_2	oui	2/3 1/3	b-r, j-k , r-k
XGT_3D_3	non	complet	b-r, j-k , r-k
XGT_3D_4	non	2/3 1/3	b-r, j-k , r-k
XGT_5D_1	oui	complet	Fx/Fr, Fx/Fh, b-r, j-k, r-k
XGT_5D_2	oui	2/3 1/3	Fx/Fr, Fx/Fh, b-r, j-k, r-k
XGT_5D_3	non	complet	Fx/Fr, Fx/Fh, b-r, j-k, r-k
XGT_5D_4	non	2/3 1/3	Fx/Fr, Fx/Fh, b-r, j-k, r-k
XGT_8D_1	oui	complet	Fx/Fr, Fx/Fh, HR2, HR4, b-r, j-k, r-k, b-k
XGT_8D_2	oui	2/3 1/3	Fx/Fr, Fx/Fh, HR2, HR4, b-r, j-k, r-k, b-k
XGT_8D_3	non	complet	Fx/Fr, Fx/Fh, HR2, HR4, b-r, j-k, r-k, b-k
XGT_8D_4	non	2/3 1/3	Fx/Fr, Fx/Fh, HR2, HR4, b-r, j-k, r-k, b-k

de confusion (*confusion matrix*). La matrice de confusion est un histogramme 2D qui contient sur les lignes les classes d'origine des sources de l'échantillon d'apprentissage et sur les colonnes les classes attribuées à ces même sources. Si la classification était parfaite (et si il n'y avait pas de mauvaise classification dans l'échantillon d'apprentissage), seules les cases situées sur la diagonale comporteraient des valeurs non nulles. Le tableau IV.17 représente cette matrice de confusion en présence de deux classes d'objets.

TAB. IV.17 – Définition des grandeurs dans la matrice de confusion. « définie » signifie la classe de l'objet définie dans l'échantillon d'apprentissage et « attribuée » signifie la classe attribuée par l'algorithme de classification. Les grandeurs VC1, FC1, FC2 et VC2 sont des nombres d'objets.

définie\attribuée	Classe 1 attribuée	Classe 2 attribuée
Classe 1 définie	VC1 (Vraie Classe 1)	FC2 (Fausse Classe 2)
Classe 2 définie	FC1 (Fausse Classe 1)	VC2 (Vraie Classe 2)

À partir de cette matrice, on peut définir différentes grandeurs. On garde ici les termes anglais pour les nommer. La justesse, ou *accuracy*, est définie comme suit :

$$\text{Accuracy (Acc.)} = \frac{\text{VC1} + \text{VC2}}{\text{VC1} + \text{FC2} + \text{FC1} + \text{VC2}}$$

Les taux de vraie classe 1 et vraie classe 2 sont définis par :

$$\begin{aligned} \text{Taux de vraie classe 1 (Acc.1)} &= \frac{\text{VC1}}{\text{VC1} + \text{FC2}} \\ \text{Taux de vraie classe 2 (Acc.2)} &= \frac{\text{VC2}}{\text{VC2} + \text{FC1}} \end{aligned}$$

La précision, ou *precision* est définie comme :

$$\text{Precision} = \frac{\text{VC1}}{\text{VC1 + FC1}}$$

et sert dans le calcul de la *F-measure* :

$$\text{F-measure (FM)} = \frac{2 \text{ x Precision x Acc.1}}{\text{Precision + Acc.1}}$$

Enfin la moyenne G et la justesse pondérée sont définie comme suit :

$$\begin{aligned} \text{G-mean (GM)} &= (\text{Acc.1 + Acc.2})^{-1/2} \\ \text{Weighted Accuracy (WA)} &= \beta \text{ x Acc.1} + (1-\beta) \text{ x Acc.2} \end{aligned}$$

En pratique ici, on utilise toujours $\beta = 0.5$.

4.3 ALGORITHMES TESTÉS

Nous avons pour ce test utilisé les six algorithmes suivants, abordés au chapitre III, et dont on rappelle quelques propriétés :

- k-NN : le plus simple de tous les algorithmes ; supervisé, il attribue à chaque source la classe majoritaire parmis ses k plus proches sources voisines dans l'échantillon d'apprentissage ; surtout adapté à une classification binaire, avec k impair pour seul paramètre d'entrée ;
- Meanshift : algorithme non supervisé qui consiste en une montée de gradient dans l'espace des paramètres ; chaque source est associée à un maximum local qui a pour classe la classe majoritaire des sources qui lui sont associées ; prend en paramètre la largeur de bande du *kernel smoothing* sous-jacent ;
- KDC : algorithme supervisé qui fournit pour chaque source ses probabilités d'appartenance à chaque classe ; basé sur les probabilités Bayesiennes, il utilise des rapports de densités locales calculées à partir d'un *kernel smoothing* (pour obtenir les densités de probabilité de chaque classe) et la distribution a priori du nombre d'objets de chaque type dans l'échantillon total ; prend en paramètre la largeur de bande du *kernel smoothing* et la distribution a priori (éventuellement calculée itérativement) ;
- LVQ : algorithme supervisé basé sur les réseaux de Kohonen ; a chaque neurone est associé un type, les sources de l'échantillon d'apprentissage de même type ont un effet attractif, les autres, un effet répulsif sur le neurone ; demande en entrée des paramètres tels que le nombre d'itérations durant l'apprentissage, les nombres de plus proches voisins pour l'initialisation de classe de chaque neurone, et le nombre de neurones du réseau ; les résultats changent d'une exécution à l'autre en raison d'initialisations aléatoires ;
- OC1 : algorithme supervisé consistant en un arbre de décision oblique ; demande en entrée des paramètres tels que le nombre de sauts aléatoires pour perturber un plan, le nombre de plans à perturber ; les résultats changent d'une exécution à l'autre ;
- SVM : algorithme supervisé basé sur la maximisation de la distance entre deux plans séparateurs parallèles ; demande en entrée le coefficient λ qui assure un compromis entre taille de la marge et tolérance aux erreurs, dans certain cas il faut aussi la distribution a priori des objets de chaque type dans l'échantillon total.

Les trois premiers algorithmes – k-NN, Meanshift et KDC – ont été testé à partir de notre propre implémentation en JAVA. De nombreuses requêtes de type k-NN sont nécessaires dans ces différents algorithmes et ont été réalisées à partir de *kd-trees* pour en optimiser les performances.

Pour les trois autres algorithmes, nous avons eu recours à des codes C disponibles gratuitement sur Internet. En ce qui concerne les réseaux de neurones, nous avons utilisé le package LVQ_PAK[4] (Kohonen *et al.* 1992, 1996) qui est libre pour une utilisation scientifique. Nous avons uniquement

[4] `http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/neural/systems/som_lvq/lvq_pak/`

testé l'algorithme OLVQ1 avec différentes valeurs du nombre de neurones (option -noc). Concernant l'algorithme OC1, nous avons utilisé la seule implémentation qui existe à notre connaissance[5] et qui a été réalisée par les auteurs de l'article White (2008). Nous avons utilisé les options permettant de changer le nombre de plans aléatoires de départ et le nombre de sauts aléatoires servant à sortir des minimums locaux. Enfin, pour les SVM, nous utilisons le package SVM_Light[6] crée par l'auteur de l'article Joachims (1999). Nous avons utilisé les options permettant de choisir différentes fonctions noyaux et la valeur du paramètre qui détermine le compromis entre largeur de la marge et tolérance aux erreurs.

Dans l'algorithme des *meanshifts*, il se peut qu'aucune source de l'échantillon d'apprentissage ne tombe sur un maximum local donné. Cela peut notamment être le cas pour un petit groupe d'objets rares ou pour des *outliers* n'ayant aucun voisin à une distance inférieure à la valeur de la largeur de bande de la fonction noyau. Dans ce cas, la classe attribuée au maximum local est la classe du maximum local le plus proche auquel on a pu attribuer une classe.
Dans le cas de la KDC, si la densité locale mesurée en un point pour chaque type d'objet est nulle, on remplace la largeur de bande fixe par la distance au k^e plus proche voisin. On a pris ici $k = 7$.

4.4 RÉSULTATS

4.4.1 Analyse globale des résultats sur le premier jeu de données

Les résultats de la classification des jeux de données basés sur XS par les différents algorithmes avec diverses options sont présentés dans les tableaux IV.18[7], IV.19, IV.20, IV.21 et IV.22, IV.23, IV.24, IV.25. On remarque tout d'abord qu'on obtient d'excellents résultats pour l'ensemble des algorithmes, avec des valeurs pour les différents paramètres mesurés toujours au dessus de 90%, en particulier l'*accuracy* qui est presque constamment au dessus de 99%. On obtient logiquement de meilleurs résultats avec les jeux de données pour lesquels les CV ont été supprimées de l'échantillon d'apprentissage. C'est une des limitations des méthodes de classsification supervisée : on ne trouve que ce que l'on cherche. Les objets dont le type est absent de l'échantillon d'apprentissage seront mal classés alors que la matrice de confusion donnera l'impression d'une classification fiable.
Les résultats sont également globalement meilleurs quand le nombre de dimensions des paramètres d'entrée augmente.

Alors qu'elle est choisie par peur de surestimer le taux de bonne classification, la méthode *train-test* apporte ici des résultats légèrement meilleurs que la prise en compte de l'échantillon total à la fois pour la classification et pour le calcul de la matrice de confusion. Cela est peut-être dû à la taille relativement faible du nombre d'étoiles servant à déterminer la matrice de confusion avec la méthode *train-test*. En effet cet échantillon ne contient qu'un tiers des étoiles de départ, soit environ 180 sources. Une source représente alors un pourcentage de 0.56% dans le calcul de l'*accuracy* sur les étoiles. Il suffit alors que les sources proches de la limite de confusion entre QSO et étoiles soient dans les 2/3 des sources servant à l'apprentissage pour surestimer les différents paramètres. Ce résultat permet de montrer que pour des échantillons d'apprentissage relativement petits, la méthode *train-test* n'est pas forcément la plus adaptée.

4.4.2 Analyse algorithme par algorithme

Le KNN est le plus simple des algorithmes. Programmé et optimisé en C il serait sûrement plus rapide que tous les autres, LVQ compris. Ses performances en termes de classification sont ici globalement similaires aux autres algorithmes. Le k-NN produit pour chaque source des probabilités discrètes d'appartenance à chaque classe. De fait, si l'échantillon d'apprentissage contient plus de deux classes ou si le nombre k choisi est pair, une source ne peut pas toujours être classé sans ambiguité.

[5] `http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/learning/systems/oc1/`

[6] `http://svmlight.joachims.org/`

[7] dans les tableaux, « fb » signifie *fixed bandwidth* et « bal » signifie *balloon estimator*

Le *meanshift* a l'avantage d'être non supervisé. Pour l'échantillon XS_3D_1, le *meanshift* trouve 3 maximums avec une largeur de bande fixe de 1.0, 6 avec 0.75, 19 avec 0.50, etc. Deux maximums locaux représentent repectivement 99.99%, 99.96% et 88.36% des sources. Les autres points du *meanshift* sont pour la plupart au niveau d'*outliers* où la montée de gradient ne peut pas se faire par manque de sources dans la boîte de recherche sphérique. Plus rarement, ils correspondent à des maxima locaux dont le nombre augmente au fur et à mesure que le lissage diminue. Les résultats obtenus sont du même ordre qu'avec les autres algorithmes. Le temps d'éxecution, bien que nettement supérieur à celui des autres méthodes, excepté les SVM, reste acceptable. On retiendra surtout que cette méthode peut s'avérer très utile quand l'échantillon d'apprentissage est mal défini voir absent, ou souffre de biais importants. Le *meanshift* ne fournit pas de façon naturelle de probabilités en sortie.
On remarque que l'utilisation du *knn* plutôt que de la largeur de bande fixe est intéressante pour les régions de faible densité mais provoque de nombreux maxima locaux dans les zones de forte densité. Nous avons implémenté un algorithme mixte où l'on utilise le knn si le nombre de sources dans la largeur de bande fixe n'est pas suffisant.

Là encore, les résultats sont globalement identiques aux autres algorithmes. De plus, le temps d'éxécution est court : trois secondes en moyenne. L'immense avantage de cette méthode est de fournir, pour chaque source, des probabilités d'appartenance à chaque classe. On pourra donc constituer des échantillons plus ou moins purs en fixant des limites sur ces probabilités de sortie.

Concernant l'algorithme LVQ, on voit que ses résultats s'améliorent en général avec le nombre de neurones en entrée et qu'ils sont identiques, à une exception près (XS_9D_1), pour 500 et 1 000 neurones. C'est l'algorithme le plus rapide avec un temps d'éxécution presque toujours inférieur à la seconde. Globalement ses performances sont identiques au KNN.
On peut reprocher à l'algorithme LVQ de ne pas être déterministe – deux exécutions ne donneront pas nécessairement les mêmes résultats en sortie – et de reposer sur un fondement plus empirique que théorique ; pour les objets classés, il n'inclut pas de façon naturelle leurs probabilités d'appartenance à chaque classe. L'implémentation que nous avons testée est optimisée et le temps d'exécution est court.

Nous avons testé l'algorithme OC1 avec les valeurs par défaut et en doublant le nombre de plans de séparation initiaux et le nombre de sauts aléatoires. Les résultats sont souvent légèrement meilleurs avec la deuxième solution, mais pas systématiquement. En revanche, le temps d'exécution est parfois beaucoup plus long et peut durer plusieurs heures alors qu'avec les paramètres par défaut il n'est que de quelques secondes. Pire, pour une raison qui nous est inconnue, l'algorithme, avec ces paramètres doublés semble parfois se bloquer : on le laisse fonctionner plusieurs heures avant de l'arrêter en tuant le processus. C'est pour cette raison que ne figure parfois que la version avec les paramètres par défaut dans nos tests. Les résultats sont globalement identiques, voir peut-être légèrement meilleurs, que ceux obtenus avec l'algorithme LVQ. Mais l'algorithme OC1 souffre des mêmes reproches : il n'est pas déterministe (il est semi-déterministe), avec des résultats qui changent d'une exécution à l'autre ; il ne fournit pas naturellement de probabilités. Son temps d'exécution varie mais il est globalement plus lent que tous les autres algorithmes de classification supervisée, excepté les SVM.

Si on relâche la contrainte (ξ) sur la taille de la marge au profit d'une minimisation des erreurs, l'algorithme de SVM utilisé peut devenir très lent et son temps d'exécution peut dépasser la journée. Dans ce cas, nous enlevons le test de nos tableaux comparatifs.
Les résultats que nous obtenons avec les SVM sont au mieux comparables aux autres résultats. On voit cependant une forte dépendance entre résultat et paramètre ξ. Les temps d'exécution sont très variables, allant de 2 minutes à presque 14h pour un échantillon modeste de moins de 10 000 sources.
Contrairement à Gao *et al.* (2008), qui a comparé k-NN et SVM, nous obtenons de meilleurs résultats avec le k-NN.

4.5 Conclusions

Pourvu qu'on ajuste les paramètres d'entrée, tous les algorithmes de classification testés ici ont des performances globalement identiques sur nos échantillons. Le temps d'exécution de tous

TAB. IV.18 – Résultats des algorithmes de classification sur l'échantillon XS_3D_1.

Algo.	Options	Acc.	Acc.S	Acc.Q	Prec.	FM	GM	WA	Temps (s)
KNN	k=3	98.83	95.70	99.94	99.81	97.71	97.80	97.82	1
KNN	k=5	98.65	95.34	99.81	99.44	97.35	97.55	97.58	1
KNN	k=7	98.51	94.44	99.94	99.81	97.05	97.15	97.19	1
KNN	k=11	98.51	94.44	99.94	99.81	97.05	97.15	97.19	2
MS	knn 25	98.09	93.01	99.87	99.62	96.20	96.38	96.44	4
MS	knn 50	98.09	93.73	99.62	98.87	96.23	96.63	96.67	11
MS	knn 75	98.04	93.01	99.81	99.43	96.11	96.35	96.41	20
MS	knn 100	97.76	92.11	99.75	99.23	95.54	95.86	95.93	29
MS	knn 150	97.81	92.29	99.75	99.23	95.64	95.95	96.02	56
MS	fb 0.15	99.21	97.31	99.87	99.63	98.46	98.58	98.59	62
MS	fb 0.25	98.65	94.80	100.00	100.00	97.33	97.37	97.40	396
MS	fb 0.50	98.32	94.09	99.81	99.43	96.69	96.91	96.95	297
MS	fb 0.75	97.90	92.29	99.87	99.61	95.81	96.01	96.08	223
MS	fb 1.00	97.86	92.47	99.75	99.23	95.73	96.04	96.11	204
KDC	knn 25	98.00	93.19	99.68	99.05	96.03	96.38	96.44	2
KDC	knn 50	97.86	92.65	99.68	99.04	95.74	96.10	96.17	2
KDC	knn 75	97.72	92.29	99.62	98.85	95.46	95.89	95.96	3
KDC	knn 100	97.76	92.47	99.62	98.85	95.56	95.98	96.05	3
KDC	knn 150	97.81	92.47	99.68	99.04	95.64	96.01	96.08	3
KDC	fb 0.15	99.53	98.21	100.00	100.00	99.10	99.10	99.10	2
KDC	fb 0.25	99.25	97.31	99.94	99.82	98.55	98.62	98.62	2
KDC	fb 0.50	98.88	95.70	100.00	100.00	97.80	97.83	97.85	3
KDC	fb 0.75	98.69	94.98	100.00	100.00	97.43	97.46	97.49	5
KDC	fb 1.00	98.23	93.73	99.81	99.43	96.49	96.72	96.77	6
KDC	bal 25	98.55	94.98	99.81	99.44	97.16	97.37	97.40	2
KDC	bal 50	98.51	94.62	99.87	99.62	97.06	97.21	97.25	2
KDC	bal 75	98.28	94.09	99.75	99.24	96.60	96.88	96.92	2
KDC	bal 100	98.18	93.73	99.75	99.24	96.41	96.69	96.74	3
KDC	bal 150	98.04	93.37	99.68	99.05	96.13	96.48	96.53	3
LVQ	nn 50	98.04	93.73	99.56	98.68	96.14	96.60	96.64	1
LVQ	nn 200	98.46	94.44	99.87	99.62	96.96	97.12	97.16	<1
LVQ	nn 500	98.55	94.62	99.94	99.81	97.15	97.24	97.28	<1
LVQ	nn 1000	98.55	94.62	99.94	99.81	97.15	97.24	97.28	1
OC1	0.1 20 5	98.83	95.88	99.87	99.63	97.72	97.86	97.88	31
OC1	0.1 40 10	98.28	93.55	99.94	99.81	96.58	96.69	96.74	3994
OC1	0.3 20 5	99.02	96.59	99.87	99.63	98.09	98.22	98.23	36
SVM	0.1 0	88.58	95.70	86.07	70.73	81.34	90.76	90.89	1035
SVM	1.0 0	89.46	95.52	87.33	72.62	82.51	91.34	91.43	109
SVM	10.0 0	89.14	96.06	86.70	71.75	82.15	91.26	91.38	482
SVM	1000.0 0								>4600
SVM	0.1 1 2	89.46	96.24	87.08	72.37	82.62	91.55	91.66	2385
SVM	1.0 1 2	89.09	94.62	87.15	72.13	81.86	90.81	90.88	49988
SVM	10.0 1 2	98.32	94.27	99.75	99.25	96.69	96.97	97.01	452
SVM	1000.0 1 2	26.71	89.61	4.60	24.83	38.88	20.30	47.10	66
SVM	0.1 2 5	91.56	98.21	89.22	76.22	85.83	93.61	93.72	894
SVM	1.0 2 5								>4600
SVM	10.0 2 5								>4600

TAB. IV.19 – Résultats des algorithmes de classification sur l'échantillon XS_3D_2.

Algo.	Options	Acc.	Acc.S	Acc.Q	Prec.	FM	GM	WA	Temps (s)
KNN	k=3	99.01	96.48	99.90	99.72	98.07	98.18	98.19	1
KNN	k=5	98.73	95.39	99.90	99.72	97.51	97.62	97.65	1
KNN	k=7	98.66	95.12	99.90	99.72	97.36	97.48	97.51	1
KNN	k=11	98.66	95.12	99.90	99.72	97.36	97.48	97.51	1
MS	knn 25	98.31	94.31	99.71	99.15	96.67	96.97	97.01	5
MS	knn 50	98.31	94.04	99.81	99.43	96.66	96.88	96.92	12
MS	knn 75	98.24	93.77	99.81	99.43	96.51	96.74	96.79	20
MS	knn 100	97.88	92.68	99.71	99.13	95.80	96.13	96.20	28
MS	knn 150	97.95	92.95	99.71	99.13	95.94	96.27	96.33	57
MS	fb 0.15	99.29	97.56	99.90	99.72	98.63	98.73	98.73	62
MS	fb 0.25	98.87	95.66	100.00	100.00	97.78	97.81	97.83	399
MS	fb 0.50	98.59	95.12	99.81	99.43	97.23	97.44	97.47	299
MS	fb 0.75	98.03	92.68	99.90	99.71	96.07	96.23	96.29	225
MS	fb 1.00	97.95	92.95	99.71	99.13	95.94	96.27	96.33	193
KDC	knn 25	98.10	94.04	99.52	98.58	96.26	96.74	96.78	2
KDC	knn 50	98.03	93.50	99.62	98.85	96.10	96.51	96.56	2
KDC	knn 75	97.95	93.22	99.62	98.85	95.96	96.37	96.42	2
KDC	knn 100	97.95	93.22	99.62	98.85	95.96	96.37	96.42	3
KDC	knn 150	97.95	93.22	99.62	98.85	95.96	96.37	96.42	4
KDC	fb 0.15	99.72	98.92	100.00	100.00	99.46	99.46	99.46	1
KDC	fb 0.25	99.37	97.56	100.00	100.00	98.77	98.77	98.78	1
KDC	fb 0.50	99.15	96.75	100.00	100.00	98.35	98.36	98.37	2
KDC	fb 0.75	99.01	96.21	100.00	100.00	98.07	98.08	98.10	4
KDC	fb 1.00	98.66	95.12	99.90	99.72	97.36	97.48	97.51	4
KDC	bal 25	98.80	95.66	99.90	99.72	97.65	97.76	97.78	1
KDC	bal 50	98.66	95.12	99.90	99.72	97.36	97.48	97.51	2
KDC	bal 75	98.45	94.58	99.81	99.43	96.94	97.16	97.19	2
KDC	bal 100	98.38	94.58	99.71	99.15	96.81	97.11	97.15	2
KDC	bal 150	98.24	94.04	99.71	99.14	96.52	96.83	96.88	3
LVQ	nn 50	98.52	94.85	99.81	99.43	97.09	97.30	97.33	<1
LVQ	nn 200	98.73	95.39	99.90	99.72	97.51	97.62	97.65	<1
LVQ	nn 500	98.80	95.39	100.00	100.00	97.64	97.67	97.70	<1
LVQ	nn 1000	98.80	95.39	100.00	100.00	97.64	97.67	97.70	1
OC1	0.1 20 5	98.59	95.39	99.71	99.15	97.24	97.53	97.55	108
OC1	0.1 40 10	98.66	95.66	99.71	99.16	97.38	97.67	97.69	102
OC1	0.3 20 5	99.29	97.56	99.90	99.72	98.63	98.73	98.73	5
SVM	0.1 0	87.66	96.21	84.65	68.80	80.23	90.24	90.43	1022
SVM	1.0 0	90.06	95.39	88.18	73.95	83.31	91.72	91.79	247
SVM	10.0 0	89.99	96.75	87.61	73.31	83.41	92.06	92.18	349
SVM	0.1 1 2	89.70	96.75	87.23	72.71	83.02	91.86	91.99	934
SVM	1.0 1 2								>24301
SVM	1000.0 1 2	41.47	62.60	34.03	25.03	35.76	46.16	48.32	66
SVM	0.1 2 5	91.33	98.37	88.85	75.62	85.51	93.49	93.61	102
SVM	1.0 2 5								>13141

TAB. IV.20 – Résultats des algorithmes de classification sur l'échantillon XS_3D_3.

Algo.	Options	Acc.	Acc.S	Acc.Q	Prec.	FM	GM	WA	Temps (s)
KNN	k=3	99.81	99.25	100.00	100.00	99.62	99.62	99.62	1
KNN	k=5	99.62	98.87	99.87	99.62	99.25	99.37	99.37	1
KNN	k=7	99.72	98.87	100.00	100.00	99.43	99.43	99.44	1
KNN	k=11	99.67	98.87	99.94	99.81	99.34	99.40	99.40	1
MS	knn 25	99.24	97.37	99.87	99.62	98.48	98.61	98.62	5
MS	knn 50	99.29	98.12	99.68	99.05	98.58	98.90	98.90	11
MS	knn 75	99.24	97.56	99.81	99.43	98.48	98.68	98.68	20
MS	knn 100	98.96	96.62	99.75	99.23	97.90	98.17	98.18	28
MS	knn 150	99.01	96.80	99.75	99.23	98.00	98.27	98.28	58
MS	fb 0.15	99.95	99.81	100.00	100.00	99.91	99.91	99.91	62
MS	fb 0.25	99.58	98.31	100.00	100.00	99.15	99.15	99.15	402
MS	fb 0.50	99.43	98.12	99.87	99.62	98.86	98.99	99.00	304
MS	fb 0.75	99.06	96.62	99.87	99.61	98.09	98.23	98.25	225
MS	fb 1.00	99.01	96.80	99.75	99.23	98.00	98.27	98.28	206
KDC	knn 25	99.15	97.56	99.68	99.05	98.30	98.61	98.62	2
KDC	knn 50	99.01	96.99	99.68	99.04	98.01	98.33	98.34	2
KDC	knn 75	98.87	96.62	99.62	98.85	97.72	98.11	98.12	3
KDC	knn 100	98.91	96.80	99.62	98.85	97.82	98.20	98.21	3
KDC	knn 150	98.96	96.80	99.68	99.04	97.91	98.23	98.24	4
KDC	fb 0.15	99.95	99.81	100.00	100.00	99.91	99.91	99.91	1
KDC	fb 0.25	99.95	99.81	100.00	100.00	99.91	99.91	99.91	1
KDC	fb 0.50	99.86	99.44	100.00	100.00	99.72	99.72	99.72	3
KDC	fb 0.75	99.76	99.06	100.00	100.00	99.53	99.53	99.53	5
KDC	fb 1.00	99.29	97.56	99.87	99.62	98.58	98.71	98.72	6
KDC	bal 25	99.72	99.06	99.94	99.81	99.43	99.50	99.50	2
KDC	bal 50	99.62	98.87	99.87	99.62	99.25	99.37	99.37	2
KDC	bal 75	99.43	98.12	99.87	99.62	98.86	98.99	99.00	3
KDC	bal 100	99.24	97.74	99.75	99.24	98.48	98.74	98.75	3
KDC	bal 150	99.20	97.56	99.75	99.24	98.39	98.65	98.65	3
LVQ	nn 50	99.20	97.93	99.62	98.86	98.39	98.77	98.78	<1
LVQ	nn 200	99.67	98.87	99.94	99.81	99.34	99.40	99.40	<1
LVQ	nn 500	99.72	98.87	100.00	100.00	99.43	99.43	99.44	<1
LVQ	nn 1000	99.72	98.87	100.00	100.00	99.43	99.43	99.44	1
OC1	0.1 20 5	100.00	100.00	100.00	100.00	100.00	100.00	100.00	9
OC1	0.3 20 5	99.58	98.68	99.87	99.62	99.15	99.28	99.28	7
SVM	0.1 0	90.09	99.44	86.96	71.88	83.44	92.99	93.20	730
SVM	1.0 0	90.04	99.06	87.02	71.90	83.32	92.84	93.04	705
SVM	10.0 0	90.42	99.06	87.52	72.69	83.85	93.11	93.29	454
SVM	1000.0 0								>7807
SVM	0.1 1 2	90.28	99.44	87.21	72.27	83.70	93.12	93.32	383
SVM	1.0 1 2								>65850
SVM	10.0 1 2								>38628
SVM	1000.0 1 2	99.39	98.68	99.62	98.87	98.78	99.15	99.15	187
SVM	0.1 2 5	92.21	100.00	89.60	76.33	86.57	94.66	94.80	794
SVM	1.0 2 5								>26501

TAB. IV.21 – Résultats des algorithmes de classification sur l'échantillon XS_3D_4.

Algo.	Options	Acc.	Acc.S	Acc.Q	Prec.	FM	GM	WA	Temps (s)
KNN	k=3	99.86	99.43	100.00	100.00	99.72	99.72	99.72	1
KNN	k=5	99.57	98.87	99.81	99.43	99.15	99.34	99.34	1
KNN	k=7	99.57	98.87	99.81	99.43	99.15	99.34	99.34	1
KNN	k=11	99.50	98.58	99.81	99.43	99.00	99.19	99.20	2
MS	knn 25	99.29	99.15	99.33	98.04	98.59	99.24	99.24	5
MS	knn 50	99.36	98.30	99.71	99.14	98.72	99.00	99.01	11
MS	knn 75	99.29	97.45	99.90	99.71	98.57	98.67	98.68	20
MS	knn 100	99.00	96.32	99.90	99.71	97.98	98.09	98.11	28
MS	knn 150	99.07	96.60	99.90	99.71	98.13	98.24	98.25	58
MS	fb 0.15	100.00	100.00	100.00	100.00	100.00	100.00	100.00	62
MS	fb 0.25	99.57	98.30	100.00	100.00	99.14	99.15	99.15	401
MS	fb 0.50	99.43	98.02	99.90	99.71	98.86	98.96	98.96	297
MS	fb 0.75	99.00	96.32	99.90	99.71	97.98	98.09	98.11	226
MS	fb 1.00	99.00	96.60	99.81	99.42	97.99	98.19	98.20	196
KDC	knn 25	99.21	97.45	99.81	99.42	98.43	98.62	98.63	2
KDC	knn 50	99.14	96.88	99.90	99.71	98.28	98.38	98.39	2
KDC	knn 75	99.00	96.32	99.90	99.71	97.98	98.09	98.11	2
KDC	knn 100	99.07	96.60	99.90	99.71	98.13	98.24	98.25	2
KDC	knn 150	99.07	96.60	99.90	99.71	98.13	98.24	98.25	3
KDC	fb 0.15	100.00	100.00	100.00	100.00	100.00	100.00	100.00	1
KDC	fb 0.25	100.00	100.00	100.00	100.00	100.00	100.00	100.00	1
KDC	fb 0.50	99.93	99.72	100.00	100.00	99.86	99.86	99.86	3
KDC	fb 0.75	99.71	98.87	100.00	100.00	99.43	99.43	99.43	3
KDC	fb 1.00	99.29	97.73	99.81	99.42	98.57	98.77	98.77	4
KDC	bal 25	99.57	98.87	99.81	99.43	99.15	99.34	99.34	2
KDC	bal 50	99.43	98.30	99.81	99.43	98.86	99.05	99.05	2
KDC	bal 75	99.29	97.73	99.81	99.42	98.57	98.77	98.77	3
KDC	bal 100	99.29	97.73	99.81	99.42	98.57	98.77	98.77	3
KDC	bal 150	99.21	97.17	99.90	99.71	98.42	98.53	98.54	4
LVQ	nn 50	99.14	97.17	99.81	99.42	98.28	98.48	98.49	<1
LVQ	nn 200	99.71	99.15	99.90	99.72	99.43	99.53	99.53	<1
LVQ	nn 500	99.71	99.15	99.90	99.72	99.43	99.53	99.53	<1
LVQ	nn 1000	99.71	99.15	99.90	99.72	99.43	99.53	99.53	<1
OC1	0.1 20 5	99.50	98.87	99.71	99.15	99.01	99.29	99.29	8
OC1	0.3 20 5	99.64	98.58	100.00	100.00	99.29	99.29	99.29	7
SVM	0.1 0	89.36	99.43	85.97	70.48	82.49	92.46	92.70	4868
SVM	1.0 0	90.29	99.43	87.21	72.37	83.77	93.12	93.32	1226
SVM	10.0 0	89.15	98.58	85.97	70.30	82.08	92.06	92.28	480
SVM	1000.0 0								>12826
SVM	0.1 1 2	90.36	99.43	87.31	72.52	83.87	93.17	93.37	240
SVM	1.0 1 2	91.15	99.15	88.45	74.31	84.95	93.65	93.80	10049
SVM	10.0 1 2								>63876
SVM	1000.0 1 2	97.29	89.24	100.00	100.00	94.31	94.46	94.62	69
SVM	0.1 2 5	91.72	100.00	88.93	75.27	85.89	94.30	94.47	79
SVM	1.0 2 5								>10412

TAB. IV.22 – Résultats des algorithmes de classification sur l'échantillon XS_9D_1.

Algo.	Options	Acc.	Acc.S	Acc.Q	Prec.	FM	GM	WA	Temps (s)
KNN	k=3	99.35	97.67	99.94	99.82	98.73	98.80	98.80	2
KNN	k=5	98.88	95.88	99.94	99.81	97.81	97.89	97.91	2
KNN	k=7	98.74	95.34	99.94	99.81	97.53	97.61	97.64	3
KNN	k=11	98.60	94.80	99.94	99.81	97.24	97.34	97.37	3
MS	knn 25	98.00	92.65	99.87	99.61	96.01	96.20	96.26	93
MS	knn 50	97.76	91.40	100.00	100.00	95.51	95.60	95.70	177
MS	knn 75	97.76	91.58	99.94	99.80	95.51	95.67	95.76	222
MS	knn 100	97.76	91.58	99.94	99.80	95.51	95.67	95.76	267
MS	knn 150	97.81	91.58	100.00	100.00	95.60	95.70	95.79	307
MS	fb 0.15	100.00	100.00	100.00	100.00	100.00	100.00	100.00	20
MS	fb 0.25	100.00	100.00	100.00	100.00	100.00	100.00	100.00	28
MS	fb 0.50	99.25	97.13	100.00	100.00	98.55	98.56	98.57	388
MS	fb 0.75	98.32	93.91	99.87	99.62	96.68	96.84	96.89	401
MS	fb 1.00	98.23	93.37	99.94	99.81	96.48	96.60	96.65	378
KDC	knn 25	97.95	92.47	99.87	99.61	95.91	96.10	96.17	4
KDC	knn 50	97.81	92.11	99.81	99.42	95.63	95.89	95.96	5
KDC	knn 75	97.81	92.11	99.81	99.42	95.63	95.89	95.96	6
KDC	knn 100	97.72	91.76	99.81	99.42	95.43	95.70	95.78	7
KDC	knn 150	97.67	91.58	99.81	99.42	95.34	95.61	95.69	7
KDC	fb 0.15	100.00	100.00	100.00	100.00	100.00	100.00	100.00	1
KDC	fb 0.25	100.00	100.00	100.00	100.00	100.00	100.00	100.00	2
KDC	fb 0.50	99.81	99.28	100.00	100.00	99.64	99.64	99.64	3
KDC	fb 0.75	99.21	96.95	100.00	100.00	98.45	98.46	98.48	4
KDC	fb 1.00	98.88	95.70	100.00	100.00	97.80	97.83	97.85	5
KDC	bal 25	98.09	98.03	98.11	94.80	96.39	98.07	98.07	4
KDC	bal 50	98.41	95.88	99.31	97.99	96.92	97.58	97.59	4
KDC	bal 75	98.41	95.88	99.31	97.99	96.92	97.58	97.59	5
KDC	bal 100	98.46	95.70	99.43	98.34	97.00	97.55	97.57	5
KDC	bal 150	98.18	94.80	99.37	98.14	96.44	97.06	97.09	6
LVQ	nn 50	98.28	94.27	99.68	99.06	96.60	96.94	96.98	<1
LVQ	nn 200	98.60	94.62	100.00	100.00	97.24	97.27	97.31	<1
LVQ	nn 500	98.88	95.70	100.00	100.00	97.80	97.83	97.85	<1
LVQ	nn 1000	99.02	96.24	100.00	100.00	98.08	98.10	98.12	<1
OC1	0.1 20 5	98.41	94.27	99.87	99.62	96.87	97.03	97.07	32
OC1	0.3 20 5	98.69	95.88	99.68	99.07	97.45	97.76	97.78	21

TAB. IV.23 – Résultats des algorithmes de classification sur l'échantillon XS_9D_2.

Algo.	Options	Acc.	Acc.S	Acc.Q	Prec.	FM	GM	WA	Temps (s)
KNN	k=3	99.08	96.48	100.00	100.00	98.21	98.22	98.24	2
KNN	k=5	98.94	95.93	100.00	100.00	97.93	97.95	97.97	2
KNN	k=7	98.66	95.12	99.90	99.72	97.36	97.48	97.51	3
KNN	k=11	98.52	94.58	99.90	99.71	97.08	97.21	97.24	3
MS	knn 25	98.31	93.77	99.90	99.71	96.65	96.79	96.84	94
MS	knn 50	97.95	92.14	100.00	100.00	95.91	95.99	96.07	179
MS	knn 75	98.03	92.68	99.90	99.71	96.07	96.23	96.29	224
MS	knn 100	98.03	92.68	99.90	99.71	96.07	96.23	96.29	267
MS	knn 150	98.10	92.68	100.00	100.00	96.20	96.27	96.34	313
MS	fb 0.15	100.00	100.00	100.00	100.00	100.00	100.00	100.00	22
MS	fb 0.25	100.00	100.00	100.00	100.00	100.00	100.00	100.00	27
MS	fb 0.50	99.44	97.83	100.00	100.00	98.90	98.91	98.92	289
MS	fb 0.75	98.73	95.12	100.00	100.00	97.50	97.53	97.56	410
MS	fb 1.00	98.52	94.58	99.90	99.71	97.08	97.21	97.24	389
KDC	knn 25	97.95	92.95	99.71	99.13	95.94	96.27	96.33	4
KDC	knn 50	97.95	92.95	99.71	99.13	95.94	96.27	96.33	4
KDC	knn 75	97.81	92.41	99.71	99.13	95.65	95.99	96.06	4
KDC	knn 100	97.81	92.41	99.71	99.13	95.65	95.99	96.06	5
KDC	knn 150	97.74	92.14	99.71	99.13	95.51	95.85	95.93	5
KDC	fb 0.15	100.00	100.00	100.00	100.00	100.00	100.00	100.00	1
KDC	fb 0.25	100.00	100.00	100.00	100.00	100.00	100.00	100.00	2
KDC	fb 0.50	99.79	99.19	100.00	100.00	99.59	99.59	99.59	3
KDC	fb 0.75	99.58	98.37	100.00	100.00	99.18	99.18	99.19	3
KDC	fb 1.00	99.15	96.75	100.00	100.00	98.35	98.36	98.37	4
KDC	bal 25	98.94	97.02	99.62	98.90	97.95	98.31	98.32	3
KDC	bal 50	98.73	96.21	99.62	98.89	97.53	97.90	97.91	4
KDC	bal 75	98.73	95.93	99.71	99.16	97.52	97.81	97.82	5
KDC	bal 100	98.59	95.66	99.62	98.88	97.25	97.62	97.64	5
KDC	bal 150	98.45	95.12	99.62	98.87	96.96	97.34	97.37	5
LVQ	nn 50	98.52	94.58	99.90	99.71	97.08	97.21	97.24	<1
LVQ	nn 200	98.87	95.66	100.00	100.00	97.78	97.81	97.83	<1
LVQ	nn 500	98.80	95.39	100.00	100.00	97.64	97.67	97.70	<1
LVQ	nn 1000	98.80	95.39	100.00	100.00	97.64	97.67	97.70	1
OC1	0.1 20 5	99.79	99.19	100.00	100.00	99.59	99.59	99.59	28
OC1	0.3 20 5	99.01	96.48	99.90	99.72	98.07	98.18	98.19	8

TAB. IV.24 – Résultats des algorithmes de classification sur l'échantillon XS_9D_3.

Algo.	Options	Acc.	Acc.S	Acc.Q	Prec.	FM	GM	WA	Temps (s)
KNN	k=3	99.72	99.06	99.94	99.81	99.43	99.50	99.50	2
KNN	k=5	99.72	99.06	99.94	99.81	99.43	99.50	99.50	3
KNN	k=7	99.72	99.06	99.94	99.81	99.43	99.50	99.50	3
KNN	k=11	99.72	99.06	99.94	99.81	99.43	99.50	99.50	3
MS	knn 25	99.24	97.37	99.87	99.62	98.48	98.61	98.62	93
MS	knn 50	98.96	95.86	100.00	100.00	97.89	97.91	97.93	175
MS	knn 75	98.96	96.05	99.94	99.80	97.89	97.98	97.99	224
MS	knn 100	98.96	96.05	99.94	99.80	97.89	97.98	97.99	268
MS	knn 150	99.01	96.05	100.00	100.00	97.99	98.01	98.03	314
MS	fb 0.15	100.00	100.00	100.00	100.00	100.00	100.00	100.00	22
MS	fb 0.25	99.95	100.00	99.94	99.81	99.91	99.97	99.97	28
MS	fb 0.50	99.91	99.62	100.00	100.00	99.81	99.81	99.81	390
MS	fb 0.75	99.34	97.74	99.87	99.62	98.67	98.80	98.81	404
MS	fb 1.00	99.24	97.18	99.94	99.81	98.48	98.55	98.56	453
KDC	knn 25	99.10	96.80	99.87	99.61	98.19	98.33	98.34	5
KDC	knn 50	98.96	96.43	99.81	99.42	97.90	98.11	98.12	5
KDC	knn 75	98.96	96.43	99.81	99.42	97.90	98.11	98.12	5
KDC	knn 100	98.87	96.05	99.81	99.42	97.71	97.91	97.93	5
KDC	knn 150	98.82	95.86	99.81	99.42	97.61	97.82	97.84	6
KDC	fb 0.15	100.00	100.00	100.00	100.00	100.00	100.00	100.00	1
KDC	fb 0.25	100.00	100.00	100.00	100.00	100.00	100.00	100.00	2
KDC	fb 0.50	99.95	99.81	100.00	100.00	99.91	99.91	99.91	3
KDC	fb 0.75	99.91	99.62	100.00	100.00	99.81	99.81	99.81	4
KDC	fb 1.00	99.91	99.62	100.00	100.00	99.81	99.81	99.81	5
KDC	bal 25	99.72	99.44	99.81	99.44	99.44	99.62	99.62	4
KDC	bal 50	99.48	99.06	99.62	98.87	98.97	99.34	99.34	5
KDC	bal 75	99.53	99.44	99.56	98.69	99.06	99.50	99.50	5
KDC	bal 100	99.58	99.44	99.62	98.88	99.16	99.53	99.53	6
KDC	bal 150	99.39	98.50	99.68	99.05	98.77	99.09	99.09	6
LVQ	nn 50	99.48	98.87	99.68	99.06	98.97	99.28	99.28	>1
LVQ	nn 200	99.76	99.06	100.00	100.00	99.53	99.53	99.53	1
LVQ	nn 500	99.81	99.25	100.00	100.00	99.62	99.62	99.62	1
LVQ	nn 1000	99.81	99.25	100.00	100.00	99.62	99.62	99.62	1
OC1	0.1 20 5	98.96	95.86	100.00	100.00	97.89	97.91	97.93	27
OC1	0.3 20 5	99.29	98.12	99.68	99.05	98.58	98.90	98.90	18

TAB. IV.25 – Résultats des algorithmes de classification sur l'échantillon XS_9D_4.

Algo.	Options	Acc.	Acc.S	Acc.Q	Prec.	FM	GM	WA	Temps (s)
KNN	k=3	99.71	99.15	99.90	99.72	99.43	99.53	99.53	2
KNN	k=5	99.64	99.15	99.81	99.43	99.29	99.48	99.48	2
KNN	k=7	99.64	99.15	99.81	99.43	99.29	99.48	99.48	3
KNN	k=11	99.57	98.58	99.90	99.71	99.15	99.24	99.24	3
MS	knn 25	99.21	97.17	99.90	99.71	98.42	98.53	98.54	93
MS	knn 50	98.86	95.47	100.00	100.00	97.68	97.71	97.73	181
MS	knn 75	98.79	95.47	99.90	99.70	97.54	97.66	97.69	229
MS	knn 100	98.79	95.47	99.90	99.70	97.54	97.66	97.69	267
MS	knn 150	98.86	95.47	100.00	100.00	97.68	97.71	97.73	310
MS	fb 0.15	100.00	100.00	100.00	100.00	100.00	100.00	100.00	21
MS	fb 0.25	100.00	100.00	100.00	100.00	100.00	100.00	100.00	27
MS	fb 0.50	99.93	99.72	100.00	100.00	99.86	99.86	99.86	391
MS	fb 0.75	99.21	97.17	99.90	99.71	98.42	98.53	98.54	410
MS	fb 1.00	99.07	96.60	99.90	99.71	98.13	98.24	98.25	377
KDC	knn 25	99.07	96.60	99.90	99.71	98.13	98.24	98.25	3
KDC	knn 50	98.86	95.75	99.90	99.71	97.69	97.81	97.83	4
KDC	knn 75	98.72	95.18	99.90	99.70	97.39	97.52	97.54	5
KDC	knn 100	98.72	95.18	99.90	99.70	97.39	97.52	97.54	5
KDC	knn 150	98.57	94.62	99.90	99.70	97.09	97.23	97.26	5
KDC	fb 0.15	100.00	100.00	100.00	100.00	100.00	100.00	100.00	1
KDC	fb 0.25	100.00	100.00	100.00	100.00	100.00	100.00	100.00	1
KDC	fb 0.50	100.00	100.00	100.00	100.00	100.00	100.00	100.00	2
KDC	fb 0.75	100.00	100.00	100.00	100.00	100.00	100.00	100.00	3
KDC	fb 1.00	99.93	99.72	100.00	100.00	99.86	99.86	99.86	4
KDC	bal 25	99.50	98.87	99.71	99.15	99.01	99.29	99.29	4
KDC	bal 50	99.64	99.43	99.71	99.15	99.29	99.57	99.57	4
KDC	bal 75	99.57	99.43	99.62	98.87	99.15	99.53	99.53	4
KDC	bal 100	99.50	98.58	99.81	99.43	99.00	99.19	99.20	4
KDC	bal 150	99.29	98.02	99.71	99.14	98.58	98.86	98.87	5
LVQ	nn 50	99.50	98.30	99.90	99.71	99.00	99.10	99.10	>1
LVQ	nn 200	99.64	99.15	99.81	99.43	99.29	99.48	99.48	1
LVQ	nn 500	99.79	99.15	100.00	100.00	99.57	99.57	99.58	1
LVQ	nn 1000	99.79	99.15	100.00	100.00	99.57	99.57	99.58	1
OC1	0.1 20 5	99.57	99.15	99.71	99.15	99.15	99.43	99.43	5
OC1	0.3 20 5	99.79	99.15	100.00	100.00	99.57	99.57	99.58	7

les algorithmes est acceptable, hormis parfois les SVM avec certaines options. On notera toutefois les performances particulières du LVQ sur ce dernier point.
Nous notons un défaut majeur commun aux algorithmes LVQ et OC1 : le fait que deux exécutions ne fournissent pas les mêmes résultats. Ce point pose problème quand on a besoin de faire des tests avec des sorties identiques à chaque étape d'une chaîne de traitement par exemple. Les SVM ont pour défaut de permettre uniquement une classification binaire. Pour plus de deux classes, il faut alors exécuter successivement plusieurs fois l'algorithme en rassemblant d'abord des classes en classes principales et en les subdivisant ensuite. Dans le cas du k-NN, si on dispose de plus de deux classes dans l'échantillon d'apprentissage, une classe ne pourra pas toujours être assigné sans ambiguité à une source, à moins d'utiliser le même mécanisme dichotomique que décrit pour les SVM.
Parmi les algorithmes, seule la *kernel density classification* fournit naturellement pour chaque source les probabilités continues d'appartenance aux différentes classes (le k-NN fournit des probabilités discrètes). Cependant la qualité de l'échantillon d'apprentissage est plus importante que pour les SVM ou l'algorithme OC1 pour lesquels seules les frontières importent vraiment. C'est néanmoins l'algorithme que l'on choisira pour la suite. Enfin, dans le cas où l'échantillon d'apprentissage est fortement biaisé, incomplet, etc., l'algorithme des *meanshifts*, seul à être non supervisé dans nos tests, nous semble apporter une approche intéressante même si son temps d'exécution est globalement supérieur à celui des autres algorithmes. Il permet notamment de rechercher des maxima locaux qui correspondraient à des types d'objets absents de l'échantillon d'apprentissage, et donc de découvrir des objets qu'on ne recherche pas.

En conlusion, si on a besoin de probabilités en sortie, on utilisera la KDC. Si notre échantillon d'apprentissage est mal défini, on préférera le *meanshift*. Enfin, si le temps d'éxécution est primordial, on pourra utiliser les algorithmes LVQ, k-NN voire OC1 si le temps d'apprentissage compte peu et le nombre de sources à classer est grand pour ce dernier choix. Enfin, les SVM sont surtout adaptées à des échantillons de taille modeste mais ayant un très grand nombre de dimensions. N'étant pas notre cas, les SVM ne nous apparaissent pas particulièrement pertinentes.

4.6 PERSPECTIVES : PROPOSITION DE CLASSIFICATION

On propose ici une possibilité pour traiter les distributions contenant des *outliers*. On commence par calculer la densité locale en chaque point via un *kernel smoothing*. Un seuil en densité en dessous duquel les sources sont considérées comme des *outliers* est fixé. On classe alors séparément les *outliers* et le reste des sources, éventuellement par des algorithmes différents (OC1 pour les *outliers* et KCD pour le reste par exemple). Cela suppose évidemment qu'on connaisse la classe de certains *outliers*, ce qui est en général le cas.

Pour agrandir l'échantillon d'apprentissage, on peut choisir un échantillon réduit ne contenant que des contreparties presque sûres, avec un taux de fausses associations global de l'ordre de 1% par exemple. On classe ensuite toutes ces sources par KDC et on ne conserve que les meilleurs candidats, ceux ayant une de leurs probabilités d'appartenance à une classe > 0.7 par exemple. Ce seuil peut être ajusté en regardant son effet sur la matrice de confusion des sources de l'EA qui satisfont ce seuil. L'échantillon ainsi constitué peut alors servir de nouvel échantillon d'apprentissage.



Références

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., Allende Prieto, C., An, D., Anderson, K. S. J., Anderson, S. F., Annis, J., Bahcall, N. A., Bailer-Jones, C. A. L., Barentine, J. C., Bassett, B. A., Becker, A. C., Beers, T. C., Bell, E. F., Belokurov, V., Berlind, A. A., Berman, E. F., Bernardi, M., Bickerton, S. J., Bizyaev, D., Blakeslee, J. P., Blanton, M. R., Bochanski, J. J., Boroski, W. N., Brewington, H. J., Brinchmann, J., Brinkmann, J., Brunner, R. J., Budavári, T., Carey, L. N., Carliles, S., Carr, M. A., Castander, F. J., Cinabro, D., Connolly, A. J., Csabai, I., Cunha, C. E., Czarapata, P. C., Davenport, J. R. A., de Haas, E., Dilday, B., Doi, M., Eisenstein, D. J., Evans, M. L., Evans, N. W., Fan, X., Friedman, S. D., Frieman, J. A., Fukugita, M., Gänsicke, B. T., Gates, E., Gillespie, B., Gilmore, G., Gonzalez, B., Gonzalez, C. F., Grebel, E. K., Gunn, J. E., Györy, Z., Hall, P. B., Harding, P., Harris, F. H., Harvanek, M., Hawley, S. L., Hayes, J. J. E., Heckman, T. M., Hendry, J. S., Hennessy, G. S., Hindsley, R. B., Hoblitt, J., Hogan, C. J., Hogg, D. W., Holtzman, J. A., Hyde, J. B., Ichikawa, S.-i., Ichikawa, T., Im, M., Ivezić, Ž., Jester, S., Jiang, L., Johnson, J. A., Jorgensen, A. M., Jurić, M., Kent, S. M., Kessler, R., Kleinman, S. J., Knapp, G. R., Konishi, K., Kron, R. G., Krzesinski, J., Kuropatkin, N., Lampeitl, H., Lebedeva, S., Lee, M. G., Lee, Y. S., Leger, R. F., Lépine, S., Li, N., Lima, M., Lin, H., Long, D. C., Loomis, C. P., Loveday, J., Lupton, R. H., Magnier, E., Malanushenko, O., Malanushenko, V., Mandelbaum, R., Margon, B., Marriner, J. P., Martínez-Delgado, D., Matsubara, T., McGehee, P. M., McKay, T. A., Meiksin, A., Morrison, H. L., Mullally, F., Munn, J. A., Murphy, T., Nash, T., Nebot, A., Neilsen, E. H., Newberg, H. J., Newman, P. R., Nichol, R. C., Nicinski, T., Nieto-Santisteban, M., Nitta, A., Okamura, S., Oravetz, D. J., Ostriker, J. P., Owen, R., Padmanabhan, N., Pan, K., Park, C., Pauls, G., Peoples, J., Percival, W. J., Pier, J. R., Pope, A. C., Pourbaix, D., Price, P. A., Purger, N., Quinn, T., Raddick, M. J., Fiorentin, P. R., Richards, G. T., Richmond, M. W., Riess, A. G., Rix, H.-W., Rockosi, C. M., Sako, M., Schlegel, D. J., Schneider, D. P., Scholz, R.-D., Schreiber, M. R., Schwope, A. D., Seljak, U., Sesar, B., Sheldon, E., Shimasaku, K., Sibley, V. C., Simmons, A. E., Sivarani, T., Smith, J. A., Smith, M. C., Smolčić, V., Snedden, S. A., Stebbins, A., Steinmetz, M., Stoughton, C., Strauss, M. A., Subba Rao, M., Suto, Y., Szalay, A. S., Szapudi, I. & et al., S., 2009 ; *The Seventh Data Release of the Sloan Digital Sky Survey*. ApJS, **182**, 543

Alexander, D. M., Bauer, F. E., Brandt, W. N., Schneider, D. P., Hornschemeier, A. E., Vignali, C., Barger, A. J., Broos, P. S., Cowie, L. L., Garmire, G. P., Townsley, L. K., Bautz, M. W., Chartas, G. & Sargent, W. L. W., 2003 ; *The Chandra Deep Field North Survey. XIII. 2 Ms Point-Source Catalogs*. AJ, **126**, 539

Barcons, X., Carrera, F. J., Ceballos, M. T., Page, M. J., Bussons-Gordo, J., Corral, A., Ebrero, J., Mateos, S., Tedds, J. A., Watson, M. G., Baskill, D., Birkinshaw, M., Boller, T., Borisov, N., Bremer, M., Bromage, G. E., Brunner, H., Caccianiga, A., Crawford, C. S., Cropper, M. S., Della Ceca, R., Derry, P., Fabian, A. C., Guillout, P., Hashimoto, Y., Hasinger, G., Hassall, B. J. M., Lamer, G., Loaring, N. S., Maccacaro, T., Mason, K. O., McMahon, R. G., Mirioni, L., Mittaz, J. P. D., Motch, C., Negueruela, I., Osborne, J. P., Panessa, F., Pérez-Fournon, I., Pye, J. P., Roberts, T. P., Rosen, S., Schartel, N., Schurch, N., Schwope, A., Severgnini, P., Sharp, R., Stewart, G. C., Szokoly, G., Ullán, A., Ward, M. J., Warwick, R. S., Wheatley, P. J., Webb, N. A., Worrall, D., Yuan, W. & Ziaeepour, H., 2007 ; *The XMM-Newton serendipitous survey. IV. Optical identification of the XMM-Newton medium sensitivity survey (XMS)*. A&A, **476**, 1191

Barcons, X., Carrera, F. J., Watson, M. G., McMahon, R. G., Aschenbach, B., Freyberg, M. J., Page, K., Page, M. J., Roberts, T. P., Turner, M. J. L., Barret, D., Brunner, H., Ceballos, M. T., Della Ceca, R., Guillout, P., Hasinger, G., Maccacaro, T., Mateos, S., Motch, C., Negueruela, I., Osborne, J. P., Pérez-Fournon, I., Schwope, A., Severgnini, P., Szokoly, G. P., Webb, N. A., Wheatley, P. J. & Worrall, D. M., 2002 ; *The XMM-Newton serendipitous survey . II. First results from the AXIS high galactic latitude medium sensitivity survey*. A&A, **382**, 522

Bochanski, J. J., West, A. A., Hawley, S. L. & Covey, K. R., 2007 ; *Low-Mass Dwarf Template Spectra from the Sloan Digital Sky Survey*. AJ, **133**, 531

Brusa, M., Zamorani, G., Comastri, A., Hasinger, G., Cappelluti, N., Civano, F., Finoguenov, A., Mainieri, V., Salvato, M., Vignali, C., Elvis, M., Fiore, F., Gilli, R., Impey, C. D., Lilly, S. J., Mignoli, M., Silverman, J., Trump, J., Urry, C. M., Bender, R.,

CAPAK, P., HUCHRA, J. P., KNEIB, J. P., KOEKEMOER, A., LEAUTHAUD, A., LEHMANN, I., MASSEY, R., MATUTE, I., MCCARTHY, P. J., MCCRACKEN, H. J., RHODES, J., SCOVILLE, N. Z., TANIGUCHI, Y. & THOMPSON, D., 2007 ; *The XMM-Newton Wide-Field Survey in the COSMOS Field. III. Optical Identification and Multiwavelength Properties of a Large Sample of X-Ray-Selected Sources*. ApJS, **172**, 353

CACCIANIGA, A., SEVERGNINI, P., BRAITO, V., DELLA CECA, R., MACCACARO, T., WOLTER, A., BARCONS, X., CARRERA, F. J., LEHMANN, I., PAGE, M. J., SAXTON, R. & WEBB, N. A., 2004 ; *The XMM-Newton HBS28 sample : Studying the obscuration in hard X-ray selected AGNs*. A&A, **416**, 901

CACCIANIGA, A., SEVERGNINI, P., DELLA CECA, R., MACCACARO, T., CARRERA, F. J. & PAGE, M. J., 2007 ; *Elusive AGN in the XMM-Newton bright serendipitous survey*. A&A, **470**, 557

CARRERA, F. J., EBRERO, J., MATEOS, S., CEBALLOS, M. T., CORRAL, A., BARCONS, X., PAGE, M. J., ROSEN, S. R., WATSON, M. G., TEDDS, J. A., DELLA CECA, R., MACCACARO, T., BRUNNER, H., FREYBERG, M., LAMER, G., BAUER, F. E. & UEDA, Y., 2007 ; *The XMM-Newton serendipitous survey. III. The AXIS X-ray source counts and angular clustering*. A&A, **469**, 27

COCCHIA, F., FIORE, F., VIGNALI, C., MIGNOLI, M., BRUSA, M., COMASTRI, A., FERUGLIO, C., BALDI, A., CARANGELO, N., CILIEGI, P., D'ELIA, V., LA FRANCA, F., MAIOLINO, R., MATT, G., MOLENDI, S., PEROLA, G. C. & PUCCETTI, S., 2007 ; *The HELLAS2XMM survey. VIII. Optical identifications of the extended sample*. A&A, **466**, 31

COVEY, K. R., AGÜEROS, M. A., GREEN, P. J., HAGGARD, D., BARKHOUSE, W. A., DRAKE, J., EVANS, N., KASHYAP, V., KIM, D.-W., MOSSMAN, A., PEASE, D. O. & SILVERMAN, J. D., 2008 ; *The ChaMP Extended Stellar Survey (ChESS) : Photometric and Spectroscopic Properties of Serendipitously Detected Stellar X-Ray Sources*. ApJS, **178**, 339

COVEY, K. R., IVEZIĆ, Ž., SCHLEGEL, D., FINKBEINER, D., PADMANABHAN, N., LUPTON, R. H., AGÜEROS, M. A., BOCHANSKI, J. J., HAWLEY, S. L., WEST, A. A., SETH, A., KIMBALL, A., GOGARTEN, S. M., CLAIRE, M., HAGGARD, D., KAIB, N., SCHNEIDER, D. P. & SESAR, B., 2007 ; *Stellar SEDs from 0.3 to 2.5 μm : Tracing the Stellar Locus and Searching for Color Outliers in the SDSS and 2MASS*. AJ, **134**, 2398

DELLA CECA, R., MACCACARO, T., CACCIANIGA, A., SEVERGNINI, P., BRAITO, V., BARCONS, X., CARRERA, F. J., WATSON, M. G., TEDDS, J. A., BRUNNER, H., LEHMANN, I., PAGE, M. J., LAMER, G. & SCHWOPE, A., 2004 ; *Exploring the X-ray sky with the XMM-Newton bright serendipitous survey*. A&A, **428**, 383

DILLON, M., GÄNSICKE, B. T., AUNGWEROJWIT, A., RODRÍGUEZ-GIL, P., MARSH, T. R., BARROS, S. C. C., SZKODY, P., BRADY, S., KRAJCI, T. & OKSANEN, A., 2008 ; *Orbital periods of cataclysmic variables identified by the SDSS - III. Time-series photometry obtained during the 2004/5 International Time Project on La Palma*. MNRAS, **386**, 1568

DOWNES, R. A., WEBBINK, R. F., SHARA, M. M., RITTER, H., KOLB, U. & DUERBECK, H. W., 2001 ; *A Catalog and Atlas of Cataclysmic Variables : The Living Edition*. PASP, **113**, 764

FLEMING, T. A., SNOWDEN, S. L., PFEFFERMANN, E., BRIEL, U. & GREINER, J., 1996 ; *Catalogue and luminosity function of white dwarfs detected in the ROSAT all-sky survey.* A&A, **316**, 147

GAO, D., ZHANG, Y.-X. & ZHAO, Y.-H., 2008 ; *Support vector machines and kd-tree for separating quasars from large survey data bases*. MNRAS, **386**, 1417

GIRARDI, L., GREBEL, E. K., ODENKIRCHEN, M. & CHIOSI, C., 2004 ; *Theoretical isochrones in several photometric systems. II. The Sloan Digital Sky Survey ugriz system*. A&A, **422**, 205

GUEDEL, M., GUINAN, E. F. & SKINNER, S. L., 1997 ; *The X-Ray Sun in Time : A Study of the Long-Term Evolution of Coronae of Solar-Type Stars*. ApJ, **483**, 947

GUEDEL, M. & NAZE, Y., 2009 ; *X-Ray Spectroscopy of Stars*. ArXiv e-prints

GUILLOUT, P., 2009 ; *Low-Mass Dwarf Template Spectra from the Sloan Digital Sky Survey*. A&A

HANDS, A. D. P., WARWICK, R. S., WATSON, M. G. & HELFAND, D. J., 2004 ; *X-ray source populations in the Galactic plane*. MNRAS, **351**, 31

HÉRENT, O., MOTCH, C. & GUILLOUT, P., 2006 ; *Unveiling Galactic Stellar Populations through XMM-Newton Surveys*. Dans *The X-ray Universe 2005*, A. Wilson, réd., tm. 604 de *ESA Special Publication*

HØG, E., FABRICIUS, C., MAKAROV, V. V., URBAN, S., CORBIN, T., WYCOFF, G., BASTIAN, U., SCHWEKENDIEK, P. & WICENEC, A., 2000 ; *The Tycho-2 catalogue of the 2.5 million brightest stars*. A&A, **355**, L27

JASCHEK, M. & EGRET, D., 1982 ; *A Catalogue of Be-Stars*. Dans *Be Stars*, M. Jaschek & H.-G. Groth, réds., tm. 98 de *IAU Symposium*

JOACHIMS, T., 1999 ; *Making large-Scale SVM Learning Practical*. Advances in Kernel Methods - Support Vector Learning

KOHONEN, T., HYNNINEN, J., KANGAS, J., LAAKSONEN, J. & TORKKOLA, K., 1996 ; *LVQ PAK : The Learning Vector Quantization Program Package*. Rap. tech., Helsinki University of Technology, Laboratory of Computer

KOHONEN, T., KANGAS, J., LAAKSONEN, J. & TORKKOLA, K., 1992 ; *LVQ PAK : A program package for the correct application of Learning Vector Quantization algorithms*. IEEE

KRAUTTER, J., WICHMANN, R., SCHMITT, J. H. M. M., ALCALA, J. M., NEUHAUSER, R. & TERRANEGRA, L., 1997 ; *New "weak-line"–T Tauri stars in Lupus*. A&AS, **123**, 329

KRUMPE, M., LAMER, G., SCHWOPE, A. D., WAGNER, S., ZAMORANI, G., MIGNOLI, M., STAUBERT, R., WISOTZKI, L. & HASINGER, G., 2007 ; *The XMM-Newton survey in the Marano field. I. The X-ray data and optical follow-up*. A&A, **466**, 41

LASKER, B. M., LATTANZI, M. G., MCLEAN, B. J., BUCCIARELLI, B., DRIMMEL, R., GARCIA, J., GREENE, G., GUGLIELMETTI, F., HANLEY, C., HAWKINS, G., LAIDLER, V. G., LOOMIS, C., MEAKES, M., MIGNANI, R., MORBIDELLI, R., MORRISON, J., PANNUNZIO, R., ROSENBERG, A., SARASSO, M., SMART, R. L., SPAGNA, A., STURCH, C. R., VOLPICELLI, A., WHITE, R. L., WOLFE, D. & ZACCHEI, A., 2008 ; *The Second-Generation Guide Star Catalog : Description and Properties*. AJ, **136**, 735

LAUBERTS, A., 1982 ; *ESO/Uppsala survey of the ESO(B) atlas*. Garching : European Southern Observatory (ESO), 1982

LIU, Q. Z., VAN PARADIJS, J. & VAN DEN HEUVEL, E. P. J., 2000 ; *A catalogue of high-mass X-ray binaries*. A&AS, **147**, 25

—, 2001 ; *A catalogue of low-mass X-ray binaries*. A&A, **368**, 1021

—, 2006 ; *Catalogue of high-mass X-ray binaries in the Galaxy (4th edition)*. A&A, **455**, 1165

MATEOS, S., WARWICK, R. S., CARRERA, F. J., STEWART, G. C., EBRERO, J., DELLA CECA, R., CACCIANIGA, A., GILLI, R., PAGE, M. J., TREISTER, E., TEDDS, J. A., WATSON, M. G., LAMER, G., SAXTON, R. D., BRUNNER, H. & PAGE, C. G., 2008 ; *High precision X-ray log N - log S distributions : implications for the obscured AGN population*. A&A, **492**, 51

MCCOOK, G. P. & SION, E. M., 1999 ; *A Catalog of Spectroscopically Identified White Dwarfs*. ApJS, **121**, 1

MOTCH, C., 2006*a* ; *Galactic Plane Surveys*. Dans *The X-ray Universe 2005*, A. Wilson, réd., tm. 604 de *ESA Special Publication*

—, 2006*b* ; *Galactic Plane Surveys*. Dans *The X-ray Universe 2005*, A. Wilson, réd., tm. 604 de *ESA Special Publication*

MOTCH, C., GUILLOUT, P., HABERL, F., PAKULL, M., PIETSCH, W. & REINSCH, K., 1997 ; *The ROSAT galactic plane survey : analysis of a low latitude sample area in Cygnus*. A&A, **318**, 111

MOTCH, C., HERENT, O. & GUILLOUT, P., 2003 ; *The source content of low galactic latitude XMM-Newton surveys*. Astronomische Nachrichten, **324**, 61

PADOVANI, P., ALLEN, M. G., ROSATI, P. & WALTON, N. A., 2004 ; *Discovery of optically faint obscured quasars with Virtual Observatory tools*. A&A, **424**, 545

PTAK, A., ZAKAMSKA, N. L., STRAUSS, M. A., KROLIK, J. H., HECKMAN, T. M., SCHNEIDER, D. P. & BRINKMANN, J., 2006 ; *Type II Quasars from the Sloan Digital Sky Survey. IV. Chandra and XMM-Newton Observations Reveal Heavily Absorbed Sources*. ApJ, **637**, 147

RICHARDS, G. T., FAN, X., NEWBERG, H. J., STRAUSS, M. A., VANDEN BERK, D. E., SCHNEIDER, D. P., YANNY, B., BOUCHER, A., BURLES, S., FRIEMAN, J. A., GUNN, J. E., HALL, P. B., IVEZIĆ, Ž., KENT, S., LOVEDAY, J., LUPTON, R. H., ROCKOSI, C. M., SCHLEGEL, D. J., STOUGHTON, C., SUBBARAO, M. & YORK, D. G., 2002 ; *Spectroscopic Target Selection in the Sloan Digital Sky Survey : The Quasar Sample*. AJ, **123**, 2945

RITTER, H. & KOLB, U., 2003 ; *Catalogue of cataclysmic binaries, low-mass X-ray binaries and related objects (Seventh edition)*. A&A, **404**, 301

ROGEL, A. B., LUGGER, P. M., COHN, H. N., SLAVIN, S. D., GRINDLAY, J. E., ZHAO, P. & HONG, J., 2006 ; *Three Years of ChaMPlane Northern Field WIYN Spectroscopy*. ApJS, **163**, 160

SCHLEGEL, D. J., FINKBEINER, D. P. & DAVIS, M., 1998 ; *Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds*. ApJ, **500**, 525

SCHMIDT, M., HASINGER, G., GUNN, J., SCHNEIDER, D., BURG, R., GIACCONI, R., LEHMANN, I., MACKENTY, J., TRUMPER, J. & ZAMORANI, G., 1998 ; *The ROSAT deep survey. II. Optical identification, photometry and spectra of X-ray sources in the Lockman field*. A&A, **329**, 495

SKRUTSKIE, M. F., CUTRI, R. M., STIENING, R., WEINBERG, M. D., SCHNEIDER, S., CARPENTER, J. M., BEICHMAN, C., CAPPS, R., CHESTER, T., ELIAS, J., HUCHRA, J., LIEBERT, J., LONSDALE, C., MONET, D. G., PRICE, S., SEITZER, P., JARRETT, T., KIRKPATRICK, J. D., GIZIS, J. E., HOWARD, E., EVANS, T., FOWLER, J., FULLMER, L., HURT, R., LIGHT, R., KOPAN, E. L., MARSH, K. A., MCCALLON, H. L., TAM, R., VAN DYK, S. & WHEELOCK, S., 2006 ; *The Two Micron All Sky Survey (2MASS)*. AJ, **131**, 1163

STASIŃSKA, G., CID FERNANDES, R., MATEUS, A., SODRÉ, L. & ASARI, N. V., 2006 ; *Semi-empirical analysis of Sloan Digital Sky Survey galaxies - III. How to distinguish AGN hosts*. MNRAS, **371**, 972

STASIŃSKA, G., VALE ASARI, N., CID FERNANDES, R., GOMES, J. M., SCHLICKMANN, M., MATEUS, A., SCHOENELL, W. & SODRÉ, JR., L., 2008 ; *Can retired galaxies mimic active galaxies ? Clues from the Sloan Digital Sky Survey*. MNRAS, **391**, L29

STOCKE, J. T., MORRIS, S. L., GIOIA, I. M., MACCACARO, T., SCHILD, R., WOLTER, A., FLEMING, T. A. & HENRY, J. P., 1991 ; *The Einstein Observatory Extended Medium-Sensitivity Survey. II - The optical identifications*. ApJS, **76**, 813

STRATEVA, I., IVEZIĆ, Ž., KNAPP, G. R., NARAYANAN, V. K., STRAUSS, M. A., GUNN, J. E., LUPTON, R. H., SCHLEGEL, D., BAHCALL, N. A., BRINKMANN, J., BRUNNER, R. J., BUDAVÁRI, T., CSABAI, I., CASTANDER, F. J., DOI, M., FUKUGITA, M., GYŐRY, Z., HAMABE, M., HENNESSY, G., ICHIKAWA, T., KUNSZT, P. Z., LAMB, D. Q., MCKAY, T. A., OKAMURA, S., RACUSIN, J., SEKIGUCHI, M., SCHNEIDER, D. P., SHIMASAKU, K. & YORK, D., 2001 ; *Color Separation of Galaxy Types in the Sloan Digital Sky Survey Imaging Data*. AJ, **122**, 1861

SUNG, H., BESSELL, M. S. & SANA, H., 2008 ; *Coronal Temperature as AN Age Indicator*. Journal of Korean Astronomical Society, **41**, 1

TAYLOR, J. H., MANCHESTER, R. N. & LYNE, A. G., 1993 ; *Catalog of 558 pulsars*. ApJS, **88**, 529

VAN DER HUCHT, K. A., 2001 ; *The VIIth catalogue of galactic Wolf-Rayet stars*. New Astronomy Review, **45**, 135

VÉRON-CETTY, M.-P. & VÉRON, P., 2006 ; *A catalogue of quasars and active nuclei : 12th edition*. A&A, **455**, 773

WHITE, R. L., 2008 ; *Astronomical Applications of Oblique Decision Trees*. Dans *American Institute of Physics Conference Series*, tm. 1082 de *American Institute of Physics Conference Series*

YANNY, B., ROCKOSI, C., NEWBERG, H. J., KNAPP, G. R., ADELMAN-MCCARTHY, J. K., ALCORN, B., ALLAM, S., ALLENDE PRIETO, C., AN, D., ANDERSON, K. S. J., ANDERSON, S., BAILER-JONES, C. A. L., BASTIAN, S., BEERS, T. C., BELL, E., BELOKUROV, V., BIZYAEV, D., BLYTHE, N., BOCHANSKI, J. J., BOROSKI, W. N., BRINCHMANN, J., BRINKMANN, J., BREWINGTON, H., CAREY, L., CUDWORTH, K. M., EVANS, M., EVANS, N. W., GATES, E., GÄNSICKE, B. T., GILLESPIE, B., GILMORE, G., GOMEZ-MORAN, A. N., GREBEL, E. K., GREENWELL, J., GUNN, J. E., JORDAN, C., JORDAN, W., HARDING, P., HARRIS, H., HENDRY, J. S., HOLDER, D., IVANS, I. I., IVEZIČ, Ž., JESTER, S., JOHNSON, J. A., KENT, S. M., KLEINMAN, S., KNIAZEV, A., KRZESINSKI, J., KRON, R., KUROPATKIN, N., LEBEDEVA, S., LEE, Y. S., LEGER, R. F., LÉPINE, S., LEVINE, S., LIN, H., LONG, D. C., LOOMIS, C., LUPTON, R., MALANUSHENKO, O., MALANUSHENKO, V., MARGON, B., MARTINEZ-DELGADO, D., MCGEHEE, P., MONET, D., MORRISON, H. L., MUNN, J. A., NEILSEN, E. H., NITTA, A., NORRIS, J. E., ORAVETZ, D., OWEN, R., PADMANABHAN, N., PAN, K., PETERSON, R. S., PIER, J. R., PLATSON, J., FIORENTIN, P. R., RICHARDS, G. T., RIX, H.-W., SCHLEGEL, D. J., SCHNEIDER, D. P., SCHREIBER, M. R., SCHWOPE, A., SIBLEY, V., SIMMONS, A., SNEDDEN, S. A., SMITH, J. A., STARK, L., STAUFFER, F., STEINMETZ, M., STOUGHTON, C., SUBBA RAO, M., SZALAY, A., SZKODY, P., THAKAR, A. R., THIRUPATHI, S., TUCKER, D., UOMOTO, A., VANDEN BERK, D., VIDRIH, S., WADADEKAR, Y., WATTERS, S., WILHELM, R., WYSE, R. F. G., YARGER, J. & ZUCKER, D., 2009 ; *SEGUE : A Spectroscopic Survey of 240,000 Stars with g = 14-20*. AJ, **137**, 4377

YOUNG, M., ELVIS, M. & RISALITI, G., 2008 ; *The Reddest DR3 SDSS/XMM-Newton Quasars*. ApJ, **688**, 128

ZAKAMSKA, N. L., STRAUSS, M. A., KROLIK, J. H., COLLINGE, M. J., HALL, P. B., HAO, L., HECKMAN, T. M., IVEZIĆ, Ž., RICHARDS, G. T., SCHLEGEL, D. J., SCHNEIDER, D. P., STRATEVA, I., VANDEN BERK, D. E., ANDERSON, S. F. & BRINKMANN, J., 2003 ; *Candidate Type II Quasars from the Sloan Digital Sky Survey. I. Selection and Optical Properties of a Sample at 0.3Z0.83*. AJ, **126**, 2125

ZICKGRAF, F.-J., THIERING, I., KRAUTTER, J., APPENZELLER, I., KNEER, R., VOGES, W. H., ZIEGLER, B., CHAVARRIA, C., SERRANO, A., MUJICA, R., PAKULL, M. & HEIDT, J., 1997 ; *Identification of a complete sample of northern ROSAT All-Sky Survey X-ray sources. II. The optical observations*. A&AS, **123**, 103

Chapitre V

Interactions avec la XCAT-DB

Nous avons vu dans les chapitres précédents l'intérêt de disposer de probabilités d'identification associées aux corrélations croisées. Elles offrent, avec notamment la classification des sources qui les composent, la possibilité de créer des échantillons de référence, d'étudier des problématiques spécifiques, de rechercher des *outliers* qui sont potentiellement des objets rares, etc. Aussi, pour permettre de plus amples études et une exploitation optimale de la richesse de ces corrélations et de ces classifications, il convient de mettre à disposition de la communauté les résultats de ces travaux. La XCAT-DB étant la seule implémentation des corrélations, il nous a paru naturel d'ajouter ces informations aux données, déjà riches, qu'elle comporte.

Ce chapitre aborde l'implémentation de l'algorithme de corrélation décrit au chapitre II, ainsi que les interactions avec la XCAT-DB, à la fois pour l'extraction et l'ajout de données. Après une brève présentation de la XCAT-DB et de SAADA, nous détaillerons certains aspects du module de corrélation que nous avons développé. Nous verrons enfin de quelle manière nous avons inséré les résultats des corrélations croisées et des classifications dans la XCAT-DB.

1 Présentation de la XCAT-DB et de SAADA

1.1 La XCAT-DB

La XCAT-DB, pour *Xmm CATalogue DataBase*, est une base de données développée par Laurent Michel au sein de l'équipe Hautes Énergies de l'Observatoire Astronomique de Strasbourg. Comme le signifie l'acronyme, cette base est dédiée à l'hébergement du catalogue 2XMMi[1] (voir Tab. V.1), et aux produits associés issus du pipeline de traitement de données du SSC (Michel *et al.* 2004, Motch *et al.* 2009). Ainsi, sa spécificité par rapport aux autres points d'accès aux sources XMM est d'offrir aux utilisateurs, en plus du catalogue, les cartes de pointages, les spectres RGS, les détections EPIC, et surtout les corrélations avec de nombreux (202) catalogues d'archives et les entrées associées, des liens vers Simbad et Aladin, etc. La présence de relations entre tous ces produits permet d'effectuer facilement des recherches complexes avec un temps de réponse relativement court. On peut par exemple accéder de façon simple aux sources X qui ont un spectre EPIC et des contreparties optiques peu brillantes distantes de moins de deux secondes d'arc. L'accès aux données de la XCAT-DB peut se faire de quatre façons différentes : soit en passant par l'interface Web dédiée, soit par les protocoles de l'Observatoire Virtuel[2] (OV) implémentés dans SAADA[3] , soit par une API Java, soit directement en SQL via une connexion au serveur PostgreSQL.

La XCAT-DB est basée sur SAADA (voir section 1.2 suivante) et sur des scripts spécifiques de chargement écrits en *shell*. Ces scripts font appel au *dataloader* de SAADA ainsi qu'à de nombreuses requêtes SQL notamment pour le remplissage des relations. Nous avons eu l'occasion

[1] `http://amwdb.u-strasbg.fr/2xmmi/home`

[2] Pour une introduction sur l'OV, voir Djorgovski & Williams (2005)

[3] En l'occurrence SIA, SSA et ConeSearch

de travailler sur ces scripts pour assurer la compatibilité avec la nouvelle version de SAADA et, comme on le verra plus loin dans ce chapitre, pour y inclure le chargement de nos propres corrélations. Notons enfin que l'interface Web dédiée à la XCAT-DB, basée sur l'interface Web native de SAADA, a demandé à son auteur un développement particulier pour mettre en valeur les produits du 2XMMi. Elle contient par exemple des formulaires spécifiques qui proposent des critères de sélection propres au 2XMMi ou des filtres pour la recherche de catalogues de contreparties. Les critères de sélection portent par exemple sur les liens entre sources du catalogue et sources d'archive via l'utilisation des *matchPatterns*[4].

TAB. V.1 – Quelques chiffres sur le contenu de la XCAT-DB

Données	Nombre d'entrées
Sources 2XMMi	289 083
Catalogues d'archives	202 (de 385 000 à 4 entrées)
Sources d'archives	2 481 642
Flatfiles (FC et autres Plots)	5 386 616
Spectres EPIC	96 176
Séries temporelles EPIC	96 027
Images EPIC	∼85 000
Spectres RGS	∼25 000

1.2 SAADA

SAADA[5] est l'acronyme de « Système d'Archivage Auto-configurable de Données Astronomiques ». C'est un logiciel de création de bases de données locales orientées astronomie (Michel *et al.* 2005, 2007, 2008) qui permet également une publication rapide des données utilisateur sur l'OV (Michel *et al.* 2006). La première version de SAADA est issue d'un travail de thèse (Nguyen 2006) réalisé au sein de l'équipe Hautes Énergies de l'observatoire (Nguyen *et al.* 2004, Ngoc Nguyen *et al.* 2005, Nguyen *et al.* 2006). Depuis, SAADA continue d'évoluer par des interventions extérieures ponctuelles, mais surtout par les développements apportés par Laurent Michel. Nous avons eu l'occasion de travailler sur l'amélioration et la refonte du moteur de requêtes de SAADA, sur l'intégration des requêtes basées sur les UCD[6] et les *UTypes*[7], ainsi qu'à l'élaboration et l'implémentation du moteur de remplissage automatique de relations. Ces deux sujets ne seront pas abordés dans le présent manuscrit.

On peut identifier essentiellement quatre étapes dans la création et la publication d'une SAADA-DB (base de données crée par SAADA). En premier lieu, SAADA crée un *template* de base de données qui obéit à un modèle de données générique propre à SAADA (voir Fig. V.1). Dans un second temps, l'utilisateur définit des ensembles de données et des relations vides en fonction de ses besoins et suivant le *datamodel* de SAADA. Puis il remplit la base via le *dataloader*. Enfin, l'interface Web incluant les accès OV peut être déployée.

SAADA, qui repose sur une base de données relationnelle peut supporter des données externes dans de nouvelles tables de la base. Ces données sont simplement ignorées : il n'y a pas un *mapping* complet de la base dans SAADA, SAADA utilise juste les tables dont il a besoin et ignore les autres.

[4] Syntaxe pour exprimer une contrainte sur un ensemble de liens

[5] `http://amwdb.u-strasbg.fr/saada`

[6] Voir : `http://www.ivoa.net/Documents/latest/UCD.html`

[7] Voir : `http://www.ivoa.net/cgi-bin/twiki/bin/view/IVOA/Utypes`

1.3 Genèse, modèle de données de SAADA et unicité des sources

1.3.1 La genèse

La XCAT-DB reposait à l'origine sur le système de gestion de base de données orienté objet (OODBMS, pour *Objects-Oriented DataBase Management System*) O2. Ce système était capable de rendre persistant n'importe quel modèle orienté objet. Mais l'arrêt du projet O2 en 2000 a forcé le SSC à trouver une solution de remplacement. Deux solutions pour remplacer O2 ont d'abord été testées : Castor[8] et PostgreSQL[9]. Castor est un ORDBMS (*Object-Relational DataBase Management System*) basé sur la technologie Java et sur le format XML. Malheureusement, Castor ne gérait pas les relations NxM. Bien qu'il soit aujourd'hui essentiellement utilisé en tant que RDBMS (*Relational DataBase Management System*), PostgreSQL consiste en un ORDBMS qui gère l'héritage et permet d'avoir un vecteur comme type d'attribut. Cependant, PostgreSQL ne savait pas faire de requêtes portant sur le contenu de vecteurs, le stockage des objets prenait beaucoup d'espace disque et les requêtes comportant des objets étaient globalement assez peu performantes.

À partir de là est né le projet SAADA dont la philosophie peut se résumer à ceci : garder le meilleur du monde relationnel avec la souplesse de l'orienté objet. Cela s'est traduit par le développement d'une couche objet en Java qui repose sur une base de données relationnelle. Un modèle de données spécifique à été développé pour SAADA. Les problèmes de performances ont été résolus grâce au moteur de requêtes basé sur un système d'index spécifique et un langage de requêtes adapté au modèle de données de SAADA. Le modèle de données de SAADA, conçu à l'origine pour la XCAT-DB, est suffisamment générique pour supporter le stockage de n'importe quelles données astronomiques (ou presque). Pour pouvoir être utilisé facilement par les membres de la communauté astronomique, SAADA possède un aspect autoconfigurable. De fait, il permet à n'importe quel astronome de stocker ses données dans une base et de les publier, le tout sans écrire une ligne de code. Il existe aujourd'hui plusieurs SAADA-DB autres que la XCAT-DB à travers le monde.

1.3.2 Le modèle de données de SAADA

La notion clé du modèle de données de SAADA est la notion de « collection ». Une collection contient un ensemble de données hétérogènes rangées par catégories. Elle peut contenir un jeu de données d'une ou plusieurs des catégories de produits suivantes : *images*, *spectres*, *tables*, *flatfile* ou *misc* (pour *miscellaneous*). Toutes les données d'une même collection partagent des caractéristiques communes ou un intérêt scientifique commun. Elles peuvent par exemple provenir d'une même mission scientifique (XMM, Chandra,...), rassembler une catégorie d'objets astrophysiques (galaxie, quasar, étoile, ...), une plage de longueurs d'onde (IR, X,...) etc. C'est l'opérateur, le scientifique qui veut construire une base de données, qui définit les collections qu'elle contiendra. Une collection peut contenir différentes catégories de « produits », mais il ne peut y avoir qu'un seul ensemble de produits par catégorie. Chaque catégorie de produits contient une ou plusieurs « classes » dans lesquelles sont stockés les produits de formats compatibles. Le schéma de la figure V.1 présente la structure de données propre à SAADA.

Il faut ajouter à cette structure la notion de « relation permanente ». Une SAADA-DB est capable de stocker des liens entre les enregistrements contenus dans les différents produits. Ce sont ces liens qui enrichissent le contenu scientifique d'un ensemble de données hétérogènes. Une relation permanente relie un ensemble de produits d'une catégorie et d'une collection données à un autre ensemble de produits d'une catégorie et d'une collection données. La relation est un ensemble de liens partant d'un enregistrement donné vers des enregistrements cibles différents. Les relations sont nommées et chaque lien possède des « *qualifiers* » définis par l'utilisateur. Un *qualifier* est une valeur numérique qui qualifie chaque lien. Par exemple, le lien entre deux sources peut posséder comme *qualifier* la distance entre ces sources. Un *qualifier* est un attribut d'une relation. Dans le langage SAADA-QL[10], l'élément « *matchPattern* » permet de poser des contraintes

[8] `http://www.castor.org/index.html`

[9] `http://www.postgresql.org/`

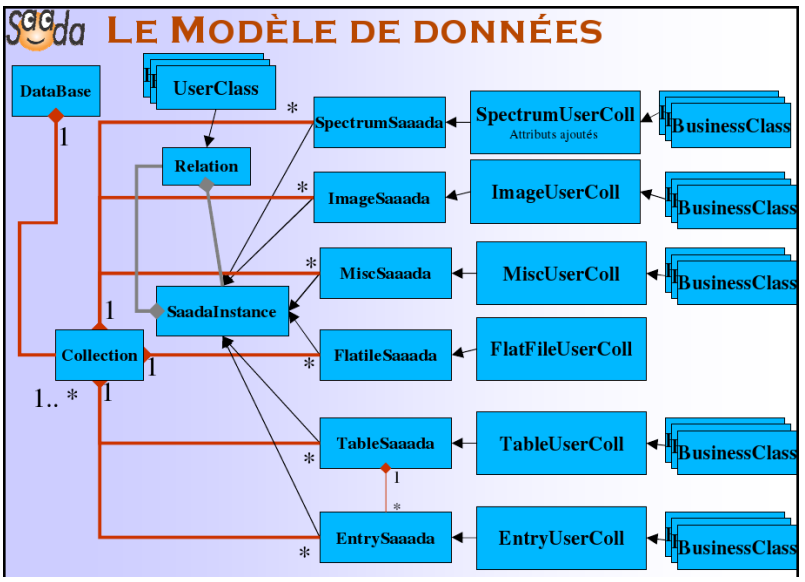


FIG. V.1 – Schéma de la structure de données d'une SAADA-DB (crédits Laurent Michel).

sur une relation et les *qualifiers* qui la composent.

Prenons l'exemple d'une SAADA-DB qui contiendrait une liste de sources X dans une collection nommée « SourcesX » et une liste de sources optiques dans une collection nommée « SourcesOpt ». Une corrélation sera par exemple composée de tous les liens entre une source X et les sources optiques dans un rayon de 5 secondes d'arc. La relation sera alors composée de l'ensemble des corrélations des sources X avec les sources optiques. On peut alors utiliser la clause *matchPattern* du langage de requêtes SAADA-QL pour sélectionner, par exemple, toutes les sources X ayant une contrepartie optique distante de moins de 2 secondes d'arc.

1.3.3 Unicité des sources d'archive

Un des problèmes auxquels on est confronté quand on charge les produits de sortie du pipeline de traitement de données du SSC concerne la multiplicité de certaines sources d'archive. Cette multiplicité peut résulter soit de la corrélation de deux sources X d'un même champ de vue avec une même source d'archive, soit du recouvrement de différents champs de vue XMM. On peut voir les données à charger comme deux tables (voir Fig. V.2) : la première contient les sources X et la seconde contient les sources d'archive avec leurs différents paramètres, plus l'identifiant de la source X avec laquelle elles sont corrélées, plus les paramètres de la corrélation (distance en secondes d'arc par exemple). Il faut alors transformer la deuxième table en deux tables distinctes (voir Fig. V.3) : la première ne contenant que les sources optiques uniques et leurs paramètres, la seconde consiste en une table de jointure ne contenant que les identifiants X, les identifiants optiques et les paramètres des corrélations.

Pour réaliser cette tâche (voir Fig. V.2), on calcule pour chaque source d'archive un code MD5[11] qui porte sur l'ensemble de ses paramètres, sauf les paramètres de corrélation. Deux sources identiques auront donc un même code MD5. On charge alors l'ensemble des sources d'archive d'un catalogue, plus leurs codes MD5, dans SAADA. Chaque source se voit attribuer un *oidsaada*. On trie ensuite les sources sur la base de leurs MD5, puis, pour un même MD5, sur leur *oidsaada* et on écrit le résultat du tri dans un fichier. À l'aide d'un script Perl, on remplace dans le fichier les *oidsaada* des sources ayant un même MD5 par l'*oidsaada* de la première source ayant cet MD5. Le tri permet d'optimiser le procédé en ne parcourant le fichier séquentiellement qu'une seule fois. On charge les sources du fichier, avec les *oidsaada* modifiés, dans une table temporaire. A ce

[10] `http://xcatdb.u-strasbg.fr/saada/spip.php?rubrique124`

[11] L'algorithme MD5, pour *Message Digest 5*, est une fonction de hachage cryptographique qui permet d'obtenir l'empreinte numérique d'une suite de caractères.

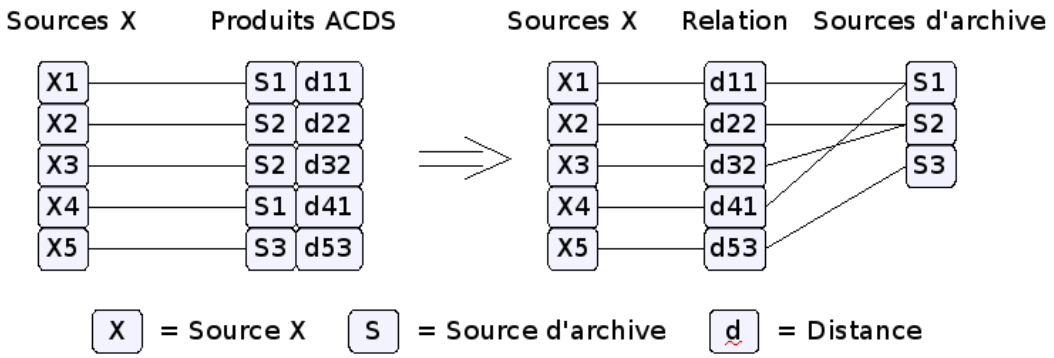


FIG. V.2 – Ce schéma montre le problème d'unicité des sources d'archive au chargement des produits de l'ACDS. On souhaite supprimer les sources d'archive qui présentent des doublons car corrélées à plusieurs sources X.

stade, les sources d'archive sont toujours multiples, mais chaque lien de la table est différent car possède des paramètres de corrélation différents. On crée alors via une requête SQL la relation à partir de la table temporaire qui contient parmi les paramètres de chaque ligne l'identifiant de la source X avec laquelle la source d'archive est corrélée. Une fois la relation créée, il faut se débarrasser des doublons dans la table où sont stockées les sources d'archive. On utilise pour cela un second script Perl qui supprime dans le fichier précédent les doublons. On ne conserve donc qu'une ligne par MD5. On charge ce fichier dans la table de la base contenant les source d'archive en écrasant les données précédentes. Ces opérations se font sur les tables de niveau « classe ». Il ne reste maintenant qu'à supprimer au niveau « collection » les doublons qui ont été supprimés au niveau « classe ». Il suffit pour cela de faire une jointure sur les *oidsaada* entre la nouvelle table de niveau classe et celle de niveau collection.

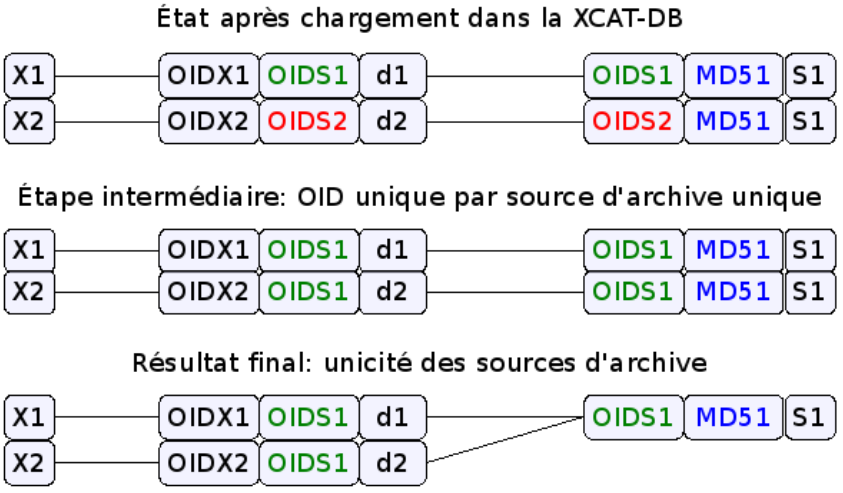


FIG. V.3 – Schéma synthétique du traitement des sources d'archive qui sont en doublons.

L'ensemble des opérations décrites ci-dessus aurait pu être réalisé en SQL uniquement. Cependant, les tests effectués ont montré que pour ces tâches les outils UNIX sont à la fois plus simples et plus performants.

2 Module de corrélation et XCAT-DB

La première partie de notre travail a concerné l'étude et la réalisation d'un algorithme de corrélations croisées fournissant des probabilités d'identification pour chaque association entre une source X et une source d'archive (voir chapitre II). L'algorithme qui y est décrit a été implémenté sous forme d'un module de corrélation indépendant écrit en langage JAVA. Ce module a été intégré aux scripts de chargement de la XCAT-DB afin d'y inclure de nouveaux catalogues et de mettre à jour les corrélations existantes réalisées par l'ACDS. On a vu dans la section précédente le problème d'unicité des sources d'archive produit par l'ACDS. Les liens calculés par l'ACDS ne comportent qu'un seul *qualifier*, la distance en secondes d'arc entre deux sources. Nous souhaitons ajouter des informations supplémentaires, notamment des probabilités d'identification et les distances en sigma. Il faut donc ajouter de nouveaux *qualifiers* aux relations. On souhaite également ajouter des catalogues d'archive qui ne sont pas traités par l'ACDS. Il est alors nécessaire d'ajouter de nouvelles sources d'archive et de nouveaux liens de corrélation dans la XCAT-DB.

On aborde ici quelques aspects généraux du module de corrélation et on décrit, sans entrer dans les détails, les interactions de ce module avec la XCAT-DB.

2.1 Module de corrélation

Le module de corrélation utilise comme support une base de données PostgreSQL (voir §2.1.2) dans laquelle il stocke ses résultats. Ses actions sont découpées en trois tâches principales (pour plus de détails, voir le §2.1.4) :

- il crée un *template* de base de données contenant certaines informations récupérées dans la base de connaissances (voir section suivante) et dans la XCAT-DB ;
- à partir d'un nom de catalogue, il se connecte à la XCAT-DB et à l'OV, il stocke dans la base les sources XMM et d'archive, il produit et stocke les liens de corrélation et enfin il calcule leurs rapports de vraisemblance ;
- il regroupe à partir de la latitude galactique les champs de vue pour calculer les fonctions de fiabilité et les probabilités d'identification.

À ce point, la XCAT-DB n'a pas été modifiée. Elle a juste été accédée pour l'extraction des sources X. Cependant, pour faciliter l'insertion de nouveaux catalogues dans la XCAT-DB et surtout pour simplifier sa mise à jour, la base de données PostgreSQL utilisée par le module de corrélation pour stocker ses résultats n'est autre que la base hébergeant la XCAT-DB. Cette incursion est totalement transparente pour SAADA et ne perturbe en aucun cas le bon fonctionnement de la XCAT-DB.

Dans la suite, on s'attache à décrire les entrées-sorties et quelques aspects particuliers du module de corrélation. On verra en premier lieu en quoi consiste la « base de connaissances ». On décrira ensuite le schéma de la base de données utilisée avant d'aborder les actions et les entrées-sorties.

On traitera uniquement des corrélations avec le catalogue XMM. Cependant, au moyen de quelques modifications mineures, le module de corrélation a été utilisé avec succès pour corréler les sources Chandra, les sources du *Slew Survey* d'XMM-Newton ainsi que les sources brillantes et faibles du ROSAT *All Sky Survey*.

2.1.1 La « base de connaissances »

a Nécessité d'une base de connaissances : limites actuelles de l'OV

Les données accessibles via l'observatoire virtuel (OV, ou VO pour *Virtual Observatory* en anglais) sont hétérogènes, l'OV n'ayant pour but que d'en uniformiser l'accès (Djorgovski & Williams 2005, §5). Mais les protocoles actuels, notamment le protocole *ConeSearch*[12], ne retourne pas d'information sémantique. Le sens d'une grandeur dans différents catalogues peut être différent. Pour illustrer ce propos, prenons l'exemple des erreurs sur les positions. En entrée de notre algorithme de corrélation, nous avons besoin des erreurs sur les positions de chaque source. Les

[12] http://www.ivoa.net/Documents/latest/ConeSearch.html

erreurs retournées par le protocole *ConeSearch* peuvent être définies à 1 σ, à 3 σ, peuvent contenir une erreur systématique ou juste correspondre à l'incertitude du positionnement due à la PSF, etc. Le format de retour, la VOTable[13], contient bien les unités, mais pas ce type d'informations.

L'utilisation des UCD[14] (Derriere *et al.* 2004) dans les VOTable permet dans une certaine mesure de catégoriser individuellement les colonnes. Mais les UCD ne permettent pas de faire de liens entre colonnes. À moins qu'un groupement entre deux colonnes ne soit fait explicitement (par exemple entre une grandeur et son erreur associée), on ne peut pas s'affranchir de possibles ambiguïtés dans les associations entre colonnes. On ne peut pas non plus restreindre la liste retournée aux sources de bonne qualité, évitant celles qui ont de fortes chances d'être de fausses détections. Il faut donc connaître le sens des différents *flags* contenus dans le catalogue interrogé et procéder soi-même à la sélection. La connaissance du catalogue auquel on accède via le VO est donc indispensable. Nous avons donc choisi de stocker les informations utiles dans une « base de connaissances ».

Jusqu'à récemment, le *registry VO*[15] contennait de nombreuses URL inactives et des liens vers des services de qualité différente. Il était donc prudent (et le reste encore en partie) de choisir avec soin le centre de données qu'on va interroger. On a par exemple voulu récupérer les URL de *ConeSearch* automatiquement à partir du seul nom de catalogue en utilisant l'API *Astroruntime*[16] (Winstanley *et al.* 2007). Une option manquait aux URL pointant vers ViZieR, ayant pour conséquence de tronquer systématiquement la liste retournée par VOTable à 999 sources. Cette expérience nous a poussé à ajouter à notre « base de connaissances » les URL OV des catalogues interrogés par *ConeSearch*.

Cette base contient au final les informations suivantes :

- la base de l'URL *ConeSearch* du catalogue à interroger ;
- le nom de la colonne contenant l'identifiant unique de la source ;
- le type d'erreur sur les positions, le nom des colonnes correspondantes et une éventuelle erreur systématique ;
- le nom des colonnes contenant les magnitudes à prendre en compte pour le calcul des rapports de vraisemblance et les erreurs correspondantes ;
- le nom des colonnes à conserver dans notre base de données locale ;
- des contraintes de suppression de sources dans un format que nous avons défini ;
- des contraintes permettant d'ignorer certaines sources pour une bande de magnitude donnée dans un format que nous avons défini.

Les magnitudes peuvent être trouvées automatiquement grâce aux UCD. Cependant, pour certains catalogues, le GSC par exemple, certaines magnitudes ne sont mesurées que pour un petit nombre de sources. évitées par l'algorithme de corrélation.

Nous espérons qu'avec les évolutions futures de l'OV la majorité des informations contenues dans cette « base de connaissances » deviendront inutiles.

b Structure des fichiers de la « base de connaissances »

En accord avec les préceptes de la philosophie UNIX (Raymond 2004), nous avons choisi de simples fichiers ASCII pour stocker les informations de cette « base de connaissances ». Les fichiers ASCII ont l'avantage d'être facilement visualisables et modifiables. Nous avons décidé de mettre l'ensemble des informations relatives à un catalogue dans un même fichier. Chaque fichier ne contient les données que d'un seul catalogue. Chaque ligne de ces fichiers commence par un mot clé indiquant le type d'information contenu sur la ligne. Le nom du fichier est constitué du nom du catalogue suivi de l'extension « txt ». La base de connaissances n'est autre que le dossier contenant l'ensemble des fichiers associés à chaque catalogue. Nous donnons ci-dessous un exemple

[13] Format de fichier XML contenant des données astronomiques tabulaires.
Voir : `http://www.ivoa.net/Documents/latest/VOT.html`

[14] Les UCD, pour *Unified Content Descriptors* sont des métadonnées utilisées en astronomie et dont le vocabulaire est contrôlé (Derriere *et al.* 2004).
Voir : `http://www.ivoa.net/Documents/latest/UCD.html`

[15] Annuaire qui recense l'ensemble des ressources OV disponibles.
Voir : `http://www.ivoa.net/Documents/cover/RegistryInterface-20061107.html`

[16] `http://www.astrogrid.org/wiki/Help/AstroRuntime`
`http://www.ivoa.net/internal/IVOA/InterOpMay2006Applications/TheAstroRuntime.pdf`

de fichier de la « base de connaissances » correspondant au catalogue 2MASS. Comme en *shell script*, et selon le principe de moindre surprise cher au monde UNIX, les lignes de commentaires commencent par un #.

```
#--------------------
# Base URL for ConeSearch
#--------------------
cs_url: http://newviz.u-strasbg.fr/viz-bin/votable/-dtd/-A?-source=2MASS&-out.max=999999
#--------------------
# Unique source identifier (<=> UCD1+ ID_MAIN )
#--------------------
src_id_col: 2MASS
#--------------------
# pos_err_info: type ColSigx ColSigy ColTheta ToArcsec ToDeg To1Sigma SysErrSigx SysErrS
#--------------------
pos_err_info: 3 errMaj  errMin   errPA     1.0     1.0    1.0      0.0        0.0
#--------------------
# Mag cols name to correlate with: ( colMag  [colErrMag | null ]  ToSigma SysErr )
#--------------------
col_mag: Jmag  e_Jmag  1.0  0.0
col_mag: Hmag  e_Hmag  1.0  0.0
col_mag: Kmag  e_Kmag  1.0  0.0
#--------------------
# Added Cols (Ra,Dec) found automaticaly
#--------------------
add_cols: prox | Cntr | JD
#----------------------
# delete_constraint: colName op val1 val2
#----------------------
# Vizier: Aflg = 1 - Source is associated with the predicted position of a known solar s
delete_constraint: Aflg = 1 null
#---------------------------------------
# ignore_constraint: colName op val1 val2
#---------------------------------------
# Vizier: Rflg = 0 - Source is not detected in this band
ignore_constraint: Jmag | Rflg   MATCHES  0\w\w
ignore_constraint: Hmag | Rflg   MATCHES  \w0\w
ignore_constraint: Kmag | Rflg   MATCHES  \w\w0
```

2.1.2 Structure de la base de données

La base de données que nous utilisons pour stocker les résultats de corrélation (Fig. V.4) est composée de neuf tables auxquelles s'ajoutent deux tables par catalogue corrélé. Ces tables sont les suivantes (dans ce qui suit, *FK* signifie *Foreign Key*, ou clef étrangére » en français) :

- *catalogue* (cat_name), qui contient uniquement les noms des catalogues présents dans la base de connaissances ;
- *magnitude* (cat_name$_{FK}$, mag_name), qui contient pour chaque catalogue les noms des magnitudes qui peuvent être utilisées pour la corrélation ;
- *cat_meta* (cat_name$_{FK}$, num, name, data_type, ucd, ucd1 , ucd1p, unit), qui stocke les métadonnées des colonnes stockées de chaque catalogue ;
- *density* (obs_id$_{FK}$, cat_name$_{FK}$, density), qui pour chaque champ de vue de chaque catalogue stocke la densité moyenne de sources dans le champ de vue ;
- *fov* (obs_id, ra, dec, l, b, syserr, flux_min, nbrSrcX, epoch), qui contient un certain nombre de paramètres associés à chaque champ de vue XMM, l'erreur systématique par exemple ;
- *histo_lr* (obs_id$_{FK}$, cat_name$_{FK}$, mag_name$_{FK}$, log_lr, count_lr, count_noise), qui contient

pour chaque champ de vue l'histogramme des LR des associations et des fausses associations ;
- *group_obsid* (group_name, obs_id_{FK}), qui contient les champs de vue associés à chaque groupe d'observation ;
- *group_histo* (group_name, cat_name_{FK}, mag_name_{FK}, log_lr, count_lr, count_noise, reliability, reliability_fit), qui stocke pour chaque groupe d'observations les histogrammes des LR des associations et des fausses associations ;
- *lognlogs* (obs_id_{FK}, cat_name_{FK}, mag_name_{FK}, mag, count), qui stocke pour chaque champ de vue la loi logN-logS du catalogue de contreparties.

Quand la base de données qui stocke les résultats n'est pas la XCAT-DB, il faut ajouter une table qui contient les sources XMM ayant une contrepartie dans un des catalogues. Il faut également ajouter une table par catalogue contenant les sources du catalogue d'archive associées à au moins une source XMM ainsi qu'une table par catalogue qui contient les associations entre sources XMM et sources uniques du catalogue d'archive.

Si par exemple, le base contient les corrélations avec la DR7 du SDSS, la base contient également les tables :
xmm (obs_id_{FK}, oid_x, ra, dec, ...)
dr7 (obs_id_{FK}, cp_name, ra, dec, ...)
join_xmm_dr7 (obs_id_{FK}, oid_x_{FK}, cp_name_{FK}, proba_id, lr_mag_i, r_mag_i, ... , dist_arcsec, dist_1sigma, lr_max, variability, dx, dy)

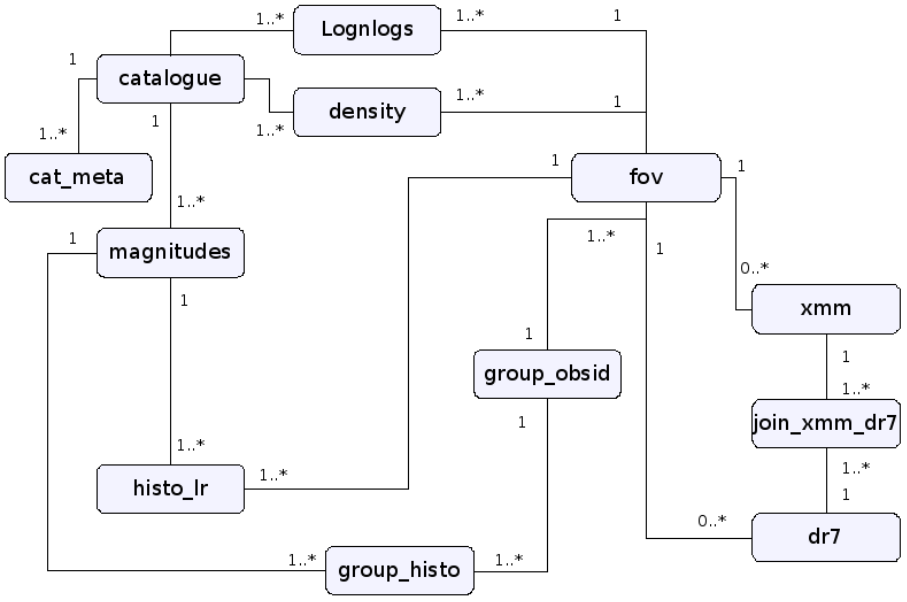


FIG. V.4 – Schéma entités-relations de la base de données du module de corrélation. Les chiffres représentent la cardinalité des tuples.

2.1.3 Actions et entrées-sorties

Le module de corrélation comporte une séquence de six tâches appelées via un script *ant*[17]. Ces actions sont :
- la création de la structure relationnelle de la base et l'insertion de données dans les tables *catalogue*, *magnitude*, *fov* et *group_obsid* (tâche *tables.create*) ;
- la suppression du cache de données qui contient les sources retournées par les requêtes *ConeSearch* après une première corrélation (tâche *cachexmm.clean*) ;

[17] *ant* est un outil de type *Make* basé sur Java et dont les instructions se trouvent dans un fichier au format XML. Pour plus d'informations, voir le site `http://ant.apache.org/` .

– la corrélation avec un catalogue (tâche *correlate*) ;
– le calcul des fiabilités qui est fait dans l'étape précédente mais qui peut également être refait indépendamment (tâche *reliability.only*) ;
– la création d'une VOTable contenant l'ensemble des données stockées d'un catalogue et pouvant être chargée par SAADA (tâche *correlated.out.votable*) ;
– la suppression de toutes les tables de la base (tâche *tables.delete*).

Lors de la création de la structure de la base, le module de corrélation a besoin en entrée :
– des paramètres de connexion à la base vierge ;
– du chemin d'accès du dossier contenant la base de connaissances ;
– des paramètres de connexion à la XCAT-DB pour remplir la table *fov*.

Cette action n'a aucune sortie, les données étant écrites dans la base. La corrélation demande en entrée :
– le nom du catalogue que l'on souhaite corréler ;
– les paramètres de connexion à la base ;
– le chemin d'accès du dossier contenant la base de connaissances ;
– les paramètres de connexion à la XCAT-DB pour récupérer les sources XMM.

Pendant l'exécution de cette tâche, le logiciel accède au cache de données. Si celui-ci ne contient pas les informations utiles, il se connecte au VO pour récupérer les sources d'archive. Hormis l'écriture des résultats dans la base de données et, éventuellement, dans le cache, cette action ne possède pas d'autres sorties.
Toutes les interactions avec les bases de données (base locale et XCAT-DB) se font grâce à des connexions JDBC[18].

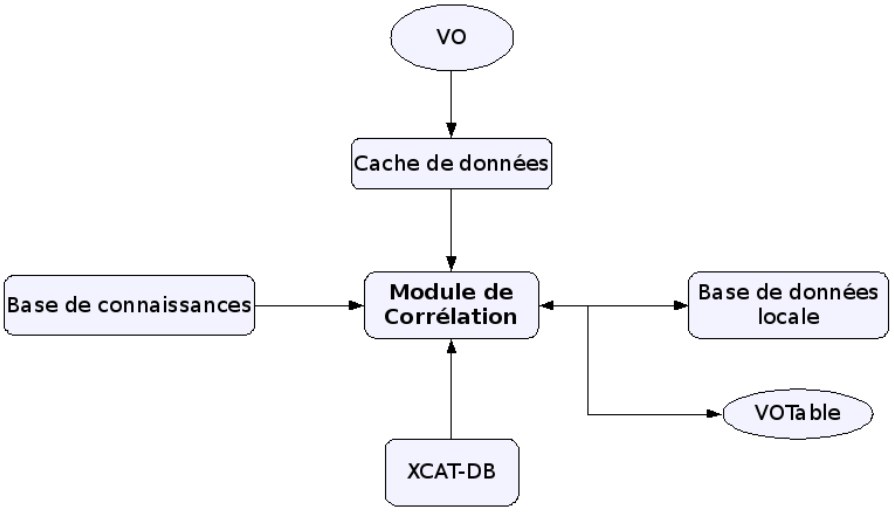


FIG. V.5 – Schéma représentant, par le sens des flèches, les entrées-sorties du module de corrélation.

2.1.4 Séquence des opérations de corrélation

Sans entrer dans les détails de programmation et les astuces permettant d'économiser du temps, voici le déroulement global des opérations de corrélation :

1. interrogation de la base pour récupérer la liste de FOV avec les coordonnées des centres et les numéros *obs_id* ;
2. pour chaque FOV :

[18] JDBC, pour *Java DataBase Connectivity* est une API fournie avec Java depuis sa version 1.1 permettant de se connecter à des bases de données. Documentation :
`http://java.sun.com/javase/6/docs/technotes/guides/jdbc/`

   (a) chargement des sources XMM du FOV via la XCAT-DB ;
   (b) chargement des données d'archive via le cache ou le VO ;
   (c) création de la jointure pour ce FOV :
      i. recherche des candidats ;
      ii. calcul des densités locales pour chaque candidat et chaque magnitude ;
      iii. calcul des LR ;
      iv. création de l'histogramme des LR des candidats pour le FOV ;
      v. calcul de la loi logN-logS ;
      vi. calcul de l'histogramme estimé des fausses associations ;
      vii. sauvegarde dans la base.
3. pour chaque groupe d'observations créé en fonction de la latitude galactique des FOV :
   (a) sommation et récupération des histogrammes de LR par une requête SQL ;
   (b) sauvegarde dans la base ;
   (c) ajustement de la courbe de fiabilité pour chaque magnitude ;
   (d) calcul des fiabilités et des probabilités d'identification (voir chapitre II) ;
   (e) mise à jour des relations dans la base.

2.2 Interactions entre module de corrélation et XCAT-DB

On a vu dans la section précédente que le module de corrélation interroge la XCAT-DB pour récupérer des informations sur les champs de vue et les sources du 2XMMi. Cette section traite de l'intégration des résultats de la corrélation à la XCAT-DB. Nous appellerons « système de corrélation » le système comprenant le module de corrélation et effectuant l'ajout de données de plusieurs catalogues dans la XCAT-DB. Ce système est destiné à être mis en place par seulement quelques personnes et à être inclus dans le chargement de la XCAT-DB. On a donc choisi pour ce système un simple script *bash* contenant des appels aux tâches *ant* de fichiers *build.xml* associés à Saada et au module de corrélation. Le module de corrélation prend la forme d'un package Java nommé *crosscorr*.

Les principales tâches que doit remplir notre système de corrélation sont :

- le calcul de liens de corrélations entre les sources EPIC de la XCAT-DB et des catalogues d'archive, indépendamment de ceux hébergés dans la XCAT-DB. C'est là le rôle du module de corrélation décrit précédemment. Ces liens sont donc crées et modifiés dans la base de données dédiée au module de corrélation.
- la mise à jour des relations de la XCAT-DB pour les catalogues présents dans l'ACDS à partir des liens créés par le module de corrélation : mise à jour des relations (EpicSrcToArchSrc, ArchSrcToEpicSrc, etc) avec les données issues de la corrélation (distances en sigma, probabilités d'identification, densités locales, LR, etc).
- l'ajout de nouveaux catalogues à partir des données stockées dans la base dédiée au module de corrélation : il faut ajouter les catalogues absents de l'ACDS, la DR7 du SDSS par exemple, à la collection *ARCH_CAT* de la XCAT-DB et ajouter les nouveaux liens de corrélation aux relations.
- permettre une restauration de la XCAT-DB : ce qui implique la mise en place d'un mécanisme permettant de restaurer l'état de la base avant chargement des nouveaux catalogues et avant la mise à jour des relations.
- vérifier la cohérence entre les données ACDS et les résultats du module de corrélation, ce qui implique :
  - la mise en place de mécanismes qui détectent la présence de liens issus du corrélateur et qui ne figurent pas dans les résultats de l'ACDS, et inversement ;
  - de vérifier que les distances sont compatibles ;

– de vérifier que toutes les sources des relations sont présentes dans les données d'archives et inversement.

Le système de corrélation que nous avons mis en place permet de faire remonter de façon claire les éventuelles erreurs.

2.2.1 Calcul des corrélations

a Initialisation

Le module de corrélation est délivré dans un package Java nommé *crosscorr*. Ce package contient les classes mentionnées ci-après.
L'appel à la classe *LocalDB* avec les options *create* ou *recreate* permet de générer toutes les tables nécessaires au fonctionnement du module. Ces tables se distinguent des tables Saada par un préfixe commun. Elles sont directement créées par des requêtes SQL, sans utiliser Saada. Elles ne sont donc pas accessibles via l'API de Saada, Saada ne les voit pas.
L'option *delete* permet d'effacer toutes les tables SQL générées par le module de corrélation.

b Corrélation d'un catalogue

La classe *CorrelationModule* appelée avec un nom d'un catalogue en paramètre réalise la corrélation, champ de vue par champ de vue, de toutes les sources EPIC présentes dans la XCAT-DB.

La table de jointure reliant les *oidsaada*[19] aux noms des sources d'archive, contenant également tous les paramètres de corrélation tel que les distances ou les probabilités d'identification, est sauvegardée localement dans une table SQL. La table contenant toutes les sources du catalogue de contreparties est également sauvegardée. Ces tables sont créées par des requêtes SQL, indépendamment de Saada.

c Vérification de la cohérence des corrélations avec l'ACDS

On vérifie plusieurs points de cohérence avant la mise à jour des relations.

– on vérifie la cohérence entre les distances calculées par le module de corrélation et celles fournies par l'ACDS : on fait une jointure interne entre nos tables de jointure et les relations de l'ACDS sur les OIDs primaires et secondaires ; on s'assure qu'aucune ligne ne comporte un écart de distance supérieur à 0.05 arcsec en valeur absolue ;
– on vérifie que toutes les associations entre les sources XMM et les sources d'archive issues du module de corrélation sont également présentes dans les relations de la XCAT-DB. On utilise pour ce faire l'opérateur SQL EXCEPT (voir exemple ci-dessous) ;
– on vérifie que tous les identifiants de conteparties des tables de jointure issues du module de corrélation se trouvent également dans les tables d'archive de la XCAT-DB. On utilise là aussi le mot clef EXCEPT. Exemple pour le catalogue 2MASS :

```
SELECT count(*) as nbr
FROM (
    SELECT cp_name FROM fx_join_xmm_TWOMASS
    EXCEPT
    SELECT _2mass FROM arch_2246aentry
) q;
```

On s'assure que la valeur *nbr* retournée vaut 0. Dans le cas contraire, on génère une erreur.

2.2.2 Mise à jour des relations

On travaille principalement sur la relation *EpicSrcToArchSrc*. Elle relie les sources XMM détectées par la caméra EPIC aux sources d'archive. Les autres relations seront recalculées à partir de celle-ci.

Avant la première mise à jour, on sauvegarde toutes les relations sujettes à modification : on les copie dans des tables que l'on nomme *relation_name_org* où *relation_name* est le nom de la relation

[19] Un *oidsaada* est un numéro unique attribué à chaque enregistrement dans une base de données génerée par Saada (une Saada-DB). Autrement dit, l'*oidsaada* est une clef primaire dans les tables de niveau Colletion et à la fois une clef primaire et étrangère (*foreign key*) dans les tables de niveau Classe.

et *org* sert de diminutif pour *origine*.
L'appel à la fonction *restor* rétablit les relations d'origine si nécessaire.

a Mise à jour de EpicSrcToArchSrc

La fonction de mise à jour, *xcatdb_relations_update*, possède comme paramètre d'entrée la liste des catalogues dont les liens vont être mis à jour. Cette mise à jour se fait par des requêtes SQL selon les étapes suivantes :

- pour chaque catalogue on génère une requête SQL dont le format de sortie est le même que le schéma des relations : *(oidprimaire,oidsecondaire,distance,...)*. On effectue pour cela une jointure interne entre la table de jointure du catalogue concerné issue du module de corrélation et la table d'archive correspondante dans la XCAT-DB.
- on assemble ces requêtes via des UNIONs SQL ;
- on fait une jointure externe gauche de la relation à mettre à jour avec la requête obtenue ; on sauvegarde le résultat dans une table temporaire ;
- on supprime la relation ;
- on renomme la table temporaire par le nom de la relation.

b Mise à jour des autres relations

Les relations *ArchSrcToEpicSrc*, *ArchSrcToCatSrc* et *CatSrcToArchSrc* sont mises à jour à partir de *EpicSrcToArchSrc*.

ArchSrcToEpicSrc est une copie de *EpicSrcToArchSrc* pour laquelle les colonnes *oidprimary* et *oidsecondary* sont interchangées :

```
DELETE FROM ArchSrcToEpicSrc;
INSERT INTO ArchSrcToEpicSrc
    SELECT j.oidsecondary  as oidprimary   ,
           j.oidprimary    as oidsecondary ,
           j.epic_cat_dist as epic_cat_dist,
           ...
    FROM EpicSrcToArchSrc j;
```

La relation *ArchSrcToCatSrc* s'obtient par une double jointure : on a besoin des liens entre *ArchSrcToEpicSrc* et *EpicSrcToCatSrc* ; on doit également récupérer des attributs de la table *Catalogue_Entry* pour remplir certains qualifiers de notre relation. On met à jour la relation par des requêtes SQL comme suit :

```
DELETE FROM ArchSrcToCatSrc;
INSERT INTO ArchSrcToCatSrc
    SELECT ae.oidprimary      as oidprimary,
           c.oidsaada         as oidsecondary,
           ae.epic_cat_dist   as epic_cat_dist,
           ae.epic_cat_sdist  as epic_cat_sdist,
           c.xcat_obsid::real as obs_id,
           c.xcat_srcnum      as cat_src_num,
           ...
    FROM ArchSrcToEpicSrc ae
    INNER JOIN EpicSrcToCatSrc ec
    ON ae.oidsecondary=ec.oidprimary
    INNER JOIN Catalogue_Entry c
    ON ec.oidsecondary = c.oidsaada;
```

La relation *CatSrcToArchSrc* est mise à jour à partir de *ArchSrcToCatSrc*. On utilise le même type de requêtes que pour calculer *ArchSrcToEpicSrc* à partir de *EpicSrcToArchSrc*.

2.2.3 Insertion de nouveaux catalogues

a Chargement du nouveau catalogue

L'appel à la classe Java *CorrelationModule* avec l'option *-vot outfilename* permet, après corrélation complète d'un catalogue, de sauvegarder les sources de contreparties dans un fichier VOTable.

On charge ce fichier dans Saada via une tâche *ant*. Cette tâche est contenue dans un fichier XML créé spécifiquement pour un catalogue donné. Elle fait appel au *dataloader* de Saada.

b Ajout des liens aux relations

L'ajout de liens se fait par une requête SQL. On fait un INSERT dans la relation *EpicSrcToArchSrc* du résultat de la jointure entre la table nouvellement chargée et la table de jointure issue du module de corrélation.

Après cette opération, il faut également mettre à jour les autres relations, comme décrit dans le paragraphe 2.2.2.b.

2.2.4 Gestion des erreurs

Pour faciliter le déboguage et assurer la fiabilité des résultats, il est important de détecter et de repérer toutes les erreurs pouvant apparaître pendant les processus de corrélation et d'intégration des résultats à la XCAT-DB. On peut rencontrer des erreurs dans les trois niveaux d'exécution suivants :

- dans le code *bash* ;
- dans le script *ant* ;
- dans le script *Java*.

Il est donc important que chaque niveau traite proprement les erreurs, et le cas échéant, les remonte au niveau supérieur.
On va voir ici comment sont gérées ces erreurs dans les scripts *ant* et *bash*.

Le script *ant* détecte les erreurs d'exécution des classes *Java* qu'il appelle. Ce comportement est obtenu grâce à l'option *failonerror* mis à *true* dans les tâches de type *Java*. Exemple d'une tâche *ant* de type *Java* utilisant cette option :

```
<target name="correlate">
    <java fork="true" classname="correlation.CorrelationModule"
                      failonerror="true">
        <classpath refid="libs"/>
        <arg value="${cat_name}"/>
        <jvmarg value="-Xmx1280M"/>
    </java>
</target>
```

Le statut de retour de notre script *ant*, en cas d'erreur, est alors différent de zéro. Cet état d'exécution est récupéré dans le script *bash* appelant. S'il contient une valeur non nulle, le script *bash* génère lui aussi une erreur. Cette erreur affiche un message à l'écran, affiche les n dernières lignes du fichier de log du script *bash* et quitte avec un statut différent de zéro.
Exemple de la gestion d'une erreur d'une classe *Java* appelée via *ant* :

```
ant correlate -Dcat_name=$1  >> ${LOG_FILE} 2>&1
[[ $? != 0 ]] && { error "error correlating the cat \"$1\"" ${LOG_FILE}; }
```

Cet exemple fait appel à la fonction *bash* « error », ici $n = 25$:

```
# params: 1:msg (the error msg)
#         2 (optional):log file (the log file complete path)
error(){
```

```
    echo "$(date) ERROR: $1"
    [[ -f $2 ]] && { echo "tail -25 $2"; tail -25 $2; }
    exit 1;
}
```

Dans le script *bash*, tous les statuts de retour d'exécution de commandes sont vérifiés. Voici un autre exemple avec la fonction chargée de l'exécution des requêtes SQL :

```
# params: 1:query (the query to exec)
#         2 (optionel): log file (the log file complete path)
exec_query(){
    local LOG=${2:-${LOG_DIR}/sql.queries.log}
    psql -h ${DB_HOST} -U ${DB_USER} ${DB_NAME} -c "$1" >> ${LOG} 2>&1
    [[ $? != 0 ]] && { error "Query failed! \"$1\"" ${LOG}; }
}
```

2.2.5 Schéma d'ensemble

La figure V.6 consiste en un schéma récapitulatif. Elle présente une vue d'ensemble du processus d'insertion des corrélations dans la XCAT-DB.
On y retrouve les différents éléments traités dans les paragraphes précédents :
- l'initialisation ;
- le calcul des corrélations par le module de corrélation ;
- la vérification de la cohérence des résultats par rapport à l'ACDS ;
- la mise à jour des relations ;
- l'insertion dans la XCAT-DB de nouveaux catalogues.

3 Classification et XCAT-DB

Le système de classification des sources XMM est constitué d'un script *bash* qui enchaîne les différentes tâches avec pour entrées-sorties des fichiers ASCII. Les données d'entrées sont récupérées dans la XCAT-DB via une requête SQL dont la sortie est formatée par une fonction *bash*. Le script comporte de multiples options permettant l'utilisation des différents algorithmes de classification vus dans le chapitre III. Ces algorithmes peuvent être écrits en C ou en JAVA quand nous les avons nous-mêmes implémenté. On peut également choisir l'utilisation d'une ACP. Dans ce cas, on peut choisir le nombre de composantes principales à conserver en entrée de l'algorithme de classification.

Il nous faut une stratégie pour stocker la classification des sources dans la XCAT-DB. Nous avons décidé de stocker plusieurs résultats en fonction des algorithmes utilisés. Ces résultats sont stockés sous forme d'une relation entre les sources du catalogue et une collection contenant les types d'objets. Chaque type d'objet est inséré dans la base sous forme d'un *flatfile* contenant la description de la classe qu'il représente. Les *qualifiers* de cette relation contiennent les probabilités et permettent de différencier les algorithmes utilisés. Il nous a fallu créer cette relation et la collection contenant les types d'objets par l'intermédiaire de Saada. Voici les actions nécessaires au stockage des probabilités issues de la classification des sources :
- création d'une collection qui contient les différents types d'objets :
  - nommé OBJECT_TYPE ;
  - seule la catégorie FLATFILE est utilisée ;
  - des fichiers décrivant les différents types d'objets y sont chargés.
- création des liens qui relient les sources EPIC et les sources CATALOGUE avec les entrées de la collection OBJECT_TYPE, et inversement :
  - création d'une relation entre les collections EPIC et OBJECT_TYPE que l'on nomme *EpicSrcToObjType* ;

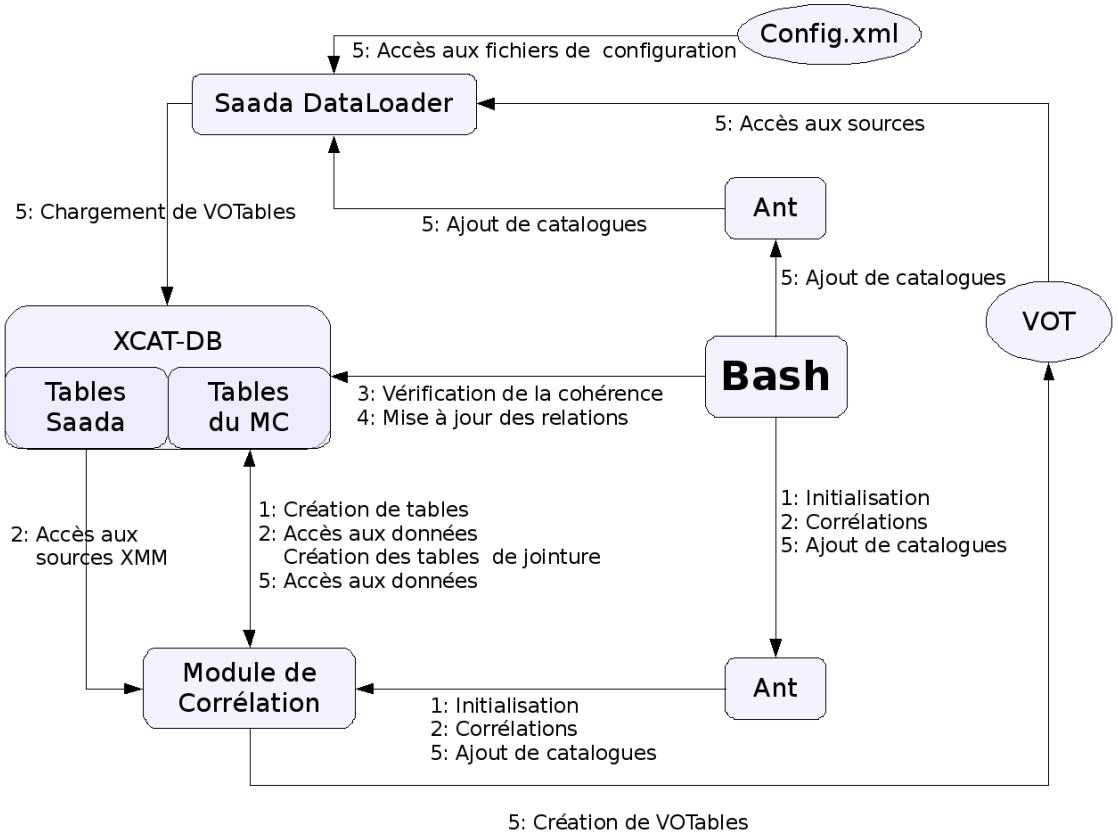


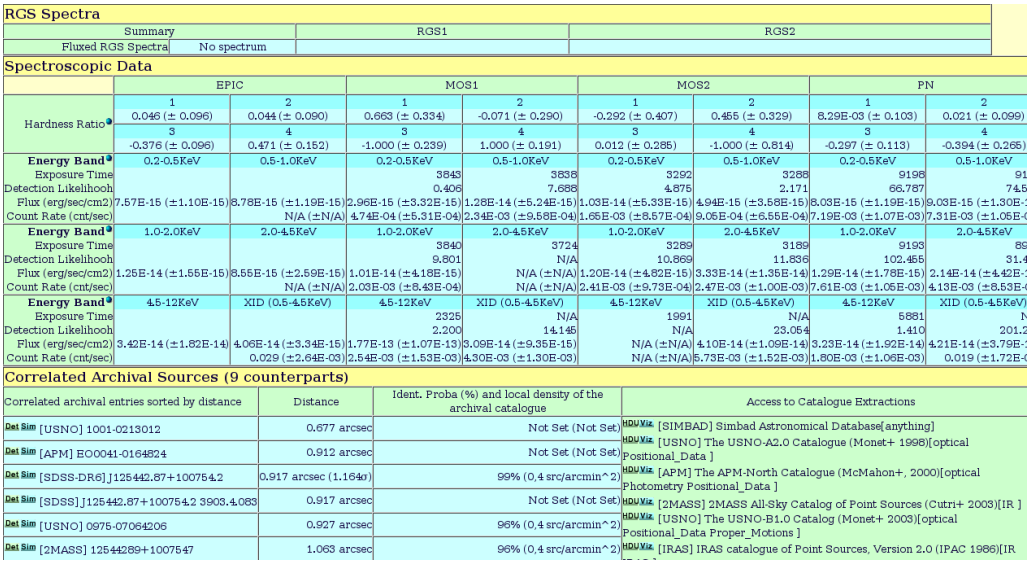
FIG. V.6 – Processus d'insertion des corrélations dans la XCAT-DB. Les numéros correspondent aux tâches du paragraphe 2.1.3.

  - création d'une relation entre les collections OBJECT_TYPE et EPIC que l'on nomme *ObjTypeToEpicSrc* ;
  - création d'une relation entre les collections CATALOGUE et OBJECT_TYPE que l'on nomme *CatSrcToObjType* ;
  - création d'une relation entre les collections OBJECT_TYPE et CATALOGUE que l'on nomme *ObjTypeToCatSrc*.
- mise en place du système de remplissage et de mise à jour des liens :
  - remplissage via des requêtes SQL de la relation *EpicSrcToObjType* avec les probabilités calculées, à partir des sorties des algorithmes de classification ;
  - mise à jour des autres relations à partir de *EpicSrcToObjType*.

4 Accès aux données hebergé dans la XCAT-DB

Les données que nous avons insérés dans la XCAT-DB sont accèssible via l'interface Web de la base. Lorsque l'on consulte la page qui contient les détails sur une source X, les corrélations de cette source avec les différents catalogues d'archives apparaîssent en bas, comme on peut le voir Fig. V.7. Les probabilités d'identifications que nous avons calculées,la distance en sigma et la densité locale de source sont affichées dans les paramètres de corrélation.

La XCAT-DB comporte des formulaires qui aident l'utilisateur à créer une requête SAADA-QL. En utilisant ces formulaires, ou directement une requête SAADA-QL, on peut par exemple contraindre la recherche de sources XMM ayant une contrepartie dans le SDSS DR7 avec une probabilité d'identification supérieure à un seuil donné. Le formulaire traduira automatiquement



RGS Spectra

Summary	RGS1	RGS2
Fluxed RGS Spectra: No spectrum		

Spectroscopic Data

	EPIC		MOS1		MOS2		PN	
Hardness Ratio	1: 0.046 (± 0.096)	2: 0.044 (± 0.090)	1: 0.663 (± 0.334)	2: -0.071 (± 0.290)	1: -0.292 (± 0.407)	2: 0.455 (± 0.329)	1: 8.29E-03 (± 0.103)	2: 0.021 (± 0.099)
	3: -0.376 (± 0.096)	4: 0.471 (± 0.152)	3: -1.000 (± 0.239)	4: 1.000 (± 0.191)	3: 0.012 (± 0.285)	4: -1.000 (± 0.814)	3: -0.297 (± 0.113)	4: -0.394 (± 0.265)
Energy Band	0.2-0.5KeV	0.5-1.0KeV	0.2-0.5KeV	0.5-1.0KeV	0.2-0.5KeV	0.5-1.0KeV	0.2-0.5KeV	0.5-1.0KeV
Exposure Time			3843	3838	3292	3288	9198	91
Detection Likelihood			0.406	7.688	4.875	2.171	66.787	74.5
Flux (erg/sec/cm2)	7.57E-15 (±1.10E-15)	8.78E-15 (±1.19E-15)	2.96E-15 (±3.32E-15)	1.28E-14 (±5.24E-15)	1.03E-14 (±5.33E-15)	4.94E-15 (±3.58E-15)	8.03E-15 (±1.19E-15)	9.03E-15 (±1.30E-
Count Rate (cnt/sec)		N/A (±N/A)	4.74E-04 (±5.31E-04)	2.34E-03 (±9.58E-04)	1.65E-03 (±8.57E-04)	9.05E-04 (±6.55E-04)	7.19E-03 (±1.07E-03)	7.31E-03 (±1.05E-
Energy Band	1.0-2.0KeV	2.0-4.5KeV	1.0-2.0KeV	2.0-4.5KeV	1.0-2.0KeV	2.0-4.5KeV	1.0-2.0KeV	2.0-4.5KeV
Exposure Time			3840	3724	3289	3189	9193	89
Detection Likelihood			9.801	N/A	10.869	11.836	102.455	31.4
Flux (erg/sec/cm2)	1.25E-14 (±1.55E-15)	8.55E-15 (±2.59E-15)	1.01E-14 (±4.18E-15)	N/A (±N/A)	1.20E-14 (±4.82E-15)	3.33E-14 (±1.35E-14)	1.29E-14 (±1.78E-15)	2.14E-14 (±4.42E-
Count Rate (cnt/sec)		N/A (±N/A)	2.03E-03 (±8.43E-04)	N/A (±N/A)	2.41E-03 (±9.73E-04)	2.47E-03 (±1.00E-03)	7.61E-03 (±1.05E-03)	4.13E-03 (±8.53E-
Energy Band	4.5-12KeV	XID (0.5-4.5KeV)	4.5-12KeV	XID (0.5-4.5KeV)	4.5-12KeV	XID (0.5-4.5KeV)	4.5-12KeV	XID (0.5-4.5KeV)
Exposure Time			2325	N/A	1991	N/A	5881	N
Detection Likelihood			2.200	14.145	N/A	23.054	1.410	201.2
Flux (erg/sec/cm2)	3.42E-14 (±1.82E-14)	4.06E-14 (±3.34E-15)	1.77E-13 (±1.07E-13)	3.09E-14 (±9.35E-15)	N/A (±N/A)	4.10E-14 (±1.09E-14)	3.23E-14 (±1.92E-14)	4.21E-14 (±3.79E-
Count Rate (cnt/sec)		0.029 (±2.64E-03)	2.54E-03 (±1.53E-03)	4.30E-03 (±1.30E-03)	N/A (±N/A)	5.73E-03 (±1.52E-03)	1.80E-03 (±1.06E-03)	0.019 (±1.72E-

Correlated Archival Sources (9 counterparts)

Correlated archival entries sorted by distance	Distance	Ident. Proba (%) and local density of the archival catalogue	Access to Catalogue Extractions
Det Sim [USNO] 1001-0213012	0.677 arcsec	Not Set (Not Set)	HDU Viz. [SIMBAD] Simbad Astronomical Database[anything]
Det Sim [APM] EO0041-0164824	0.912 arcsec	Not Set (Not Set)	HDU Viz. [USNO] The USNO-A2.0 Catalogue (Monet+ 1998)[optical Positional_Data]
Det Sim [SDSS-DR6] J125442.87+100754.2	0.917 arcsec (1.164σ)	99% (0.4 src/arcmin^2)	HDU Viz. [APM] The APM-North Catalogue (McMahon+, 2000)[optical Photometry Positional_Data]
Det Sim [SDSS] J125442.87+100754.2 3903.4.083	0.917 arcsec	Not Set (Not Set)	HDU Viz. [2MASS] 2MASS All-Sky Catalog of Point Sources (Cutri+ 2003)[IR]
Det Sim [USNO] 0975-07064206	0.927 arcsec	96% (0.4 src/arcmin^2)	HDU Viz. [USNO] The USNO-B1.0 Catalog (Monet+ 2003)[optical Positional_Data Proper_Motions]
Det Sim [2MASS] 12544289+1007547	1.063 arcsec	96% (0.4 src/arcmin^2)	HDU Viz. [IRAS] IRAS catalogue of Point Sources, Version 2.0 (IPAC 1986)[IR

FIG. V.7 – Interface de la XCAT-DB qui contient les probabilités d'identification.

la demande de l'utilisateur en construisant une requête SAADA-QL avec la syntaxe *matchPattern*. La requête SAADA-QL suivante sélectionnera par exemple toutes les sources X à moins de 5 minutes d'arc de la galaxie Mrk 91 ayant une contrepartie dans le SDSS à moins de 3 seconde d'arc avec une probabilité supérieure à 90% :

```
Select ENTRY From CatalogueEntry In CATALOGUE
WherePosition{
  isInCircle{"Mk 91",8.0,FK5,J2000}
}
WhereRelation{
  mathcPattern{
    ArchSrcToCatSrc,
    AssObjClass{ arch_SDSS_DR7 },
    Qualifier{ dist_arcsec < 3 },
    Qualifier{ proba_id > 0.9 }
  }
}
```

On peut de même effectuer des requêtes avec des contraintes sur les relations de classification et les *qualifiers* associés. Si par exemple on cherche les toutes les sources X XMM associées à une sources SDSS DR7 avec une probabilité d'identification supérieure à 99% et classée comme QSO par la KDC sur l'échantillon XS avec une probabilité supérieure à 90%, la requête contruite par le formulaire ressemblera à la requête suivante :

```
Select ENTRY From catalogueEntry In CATALOGUE
WhereRelation{
  mathcPattern{
    ArchSrcToCatSrc,
    AssObjClass{ SDSS DR7 },
    Qualifier{ proba_id > 0.99 }
  }
  mathcPattern{
    ClassifXmmDR7,
    AssObjClass{ SDSS DR7 },
    AssObjAttrSaada{ classe = QSO }
    Qualifier{ proba_qso > 0.9 }
```

```
    }
  }
```

Un formulaire simplifié spécialement dédié aux requêtes avec contraintes sur les corrélations et classifications sera développé dans un futur proche.
Les données sont uniquement consultables en ligne.



Références

DERRIERE, S., GRAY, N., MCDOWELL, J. C., MANN, R., OCHSENBEIN, F., OSUNA, P., PREITE MARTINEZ, A., RIXON, G. & WILLIAMS, R., 2004 ; *UCD in the IVOA context*. Dans *Astronomical Data Analysis Software and Systems (ADASS) XIII*, F. Ochsenbein, M. G. Allen & D. Egret, réds., tm. 314 de *Astronomical Society of the Pacific Conference Series*

DJORGOVSKI, S. G. & WILLIAMS, R., 2005 ; *Virtual Observatory : From Concept to Implementation*. Dans *Astronomical Society of the Pacific Conference Series*, N. Kassim, M. Perez, W. Junor & P. Henning, réds., tm. 345 de *Astronomical Society of the Pacific Conference Series*

MICHEL, L., HERENT, O., MOTCH, C., PYE, J. & WATSON, M. G., 2004 ; *XCAT-DB a Public Interface for the SSC XMM-Newton Catalogue*. Dans *Astronomical Data Analysis Software and Systems (ADASS) XIII*, F. Ochsenbein, M. G. Allen & D. Egret, réds., tm. 314 de *Astronomical Society of the Pacific Conference Series*

MICHEL, L., MILLAN, P., MOTCH, C. & PINEAU, F. X., 2007 ; *Building an Astronomical Database with Saada*. Dans *Astronomical Data Analysis Software and Systems XVI*, R. A. Shaw, F. Hill & D. J. Bell, réds., tm. 376 de *Astronomical Society of the Pacific Conference Series*

MICHEL, L., MOTCH, C. & PINEAU, F. X., 2008 ; *Building an Astronomical Database with Saada.* Dans *Astronomical Data Analysis Software and Systems XVII*, R. W. Argyle, P. S. Bunclark & J. R. Lewis, réds., tm. 394 de *Astronomical Society of the Pacific Conference Series*

MICHEL, L., NGUYEN, H. N. & MOTCH, C., 2005 ; *SAADA : Astronomical Databases Made Easier*. Dans *Astronomical Data Analysis Software and Systems XIV*, P. Shopbell, M. Britton & R. Ebert, réds., tm. 347 de *Astronomical Society of the Pacific Conference Series*

—, 2006 ; *How to Publish Local Data Into the VO with Saada*. Dans *Astronomical Data Analysis Software and Systems XV*, C. Gabriel, C. Arviset, D. Ponz & S. Enrique, réds., tm. 351 de *Astronomical Society of the Pacific Conference Series*

MOTCH, C., MICHEL, L., PINEAU, F.-X. & GONÇALVES, A., 2009 ; *The X-cat-DB : A multi-wavelength view on the 2XMMi catalogue*. Dans *Astronomical Data Analysis Software and Systems (ADASS) XVIII*, Astronomical Society of the Pacific Conference Series

NGOC NGUYEN, H., MICHEL, L. & MOTCH, C., 2005 ; *Saada : A Generator of Astronomical Database*. Dans *SF2A-2005 : Semaine de l'Astrophysique Francaise*, F. Casoli, T. Contini, J. M. Hameury & L. Pagani, réds.

NGUYEN, H. N., MICHEL, L. & MOTCH, C., 2006 ; *Building an Astronomical Database with Saada.* Dans *Astronomical Data Analysis Software and Systems XV*, C. Gabriel, C. Arviset, D. Ponz & S. Enrique, réds., tm. 351 de *Astronomical Society of the Pacific Conference Series*

NGUYEN, N. H., 2006 ; *Concecption et réalisation d'un générateur de base de données astronomiques : Saada*. Thèse de doctorat, Strasbourg, France

NGUYEN, N. H., MICHEL, L. & MOTCH, C., 2004 ; *SAADA : An Automatic Archival System for Astronomical Data*. Dans *Astronomical Data Analysis Software and Systems (ADASS) XIII*, F. Ochsenbein, M. G. Allen & D. Egret, réds., tm. 314 de *Astronomical Society of the Pacific Conference Series*

RAYMOND, E. S., 2004 ; *L'Art de la programmation Unix*. Vuibert, Paris

WINSTANLEY, N., TAYLOR, J. D., TAYLOR, M. B., NODDLE, K., GONZALEZ-SOLARES, E. & LINDROOS, J., 2007 ; *Astro Runtime : An API to the Virtual Observatory*. Dans *Astronomical Data Analysis Software and Systems XVI*, R. A. Shaw, F. Hill & D. J. Bell, réds., tm. 376 de *Astronomical Society of the Pacific Conference Series*

Conclusion & Perspectives

Conclusion

Durant ce travail de thèse, nous avons mis en place un algorithme de corrélation croisée basé sur les probabilités bayesiennes, sur une convolution des ellipses d'erreurs sur les positions robuste aux pôles, sur le *kernel smoothing* pour le calcul des densités locales et sur une estimation originale du taux de fausses associations. La communauté devient de plus en plus demandeuse de ce type d'algorithme. Dans ce contexte, nous avons entamé une collaboration avec l'équipe du CDS pour inclure une partie de ce code dans Aladin, logiciel phare pour nombre d'astronomes.

L'outil que nous avons développé, et qui implémente cet algorithme, nous a permis de corréler le catalogue 2XMMi, plus grand catalogue de sources X à l'heure actuelle, avec des catalogues d'archive *all sky* (couvrant l'intégralité du ciel) tels que le 2MASS, les catalogues USNO ou encore le GSC2. Il nous a également permis d'effectuer une corrélation croisée entre le 2XMMi et le SDSS DR7. Le SDSS présente à l'heure actuelle le plus grand échantillon homogène de sources optiques avec des identifications spectroscopiques pour un certain nombre de ces sources. Les résultats de cette corrélation font l'objet de la discussion scientifique du chapitre IV, section 1.3. Nous montrons par exemple un effet de décalage vers le bleu dans la bande $g - r$ des étoiles qui ont la plus forte luminosité X. Ce décalage pourrait être dû à des *flares* optiques. Nous confirmons à partir de nos données le durcissement spectral des couronnes actives avec l'augmentation de la luminosité X. De même, nous confirmons l'absence de dépendance entre dureté des spectres X des QSO et redshift.

L'analyse en composantes principales (ACP) est un algorithme qui permet de faire ressortir parmi les paramètres d'un jeu de données des axes qui maximisent la variance des données projetées dessus. Ces axes sont des combinaisons linéaires des paramètres d'entrée. Souvent, ce sont ces axes qui permettent de mieux distinguer visuellement les différentes classes d'objets qui peuplent une distribution de points. L'ACP classique n'est pas robuste et quelques objets éloignés de la distribution, des *outliers*, peuvent biaiser les résultats. Nous avons développé un algorithme qui prend en compte les erreurs de mesure pour diminuer l'effet dans le calcul des axes principaux des *outliers* qui possèdent de grandes incertitudes sur leurs valeurs. L'ACP permet de trouver dans un jeu de paramètres ceux qui sont les plus pertinents pour séparer les différentes classes d'objets. Cela nous a par exemple permis de minimiser l'importance des rapports de dureté comparés au rapport Fx/Fopt et aux couleurs optiques pour la classification des sources 2XMMi/SDSS DR7. L'ACP sert également à réduire la dimensionnalité d'un jeu de données en maîtrisant la perte d'information. Elle autorise donc par exemple une inspection visuelle de données projetées dans le sous-espace de dimension 3 dans lequel elles sont le mieux séparées. L'ACP améliorée que nous avons developpée et implémentée a notamment été utilisée pour la visualisation de données et pour réduire le nombre de dimensions en entrée des algorithmes de classification.

Nous avons étudié, implémenté et comparé différents algorithmes de classification : le k-NN, les *meanshifts*, les réseaux de neurones de Kohonen (SOM et LVQ), la *kernel density classification*, les arbres de décision avec l'algorithme OC1, et les SVM. Il ressort que les différents algorithmes sont plus ou moins équivalents en terme de résultats de classification. Les différences sont souvent plus

importantes en utilisant des paramètres différents pour un même algorithme qu'en utilisant deux algorithmes différents. Pour les algorithmes supervisés, la qualité de la classification dépend surtout de la qualité de l'échantillon d'apprentissage. Si cet échantillon est mal défini, qu'il ne couvre pas l'ensemble des zones occupées par les sources, on peut utiliser un algorithme non supervisé comme les *meanshifts*. Le *meanshift* permet également la recherche de populations particulières alors que les algorihtmes supervisés ne permettent de trouver que les populations déjà connues. Parmi toutes les méthodes que nous avons testées, seule la *kernel density classification* (KDC) fournit en sortie pour chaque objet la probabilité qu'il appartienne à chacune des classes. On peut par exemple jouer sur ces probabilités pour constituer des échantillons plus ou moins fiables, ou pour chercher les objets qui se trouvent à la frontière entre deux classes. C'est cet algorithme, la *kernel density classification* (KDC), que nous avons choisi pour les différentes classifications des sources du 2XMMi.

Pour faire de la classification supervisée avec la KDC, il nous a fallu construire un échantillon d'identifications. Cet échantillon d'identifications à permis la construction des différents échantillons d'apprentissage par de simples jointures avec les différents jeux de données. Ces jointures portent sur les identifiants des sources X. Nous avons pour cela utilisé certaines corrélations réalisées par l'ACDS, des données de la littérature et les résultats de la corrélation avec la DR7 du SDSS pour construire cet échantillon. Le SDSS contient des identifications spectroscopiques qui ont servi de base pour la définition de nos classes d'objet. Il manquait cependant des sources stellaires. Nous avons remédié à ce problème en classant les sources optiques SDSS corrélées aux sources X par une *kernel density classification* sur les couleurs optiques du SDSS.

Nous avons essentiellement classé les sources de deux échantillons : le 2XMMi/SDSS DR7 (XS) et le 2XMMi/GSC2/2MASS (XGT). Ces deux échantillons sont complémentaires : le XGT contient les sources brillantes qui sont saturées dans le XS ; le XS est plus profond est permet de classer de nombreux quasars trop faibles pour être dans le XGT. Les sources du 2XMMi ont également été classées sur la base de leurs propriétés X. Comme on pouvait s'y attendre, dans ce dernier cas la confusion entre les différentes classes est plus importante que pour les échantillons XS et XGT.

Enfin, nous avons mis en place des mécanismes permettant de mettre à jour la XCAT-DB en apportant de nouveaux catalogues et des paramètres comme la distance en sigma et la probabilité d'identification à certains liens de corrélation existant. De même, nous avons inclus les résultats de classification sous forme de relations entre les sources du catalogue 2XMMi et une nouvelle collection contenant uniquement des types d'objet.
Les résultats de nos travaux sont donc par ce biais accessibles par le reste de la communauté.

Perspectives

Les données du XID, qui regroupe différents programmes de suivi optique de champs de vue XMM menés au sein du SSC, devraient bientôt être disponibles. Parmi ces programmes, notre échantillon d'apprentissage (d'identifications) ne contient actuellement que les sources du XMS et du XBS. Nous pensons que l'utilisation de données plus complètes pourrait améliorer la qualité de cet échantillon, notamment en apportant des identifications spectroscopiques pour les sources plus faibles que 19-20 mag dans les bandes optiques, ces sources étant trop faibles pour figurer dans les cibles spectroscopiques du SDSS. Un échantillon conséquent pourrait également nous permettre de nous affranchir de certains catalogues comme le Véron dont les données sont fortement hétérogènes (Véron-Cetty & Véron 2006). En effet, le Véron est une compilation de sources d'origines diverses : littérature, catalogues, etc. Le but serait d'obtenir un échantillon le plus complet et le plus homogène possible. On pourrait également inclure dans notre échantillon d'autres données comme celles venant de GLIMPSE (pour *Galactic Legacy Infrared Midplane Survey Extraordinaire*), relevé réalisé à partir du télescope infrarouge Spitzer (Benjamin *et al.* 2003).

Concernant l'analyse en composantes principales, les résultats que nous avons obtenus avec

notre algorithme nous ont paru satisfaisants. Il pourrait cependant être intéressant d'étudier les méthodes connues sous le nom de *Total Least Square* (TLS Branham 1999), *orthogonal regression*, *error-in-variable regression* ou *measurement error modeling*. Ces méthodes prennent en compte les erreurs de mesures et semblent traiter proprement les problèmes liés à l'hétéroscédasticité.
On pourrait également utiliser les résultats de classifications pour prendre en compte les types d'objet dans l'ACP : chaque source serait pondérée en fonction de sa population d'appartenance. Le but serait d'éviter qu'une population d'objet majoritaire n'influence fortement la détermination des axes principaux.

La corrélation croisée avec le SDSS est actuellement basée uniquement sur les positions avec un domaine de recherche proportionnel à l'ellipse de convolution à 1σ des erreurs sur les positions. On exclut donc les sources X non résolues en optique et situées en dehors du centre de la galaxie hôte. Pour corréler ces sources, il faut prendre en considération l'extension des galaxies. Le choix d'un rayon de recherche n'est pas trivial en raison du flou existant sur la limite de l'extension spatiale des galaxies. On peut choisir plus ou moins arbitrairement un rayon, par exemple en fixant le pourcentage de la luminosité totale de la galaxie qu'il contient. Mais il faut adapter ce pourcentage en fonction du profil de luminosité de la galaxie. Par exemple, un rayon qui contient 90% de la luminosité totale d'une galaxie représentera une part importante de l'extension d'une galaxie elliptique « normale » alors qu'il ne contiendra que la partie centrale d'un AGN, le noyau étant particulièrement brillant dans ce dernier cas.
De même nous avons ignoré les sources X étendues qui représentent un peu moins de 10% des sources du 2XMMi. Il peut par exemple s'agir de l'émission diffuse d'un centre galactique, de restes de supernova qui ne possèdent pas forcément de contrepartie optique, ou du gaz chaud tombant dans le centre de gravité d'un amas de galaxies. Dans ce dernier, s'il y a une contrepartie optique, il peut s'agir d'une galaxie supermassive telle qu'on en trouve au centre de certains amas, soit des galaxies environnantes alors que l'émission X vient du gaz chaud.
Dans ces deux cas, si on prend en compte l'extension spatiale des sources et/ou des sources d'archives étendues, il faut revoir le formalisme qui a été défini uniquement pour les sources ponctuelles.

La classification pourrait être améliorée pour classer plus efficacement les *outliers*. On propose pour cela dans le chapite IV section 4.6 de définir des *outliers* à partir d'une limite en densité locale et de classer ces *outliers* et le reste des sources de façon indépendante.
De même, comme proposé section 1.2.2.b du chapitre IV, on pourrait améliorer notre algorithme actuel de *kernel density classification* en calculant par la méthode MCMC (voir chapitre II, section5.5.2) une matrice de largeur de bande pour chaque classe de l'échantillon d'apprentissage.
Une autre amélioration concerne les données manquantes. On pourrait éventuellement remplacer un paramètre non défini pour une source par la valeur moyenne de ceux des sources voisines les plus proches dans l'espace de dimension inférieure.

Les résultats de la corrélation entre le 2XMMi et le SDSS DR7 vont probablement servir de base pour le choix de nouvelles cibles d'observations spectroscopiques du SDSS. Ces nouvelles observations pourraient apporter des identifications pour les quasars rouges ou les étoiles M les plus lumineuses en X.
Des études scientifiques plus poussées pourraient être envisagées sur nos différents échantillons : recherche de QSO de type II, liens entre raies d'émissions optiques des galaxies et propriétés X, recherche d'un lien entre *redshift* et absorption intrinséque des QSO, etc.

Les méthodes développées ne sont pas propres au catalogue d'XMM. Nous avons par exemple appliqué avec succès l'algorithme de corrélation entre le catalogue de Chandra et des catalogues optiques, entre le XMM Slew Survey et l'USNOB1, et entre le Rosat All Sky Survey, le 2MASS et le GSC2. L'algorithme des *meanshifts* a également été appliqué sur ce dernier échantillon pour séparer sources stellaires et sources extragalactiques. Le résultat pourrait par exemple être utilisé pour mieux faire ressortir la ceinture de Gould ou pour étudier la structure du plan galactique. Toujours avec cet échantillon, nous avons déjà obtenu des résultats dans le cadre de la recherche de courants d'étoiles jeunes.



Références

BENJAMIN, R. A., CHURCHWELL, E., BABLER, B. L., BANIA, T. M., CLEMENS, D. P., COHEN, M., DICKEY, J. M., INDEBETOUW, R., JACKSON, J. M., KOBULNICKY, H. A., LAZARIAN, A., MARSTON, A. P., MATHIS, J. S., MEADE, M. R., SEAGER, S., STOLOVY, S. R., WATSON, C., WHITNEY, B. A., WOLFF, M. J. & WOLFIRE, M. G., 2003 ; *GLIMPSE. I. An SIRTF Legacy Project to Map the Inner Galaxy*. PASP, **115**, 953

BRANHAM, JR., R. L., 1999 ; *A Covariance Matrix for Total Least Squares with Heteroscedastic Data*. AJ, **117**, 1942

VÉRON-CETTY, M.-P. & VÉRON, P., 2006 ; *A catalogue of quasars and active nuclei : 12th edition*. A&A, **455**, 773

Annexes

Annexe A

Les Quaternions

Cette annexe décrit quelques aspects techniques sur les quaternions en astronomie. Les quaternions sont des objets géométriques 4D qui permettent de calculer des rotations dans un espace en 3D. Ils sont notamment utilisé dans l'aérospatial pour décrire précisément et sans ambiguïté le positionnement de satellite ou encore dans les moteurs 3D de jeux vidéo pour diminuer les temps de calcul.

Soit le repère $(O, \vec{x}, \vec{y}, \vec{z})$, $\vec{x}$ représentant la direction du point vernal, $\vec{y}$ celle de l'est et $\vec{z}$ celle du pôle Nord. Dans ce repère, le point vernal a donc pour coordonnées $(1, 0, 0)$ et le pôle Nord $(0, 0, 1)$.

Un quaternion permettant de passer du point vernal γ à un point quelconque de coordonnées équatoriales (α, δ) peut être défini par :

$$\boxed{Q_{\gamma\to(\alpha,\delta)} = \begin{pmatrix} +\cos\frac{\alpha}{2}\cos\frac{\delta}{2} \\ +\sin\frac{\alpha}{2}\sin\frac{\delta}{2} \\ -\cos\frac{\alpha}{2}\sin\frac{\delta}{2} \\ +\sin\frac{\alpha}{2}\cos\frac{\delta}{2} \end{pmatrix}} \tag{A.1}$$

Démonstration : Pour placer le point vernal γ à un point quelconque de coordonnées équatoriales (α, δ), il faut faire une rotation d'angle $-\delta$ autour de l'axe $\vec{y}$ suivi d'une rotation d'angle α autour de l'axe $\vec{z}$. On associe respectivement à ces deux transformations les quaternions :

$$Q_\delta = \begin{pmatrix} +\cos\frac{\delta}{2} \\ 0 \\ -\sin\frac{\delta}{2} \\ 0 \end{pmatrix}, \text{ et } Q_\alpha = \begin{pmatrix} +\cos\frac{\alpha}{2} \\ 0 \\ 0 \\ +\sin\frac{\alpha}{2} \end{pmatrix}$$

La succession de ces deux opérations peut s'exprimer sous la forme d'un quaternion résultant du produit, dans cet ordre, de Q_α par Q_δ :

$$Q_{\gamma\to(\alpha,\delta)} = Q_\alpha Q_\delta \tag{A.2}$$

Pour les formules du produit de quaternions, voir par exemple la documentation du site geometrictools[1]

Il peut aussi être défini par :

$$Q_{\gamma\to(\alpha,\delta)} = \begin{pmatrix} \rho \\ 0 \\ -\cos\theta \\ +\sin\theta \end{pmatrix} \tag{A.3}$$

[1] http://www.geometrictools.com/Documentation/Quaternions.pdf

Avec ρ la distance angulaire entre le point γ et le point (α, δ) et θ l'angle en γ entre le pôle Nord et la direction du point (α, δ). Dans ce cas particulier, on a :

$$\rho = 2 \arcsin \sqrt{\sin^2 \frac{\delta}{2} + \sin^2 \frac{\alpha}{2} \cos \delta} \tag{A.4}$$

et :

$$\theta = \begin{cases} 180 - \arcsin \frac{(90^\circ - \delta) \cdot \sin \alpha}{\rho} & \text{si} \quad \sqrt{\frac{\cos(90+\rho-\delta)}{\sin \rho}} > 1 \quad \text{et} \quad \arcsin \frac{(90^\circ - \delta) \cdot \sin \alpha}{\rho} \geq 0 \\ -180 - \arcsin \frac{(90^\circ - \delta) \cdot \sin \alpha}{\rho} & \text{si} \quad \sqrt{\frac{\cos(90+\rho-\delta)}{\sin \rho}} > 1 \quad \text{et} \quad \arcsin \frac{(90^\circ - \delta) \cdot \sin \alpha}{\rho} < 0 \\ \arcsin \frac{(90^\circ - \delta) \cdot \sin \alpha}{\rho} & \text{sinon} \end{cases} \tag{A.5}$$

Soit $\vec{V} = (v_1, v_2, v_3)$ l'image du vecteur $\vec{U} = (u_1, u_2, u_3)$ par la rotation décrite par le quaternion $Q = [q_0, q_1, q_2, q_3]$ ayant pour axe $\vec{A} = (q_1, q_2, q_3)$. On trouve $\vec{V}$ par la formule :

$$\vec{V} = (2q_0^2 - 1)\vec{U} + 2q_0 \vec{A} \wedge \vec{U} - 2(\vec{A}.\vec{U})\vec{A} \tag{A.6}$$

Ce qui nous donne après développement :

$$v_1 = (2q_0^2 - 1)u_1 + 2q_0(q_2 u_3 - q_3 u_2) + 2(q_1 u_1 + q_2 u_2 + q_3 u_3)q_1 \tag{A.7}$$
$$v_2 = (2q_0^2 - 1)u_2 + 2q_0(q_3 u_1 - q_1 u_3) + 2(q_1 u_1 + q_2 u_2 + q_3 u_3)q_2 \tag{A.8}$$
$$v_3 = (2q_0^2 - 1)u_3 + 2q_0(q_1 u_2 - q_2 u_1) + 2(q_1 u_1 + q_2 u_2 + q_3 u_3)q_3 \tag{A.9}$$

On peut retrouver des coordonnées (α, δ) à partir d'un quaternion $Q = [q_0, q_1, q_2, q_3]$ de façon générale en utilisant les formules :

$$\alpha = \arctan \frac{2(q_0 q_3 + q_1 q_2)}{2(q_0^2 + q_1^2) - 1} \tag{A.10}$$
$$\delta = \arcsin 2(q_1 q_3 - q_0 q_2) \tag{A.11}$$

Démonstration : On calcule l'image du vecteur $\vec{U} = (1, 0, 0)$ par le quaternion Q et on applique les formules classiques :

$$\alpha = \arctan \frac{v_2}{v_1} \tag{A.12}$$
$$\delta = \arcsin v_3 \tag{A.13}$$

On peut également retrouver les coordonnées (α, δ) à partir d'un quaternion $Q = [q_0, q_1, q_2, q_3]$ dont les coefficients vérifient au moins :

$$q_0 q_1 = -q_2 q_3 \tag{A.14}$$

en s'aidant des formules :

$$\sin \alpha = 2(q_0 q_3 - q_1 q_2) \tag{A.15}$$
$$\sin \delta = 2(q_1 q_3 - q_0 q_2) \tag{A.16}$$
$$\cos \alpha = 1 - \frac{2q_1 q_3}{q_1 q_3 - q_0 q_2} \tag{A.17}$$
$$\cos \delta = 1 + \frac{2q_1 q_2}{q_0 q_3 - q_1 q_2} \tag{A.18}$$

Démonstration : Remplacer les coefficients q_i par ceux du quaternion $Q_{\gamma \to (\alpha, \delta)}$ et se servir des formules trigonométriques du type : $\sin 2a = 2 \sin a \cos a$ et $\cos 2a = 1 - 2 \sin^2 a$.

Annexe B

Logiciel de création de SOM

Nous presentons ici le message d'aide du logiciel de création de cartes auto-organisatrices de Kohonen que nous avons developpé.

```
USAGE:
  java -cp ../extern/libs/kdtree.jar:../java/classes kohonen.Network -h
  java -cp ../extern/libs/kdtree.jar:../java/classes kohonen.Network dim inputCsv unknownLabel
          topology dimensions algo itMax alpha "alphaArgs" neighbour "neighbourArgs" outBase
          fracMin fracMax [ weightsFile | -sw ] [-cr] [-t1t2] [-sai]
PARAMETERS:
  dim              the dimension of the input parameter space
  inputCsv         the name of the csv containing the input data (cf. INFO)
  unknownLabel     the type in the inputCsv file which stands for unknown type objects
  topology         the topology of the neuron network (see TOPOLOGY for posibilities)
  dimensions       the dimensions of the network (exemple in 2D: 45x20; in 3D: 20x10x30)
  algo             the algorithm you want to use: 'adaptative' or 'adaptativeRand' or 'batch'
                   (WARNING the batch algorihtm is not yet implemented!)
  itMax            the index of the last iteration (as index start at 0, number of iteration =
                   itMax+1 = number of time that inputs are put in the network for learning)
  alpha            the alpha function you want to use (cf. ALPHA FUNCTION)
  alphaArgs        the arguments to give to the alpha function (cf. ALPHA FUNCTION)
  neighbour        the neighbour function you want to use (cf. NEIGHBOUR FUNCTION)
  neighbourArgs    the arguments to give to the neighbour function (cf. NEIGHBOUR FUNC.)
  outBase          the base name of the network outputs (for the list of outputs, cf. OUTPUTS)
  fracMin          the minimum fraction of objects of same type on a neuron to consider the
                   neuron is of that type
  fracMax          for each other type, the maximum fraction of object of a type not to consider
                   the neuron of an unknown type
  weightsFile      if you want to load pre-define random weights to use same initial conditions in
                   several try, name of the random weights file
OPTIONS:
  -h       stands for "help"                : print this message, then exit
  -sw      stands for "save weight"         : create an output containing the initial random weights
  -cr      stands for "center and reduce"   : the inputs data are centered and reduced
  -t1t2    stands for "T1 and T2"           : compute after each iteration the T1 and T2 quantities
                                              and save them (c.f. INFO) (slow the algorithm)
  -sai     stands for "save all iterations" : after each iteration the output of the network is saved
                                             (slow down the algorithm)
  -ffuid   stands for "filter for unique ID" : if several inputs have the same ID, keeps only the last
TOPOLOGY: (more info reading the JavaDoc)
  Matrix2D
```

```
  Hexagonal2D
ALPHA FUNCTION: (more info reading the JavaDoc)
  CONSTANT   1 argument : 'a' (the constant factor)
  RITTER     2 arguments: 'alpha_init' (initial value) and  'alpha_end' (final value)
NEIGHBOUR FUNCTION: (you can learn more reading the JavaDoc)
  BINARY   1 argument : 'd' (the limit distance)
  CONE     1 argument : 'd' (the limit distance)
OUTPUTS:
  outBase.som.csv             final output
  outBase.som.[0-9]+.csv      output after each iterations
  outBase.rand.weights.csv    initial random weights
  outBase.t1t2.csv            values of T1 and T2 after each iteration
  outBase.classified.csv      classified inputs with associated neuron id and position
  outBase.classif.stat.csv    classication statistics
ALGO:
  adaptative          inputs are presented to the network in the same order than in the input file,
                      so they are presented in the same order for each iteration
  adaptativeRand      inputs are presented to the network in a totaly random way, at each iteration
                      all inputs are presented and each input is presented only once
  batch               the neuron weight are updated only after having presented all inputs to the
                      network, so weights are updated only one at each iteration
INFO:
  - In CSV files, columns are separated by a coma (character ',')
  - The first line which is not a comment line can contains column names
  - The column names must begin by a letter or a '_'
  - The first data column must contains a line identifier (read as a 'String')
  - The second column must contains a type of object (also read as a 'String')
  - The input number of columns 'dim' does not take into account the ID and the TYPE
      columns (nbrColTot=2+nbrDataCol)
  - The input CSV file can contains comments before the first data row or the column names
      row in the file
  - Comments lines must begin by the character '#'
EXAMPLE:
  java -cp ../extern/libs/kdtree.jar:../java/classes kohonen.Network 10 acp.out.csv UNK
    Matrix2D 40x40 adaptative 50 RITTER "0.9 0.0005" CONE "5" acp.kohonen 0.1 0.7 -t1t2
```

Annexe C

Liste des travaux publiés ou en cours de publication

Voici la liste des travaux que nous avons publié ou qui sont en cours de publication. Certaines de ces références sont fournies dans les annexes suivantes.

Publications

Nous avons collaboré en tant que premier auteur ou co-auteur aux articles (acceptés, soumis ou en préparation) suivant :

- Cross-Correlation of the 2XMMi catalogue with Data Release 7 of the Sloan Digital Sky Survey
  F.-X. Pineau, C. Motch, S. Derriere, L. Michel, et al
  En préparation
- The X-ray source content of the XGPS Galactic Plane Survey and the nature of X-ray selected sources
  C. Motch, R. Warwick, M. S. Cropper, F. Carrera, P. Guillout, **F.-X. Pineau**, M. W. Pakull, S. Rosen, A. Schwope, J. Tedds, N. Webb et al.
  En préparation

Communications

L'avancement de nos travaux à été présenté à divers conférences sous forme de poster ou de présentation orale :

- A VO Based Tool Computing Cross-Identification Probabilities for the 2XMM catalogue
  F.-X. Pineau, S. Derriere, C. Motch, L. Michel
  Poster pour la conférence "Astronomical Data Analysis Software And Systems XVII", Septembre 2007 (`www.adass.org/adass2007`)
- Building an Astronomical Database with Saada
  L. Michel, C. Motch, **F.-X. Pineau**
  Focus démo pour la conférence "Astronomical Data Analysis Software And Systems XVII", Septembre 2007 (`www.adass.org/adass2007`)
- Statistical identification of 2XMMi sources
  F.-X. Pineau, S. Derriere, L. Michel, C. Motch
  Contribution orale lors de la conférence "Classification and Discovery in Large Astronomical Surveys", Octobre 2008 (`http://www.mpia-hd.mpg.de/class2008/`)

- Statistical identification of 2XMMi sources
  F.-X. Pineau, S. Derriere, L. Michel, C. Motch
  Poster pour la conférence "Astronomical Data Analysis Software And Systems XVIII", Novembre 2008 (`www.adass.org/adass2008`)
- Building an Astronomical Database with Saada
  L. Michel, C. Motch, H. Nguyen Ngoc, **F.-X. Pineau**
  Poster et focus démo pour la conférence "Astronomical Data Analysis Software And Systems XVIII", Novembre 2008 (`www.adass.org/adass2008`)
- The X-cat-DB : A multi-wavelength view on the 2XMMi catalogue
  C. Motch, L. Michel, **F.-X. Pineau**, A. C. Gonçalves
  Poster pour la conférence "Astronomical Data Analysis Software And Systems XVIII", Novembre 2008 (`www.adass.org/adass2008`)
- Multi-wavelenghtobservations of high energy sources
  C. Motch et **F.-X. Pineau**
  Présentation orale lors du Workshop EURO-VO "Multi-wavelength Astronomy and Virtual Observatory", Décembre 2008 (`http://esavo.esa.int/MultiwavelengthVOWorkshopDec2008`)

Autres

Dans le cadre d'une collaboration, nous avons participé à deux demandes de temps téléscope :

- Searching for co-moving young stars around post-T Tauri candidates
  D. Montes, J. Lopez-Santiago, P. Guillout, A. Klutsch, E. Marilli, N. Grosso, **F.-X. Pineau**, R. Freire Ferrero, A. Frasca
  Demande de temps soumise et accéptée
- Searching for co-moving young stars around post-T Tauri candidates in Cepheus
  P. Guillout, A. Frasca, D. Montes, A. Klutsch, **F.-X. Pineau**, E. Marilli, R. Freire Ferrero, N. Grosso, J. Lopez-Santiago
  Demande soumise

Annexe D

Proceeding ADASS 2007

Astronomical Data Analysis Software and Systems XVII **P3.4**
ASP Conference Series, Vol. XXX, 2008
J. Lewis, R. Argyle, P. Bunclarck, D. Evans, and E. Gonzales-Solares, eds.

A VO based tool computing cross-identification probabilities for the 2XMM catalogue

François-Xavier Pineau, Sebastien Derriere, Laurent Michel, Christian Motch

Observatoire Astronomique de Strasbourg

Abstract. We present a tool developed to cross-correlate 2XMM sources with various other catalogues. Archival data are collected using VO protocols. The cross-correlation algorithm is based on a likelihood ratio (LR) technique. It takes into account local densities and estimates the expected number of spurious associations as a function of LR without resorting to Monte Carlo simulations. Groups of similar XMM fields of view are stacked to increase statistics and provide better fits to reliability functions. These informations are used to enrich qualified links between X-ray sources and possible counterparts in archival catalogues held in the XCAT-DB, one of the Survey Science Center interface to the 2XMM catalogue. These identification probabilities will be later used to develop an automated classification algorithm.

1. Introduction

We are developing a cross-identifier and a classifier for the 2XMM catalogue. The first step is to gather using the VO data from several astronomical catalogues at various wavelengths and make them available to the cross-identifier. 2XMM sources are then cross-correlated with entries in these catalogues on the basis of their coincidence in positions. The results from this cross-correlation process are then stored in the XCAT-DB (Michel et al. 2004).
We first present an overview of the process. We then describe the method used by the cross-identifier to attribute a probability of identification to each association. We end this presentation by giving some preliminary results.

2. Process Overview & Implementation

As shown in the Figure 1, 2XMM sources are retrieved using direct SQL queries – via a JDBC connection – on the XCAT-DB.
Archival sources are collected from the VO thanks to the ConeSearch protocol, the correlation is thus performed independently for each XMM observation. Sources are filtered and stored in quadtrees to speed up the search for counterparts and the estimate of local densities.
This process enriches the cross correlations stored in the XCAT-DB by appending new archival catalogues (e.g. SDSS5 sources and identifiers) and by providing qualified links (e.g. identification probability, local density of archival sources).

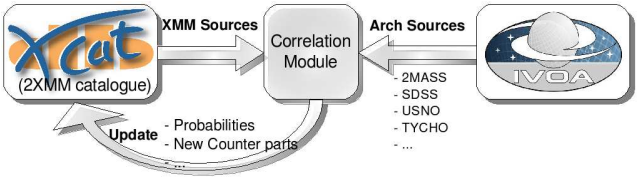


Figure 1. Cross-identification of the 2XMM: data fluxes.

3. Method

The method we have used is based on a classical likelihood ratio where local densities are computed using kernel smoothing.

3.1. Likelihood ratio

The likelihood ratio (LR) we used to score an association is inspired by the one describe in De Ruiter et al (1977):

$$LR = \frac{P_{cp}(d)}{P_{spu}(d)} = \frac{1}{2\lambda} e^{-\frac{1}{2}d^2} \tag{1}$$

It depends on:

- P_{cp}: the convolution product of the 2D Gaussian describing the positional error of the X-ray source with that of the archival source. It represents the probability that the 2 sources have to be exactly at the same location
- P_{spu}: the probability that the candidate is found by chance at this distance
- d: the distance between the 2 sources normalized by positional errors
- $\lambda = \rho\pi\sigma_{Maj}\sigma_{Min}$: local density times an elementary surface

If errors are circular: $d = \sqrt{\frac{\Delta\alpha^2+\Delta\delta^2}{\sigma_x^2+\sigma_{cp}^2}}$ and $\lambda = \rho\pi(\sigma_x^2 + \sigma_{cp}^2)$. Our algorithm handles oriented ellipses (see Appendix of Condon et al. 1995) and is robust at the poles. The local density $\rho = \rho(\alpha_x, \delta_x, mag_{cp})$ only takes into account sources at least as bright as the candidate we are considering in the association.

3.2. Local density estimation

Local densities are estimated using a variable bandwidth kernel smoothing:

- We center on each source a kernel whose basis size can be fixed or variable
- The density at a given point is defined as the sum of the contributions of all kernels at this point

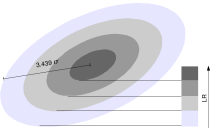
The kernel we have chosen is the 2D Epanechnikov profile:

$$K_E(\mathbf{x}) = \begin{cases} \frac{2}{\pi}(1 - ||\mathbf{x}||^2) & \text{if} \quad ||\mathbf{x}|| < 1 \\ 0 & \text{else} \end{cases} \tag{2}$$

- It maximises the quality of the reconstruction
- Its simple geometry, a truncated parabol, implies that we do not need all sources to compute the local density at a given position

3.3. Rate of spurious associations estimation

The LR distribution of spurious associations is usually computed by Monte Carlo simulations. We have implemented a new and faster method.
We consider N XMM and M candidate (CP) sources in a field of view (FOV) and assume that they are all randomly distributed on the sky. The number of spurious associations is then estimated by computing the surface a bin of LR (ΔLR) occupies with respect to the field of view area (S_{FOV}). We then compute the $N \cdot M$ convolution products of all possible associations among the N XMM and M archival sources.
In order to avoid computing all local densities (λ_j, $j \in [1, M]$), we discretize them according to the logN-logS curve. We then determine the surfaces S each bin of LR occupies inside each of the $N \cdot M$ convolution ellipses.



The sum:

$$N_{spur}(\Delta LR) = \sum_{x=1}^{N} \sum_{cp=1}^{M} S_{x,cp}(\Delta LR, \lambda_{\Delta mag_{cp}}) \,/ S_{FOV} \tag{3}$$

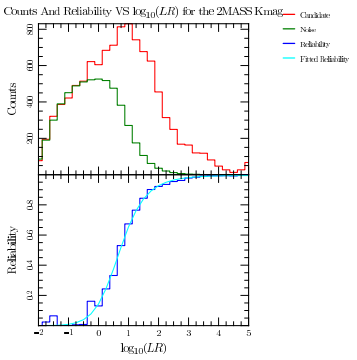
provides an estimate of the number of spurious associations for a given bin of likelihood ratio.

3.4. Reliability

Following Oyabu et al. (2005) we construct a histogram of LR for all associations and for the spurious associations estimate. The reliability:

$$R(LR) = \frac{N_{cand}(LR) - N_{spur}(LR)}{N_{cand}(LR)} \tag{4}$$

is defined as the ratio of the number of true associations to the total number of associations and can be fitted by the function:



$$R_f(LR) = \frac{1}{1 + C\frac{1}{P(LR)}} \tag{5}$$

Where C is a constant term and P is a polynomial of order 2 without zero order term. Its mathematical form derives from the exact solution of a simplified version of the problem in which all errors are identical.

3.5. Probability of identification

For a given XMM source having M candidates, the probability of identification of the ith candidate is computed according to the Rutledge et al. (2000) formula:

$$P_{id,i} = \frac{R_i}{1 - R_i} \Bigg/ \left(1 + \sum_{i=1}^{M} \frac{R_i}{1 - R_i}\right) \tag{6}$$

For each candidate, we define as many LRs and reliabilities as the number of magnitude bands in the archival catalogue. R_i corresponds to the best reliability among these different magnitudes:

$$R_i = \max_k \left(R_f(LR_{i,k})\right) \tag{7}$$

With $k \in [1, n]$, n being the number of magnitude bands defined in the archival catalogue.

3.6. Observation grouping

XMM sources are correlated FOV by FOV. In order to tail off countrate noise on individual FOV LR histograms without loosing resolution, we have to increase count statistics by stacking data coming from similar FOVs:

- We split XMM FOVs according to their systematic positional errors
- Observations of the Magellanic Clouds are treated separately
- Since they share objects of comparable nature and logN-logS distribution, observations are grouped according to their galactic latitude

4. Some preliminary results

The 2XMM contains 191870 unique sources. The table below gives:

- The number of unique X-ray sources associated with at least one archival source within the 3.439 sigma combined error ellipse
- The number of unique X-ray sources having an association with a probability of identification based on positional coincidences ≥ 0.9
- The same number expressed as a percentage of the total number of 2XMM unique sources.

	USNOB1	2MASS	SDSS 5
Source X *all*	79190	47387	42640
Source X $P_{id} \geq 90\%$	33563	19720	20033
Source X $P_{id} \geq 90\%$	18%	10%	10%

References

Condon, J. J., Anderson, E., & Broderick, J. J. 1995, AJ, 109, 2318

De Ruiter, H. R, Arp, H. C.,& Willis,A. G. 1977, A&A, 28, 211

Michel, L., Herent, O., Motch, C., Pye, J., & Watson, M. G. 2004, in ASP Conf. Ser. 314, ADASS XIII, ed. F. Ochsenbein, M. Allen, & D. Egret (San Francisco: ASP), 314, 570

Oyabu, S., Yun, M. S., Murayama, T., Sanders, D. B., Kawara, K., Taniguchi, Y., Veilleux, S., Okuda, H., Matsuhara, H., Cowie, L. L., Sato, Y., Wakamatsu, K., & Sofue, Y. 2005, AJ, 130, 2019

Rutledge, R. E., Brunner, R. J., Prince, T. A., & Lonsdale, C. 2000, ApJS, 131, 335



Annexe E

Proceeding ADASS 2008

Astronomical Data Analysis Software and Systems XVII **A.20**
ASP Conference Series, Vol. XXX, 2008
J. Lewis, R. Argyle, P. Bunclarck, D. Evans, and E. Gonzales-Solares, eds.

Statistical identification of 2XMMi sources

F.-X. Pineau, S. Derriere, L. Michel, C. Motch

CNRS, Université de Strasbourg, Observatoire Astronomique, 11 rue de l'Université, 67000 Strasbourg, France

Abstract. One of the tasks devoted to the Survey Science Center of the XMM-Newton satellite is the statistical identification of all serendipitous X-ray sources detected in the wide field of view of the EPIC cameras. For that purpose, we have cross-correlated the recently released 2XMMi source list with several major archival catalogues such as SDSS DR6 and 2MASS. For each X-ray source, probabilities of identification with associated archival entries are computed using an original method based on a classical Bayesian approach which furthermore does not rely on Monte Carlo simulations.
Identifications in the Downes catalogue of cataclysmic variables and for AGN, galaxies and stars in the SDSS DR6 form the basis of our learning sample. The multi-wavelength parameter space has been reduced by a principal component analysis which takes into account measurement errors on all data. Both a *knn* and a more elaborated kernel density smoothing approach have been tested for the supervised classification.
We report here on the current status of this project and show some illustrative results arising from the classification of the 2XMMi identifications in the SDSS DR6 catalogue.

1. Introduction

The Incremental Second XMM-Newton serendipitous source catalogue (2XMMi) is so far the largest catalogue of X-ray sources ever published. It has been compiled by the XMM-Netwon Survey Science Centre (SCC) on behalf of ESA. One of the responsibilities of the SSC is to provide the community with statistical identifications of all 2XMMi sources using multiwavelength analysis.
We presented last year (Pineau et al. 2008) an original tool that we used to cross-correlate the 2XMMi source list with various other catalogues. In the exploratory work reported here, we show how the correlation of the 2XMMi catalogue with the SDSS DR6 (DR6) and 2MASS catalogues can be used to distinguish and classify different classes of objects. In order to do so, we have first built a learning sample. Then, we have applied an improved principal component analysis (PCA) taking into account measurements errors to reduce the parameter space dimension. Finally, we have tested on our samples both a k-nearest neighbour (*knn*) and a kernel density classification.

2. Building a learning sample

The learning sample mainly consists of the spectroscopic identifications available from the DR6: Stars, Galaxies, QSOs, High Redshift QSOs (HizQSO).

We have enlarged our stellar sample by performing a *kernel density classification* (Richards et al. 2004) on the spatially unresolved (*cl=6*) objects' four main colours, cross-checked with the theoretical main sequence locus (Covey et al. 2007). Correlated entries in the Upssala (Lauberts 1982) and the Downes (Downes et al. 2001) catalogues have been used to enrich the Galaxy sample and to create a group of identified Cataclysmic Variables (CV) respectively. We have extracted from the literature additional CVs (Dillon et al. 2008) and a group of type II QSO candidates (Zakamska et al. 2003).

Being based on SDSS spectra and avoiding saturated SDSS objects, our learning sample is biased toward optical objects brighter than ∼19-20 mag and fainter than ∼14 mag respectively. Our XMM-DR6 (resp. XMM-DR6-2MASS)

Table 1. Distribution of object types in our learning samples.

Sample\Class	Star	Galaxy	QSO	HizQSO	CV	QSO II	TOTAL
XMM-DR6	513	740	1383	55	22	5	2718
XMM-DR6-2MASS	499	701	309	9	13	0	1531

sample, containing associations with probabilities of identification $\geq 95\%$, contains 9641 (resp. 956) sources which do not belong to the learning sample.

3. An improved PCA

The principal component analysis (PCA) is a well known and widely used technique to analyse high dimensional data. Its main advantages are to reduce the parameter space by finding redundancies, and to find and sort the meaningful axis according to the information – the variance – they contain. The main drawbacks are its sensitivity to extreme values – to outliers – due to the low robustness of the mean and variance estimates. Moreover, it does not account for measurement errors. Robust PCAs exist, but their usual strategy consists in minimising the effects of outliers or in estimating which points are outliers in order to reject them, ignoring measurement errors. In our case, a small population of rare objects could be significant but would be considered as outliers and be ignored by these algorithms.

We propose a method estimating the means, variances and correlation coefficients of parameters taking into account individual measurement errors on each data point. The basic idea is that for a parameter p, the variance $\sigma^2_{i,p}$ of point i is the sum of its measurement error, variance $\sigma^2_{\epsilon_{i,p}}$, plus the variance $\sigma^2_{.,p}$ due to the intrinsic scatter of the parameter: $\sigma^2_{i,p} = \sigma^2_{\epsilon_{i,p}} + \sigma^2_{.,p}$. In order to estimate $\sigma^2_{.,p}$, we assume that the total variance of centre-reduced data is equal to one and we search the zero point of Eq. 1.

$$f(\sigma_{.,p}) = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_{i,p}-\gamma_{.,p})^2}{\sigma^2_{\epsilon_{i,p}}+\sigma^2_{.,p}} - 1 \tag{1}$$

$\gamma_{.,p}$, the mean of parameter p is estimated by the weighted mean square:

$$\gamma_{.,p} = \sum_{i=1}^{n}\frac{x_{i,p}}{\sigma^2_{\epsilon_{i,p}}+\sigma^2_{.,p}} \Big/ \sum_{j=1}^{n}\frac{1}{\sigma^2_{\epsilon_{i,p}}+\sigma^2_{.,p}} \tag{2}$$

If individual errors are null or very small compared to the intrinsic scatter, our formulae reduce to the classical formulae of the mean and correlation. On the contrary, if the intrinsic scatters of the parameters are null or very small compared to measurements errors, they reduce to the classical weighted mean square formulae. Finally, if the distribution of the variances follows a Gaussian law of mean λ_{σ^2}, the total variance is $\sigma_p^2 \approx \lambda_{\sigma^2} + \sigma_{.,p}^2$.

The main interest of our improved PCA is that it takes into account measurement errors, is able to handle datasets with different typical errors and accounts for "outliers" caused by large measurement errors.

We have applied our improved PCA on 24 input parameters of the XMM-DR6 sample: 5 f_X/f_{Opt} ratios, 4 XMM-Newton EPIC hardness ratios (HR), 10 SDSS colours and 5 SDSS magnitudes. Results are shown on Figure 1. The

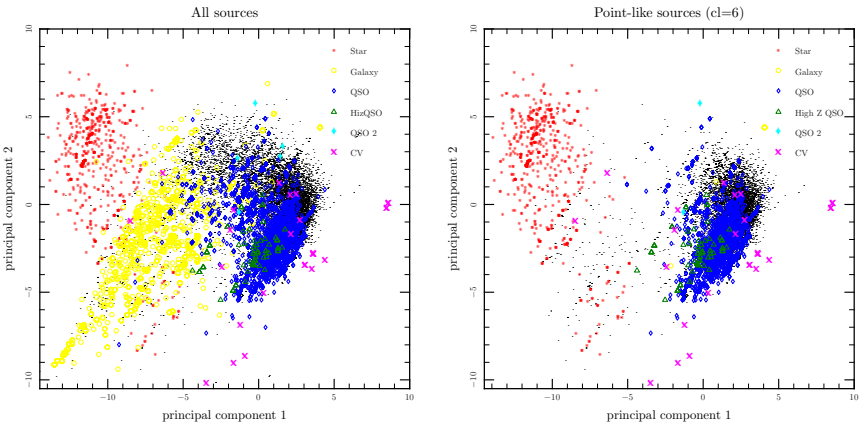


Figure 1. First principal component plane for all sources (left) and unresolved sources (right) of the XMM-DR6 sample.

main classes of X-ray emitters – stars, galaxies and AGNs – are well separated in distinct groups, whereas CVs and type II QSO candidates seem to form a 3D shell-like shape around the QSO locus (not really visible in 2D). As expected, the learning sample does not cover well the upper part (optically faint) of the Galaxy-QSO locus due to the magnitude limit of the spectroscopic observations. Although there is an overlap between stars and galaxies, the introduction of optical extension allows their separation.

We have also applied our improved PCA on 40 input parameters of the second sample:5 f_X/f_{Opt} ratios, 3 f_X/f_{IR} ratios, 4 HR, 10 SDSS colours, 3 2MASS colours, 15 optical to IR colours. The plane made by the first two principal components has the same shape as that resulting from the XMM-DR6 analysis. However, introducing optical to near infrared colours allows optimum separation between stars and galaxies in the region where optical colours and f_X/f_{Opt} ratios of the two classes overlap, without using the extension parameter.

4. *knn* and Kernel density classification

We have auto-classified our learning sample by removing for each classification the source we were classifying. We have not split the sample in two because of its small size. CVs are not properly classified and we clearly need a distinct

Table 2. Assigned class percentage (*Class*, vertically) as a function of true class in the learning sample (*True*, horizontally), using two different methods: sum of all individual class probabilities obtained with *knn*(left), and ratio of best scored objects for KC (right).

	knn proba				KC class			
True \ Class	Star	Gal	QSO	CV	Star	Gal	QSO	CV
Star	**97.6**	2.4	0.0	0.0	**97.3**	2.7	0.0	0.0
Galaxy	2.2	**90.5**	7.2	0.1	0.7	**94.6**	4.7	0.0
QSO	0.1	15.4	**83.7**	0.8	0.0	14.5	**85.5**	0.0
CV	8.8	24.2	56.0	**11.0**	16.7	16.7	66.6	**0.0**

method for outliers or shell-like shape distributions consisting of a few objects only. We remark that *knn* classification suffers from two main drawbacks: the probabilities it provides are discrete (a class cannot always be unambiguously assigned), and it does not properly handle biased learning samples (for example if the ratios between classes are different in the sample and in the result of the classification). The kernel density classification is both more elegant and convenient but requires numerous learning sample points in order to reach a good estimate of the class distributions.

5. Conclusion & prospects

In our case, a PCA with 2, 3 or 4 dimensions is sufficient to separate the main object classes and allows the detection of rare objects candidates. First trials made using all sky catalogues only – e.g. GSC2 and 2MASS – yield very promising prospects for class separation and identification.
Although the kernel density classification is more powerful than the *knn*, it is not adapted to the classification of rare objects. We plan to investigate in the near future the efficiencies of Kernel PCA and SVM.

References

Covey, K. R., Ivezić, Z., Schlegel, D., et al. 2007, AJ, 134, 2398

Dillon, M., Gänsike, B. T., Aungwerojwit, A., et al. 2008, MNRAS, 386, 1568

Downes, R. A., Webbink, R. F., Shara, M. M., et al. 2001, PASP, 113, 764

Lauberts, A. 1982, ESO/Uppsala survey of the ESO(B) atlas (Garching: European Southern Observatory (ESO), 1982)

Pineau, F.-X., Derriere, S., Michel, L., & Motch, C. 2008, in ASP Conf. Ser. XXX, ADASS XVII, ed. J. Lewis, R. Argyle, P. Bunclark, D. Evans, & E. Gonzalez-Solares (San Francisco: ASP), [P3.4]

Richards, G. T., Nichol, R. C., Gray, A. G., et al. 2004, ApJS, 155, 257

Zakamska, N. L., Strauss, M. A., Krolik, J. H., et al. 2003, AJ, 126, 2125



Annexe F

Proceeding Ringberg Castle 2008

Statistical identification of 2XMMi sources

François-Xavier Pineau, Sébastien Derriere, Laurent Michel and Christian Motch

Observatoire Astronomique, 11 rue de l'Université, 67000 Strasbourg, France

Abstract.
We present preliminary results of a project carried out by the Survey Science Centre of XMM-Newton aiming at the statistical identification of all 2XMMi catalogue sources. The 2XMMi has been cross correlated with various other large catalogues such as SDSS DR6 and 2MASS. For that purpose we have developed an original tool, based on a classical Baysian approach, which provides probabilities of identification without resorting to Monte Carlo simulations.

In order to perform supervised classifications, we have built learning samples using the Downes catalogue of cataclysmic variables, and spectroscopic identifications of AGNs, galaxies and stars in the SDSS DR6. The parameter space has been reduced by a principal component analysis. We have compared classifications using a knn approach, and a more elaborated kernel density classification. An original aspect of this work is that we take into account the heteroscedasticity of errors on each parameter.

We summarize the current status of this project and present some interesting results arising from the cross correlations and classifications.

Keywords: Statistical identification, Supervised Classification, PCA, 2XMMi, XCATDB, SDSS, 2MASS
PACS: 98.62.Tc ; 98.52.Cf ; 95.80.+p ; 98.70.Qy

INTRODUCTION

The Incremental Second XMM-Newton Serendipitous Source Catalogue (2XMMi) is so far the largest catalogue of X-ray sources ever published. It has been compiled by the XMM-Newton Survey Science Centre (SSC) on behalf of ESA. One way to access the 2XMMi catalogue and associated pipeline products is to query the XCATDB[1] interface. This database is designed to easily handle the 2XMMi sources and their cross-correlation with a large number of archival catalogues.

The SDSS DR6 (DR6) and the 2MASS are the largest sample of high quality optical sources and spectra, and the largest catalogue of infra-red sources, respectively. Large sky surveys are instrumental for the statistical study of homogeneous sample of astrophysical objects and for searching rare or extreme emitters. One of the responsibilities of the SSC is to provide the community with statistical identifications of all 2XMMi sources using a multiwavelength analysis. Such an endeavour first requires to access a well defined set of correlated data. Our cross-correlation method is shortly described in the first section. In the second section, we explain how our supervised classification

[1] `http://amwdb.u-strasbg.fr/2xmmi/home`

learning sample was built. We then present a method to account for measurement errors in a classical principal component analysis (PCA) and illustrate the outcome of our work by showing some interesting results. Finally, we compare the efficiencies of KNN and kernel density classification algorithms.

CROSS-CORRELATION OF THE 2XMMI

Our cross-correlation tool is based on a likelihood ratio [1] but uses a new method to estimate the number of spurious associations. For a given X-ray source we only consider candidate archival sources located inside the 99.7% completeness ellipse. This ellipse is derived from the convolution product of the positional error of the XMM source by that of the archival source. We asssume that positional errors are 2-D Gaussian distributions in the projection plane centered on the X-ray source.

In this frame, the *x*-axis and *y*-axis are the major axis (σ_M) and the minor axis (σ_m) of the convolution ellipse respectively. We define the dimensionless distance $r = \sqrt{x^2/\sigma_M^2 + y^2/\sigma_m^2}$ to allow the use of polar coordinates. The likelihood ratio (LR) adopted is the probability of finding the optical counterpart (cp) at a distance lying between r and $r + \mathrm{d}r$ divided by the Poissonian probability of finding a spurious object (spur) between r and $r + \mathrm{d}r$: $LR(r) = \exp(-r^2/2)/2\lambda$. We define $f(\alpha_x, \delta_x, m_o)$ as the local surface density of sources at least as bright as m_o, the magnitude of the candidate. This leads in the 1σ ellipse to a density of sources of $\lambda = \pi\sigma_M\sigma_m f(\alpha_x, \delta_x, m_o)$.

$f(\alpha_x, \delta_x, m_o)$ is estimated from a *sample-point density estimator* which consists of a variable bandwidth kernel smoothing. The bandwidth is the distance to the 100th nearest neighbour of the point (α_x, δ_x) and the kernel is the 2D Epanechnikov profile. This gives errors on $f(\alpha_x, \delta_x, m_o)$ of about 15%.

Following [2] we have built (Fig. 1) an histogram of LR for all associations and for the estimates number of spurious associations. The reliability $R(LR) = (N_{all}(LR) - N_{spur}(LR))/N_{all}(LR)$, defined as the ratio of the number of true associations to the total number of associations, has been fitted by $R_f(x = \log_{10}(LR)) = \sum_{i=1}^{n} b_i 10^{(1-i)x}/(1 + \sum_{i=1}^{n} a_i 10^{-ix})$ using the Levenberg-Marquard algorithm. R_f is the sum of n theoretical functions $p(id|r) = (1 + \frac{p(spur)}{p(cp)}\frac{1}{LR})^{-1}$ with n categories of sources having different values of $p(spur)/p(cp)$.

The most original part of our work lies in our estimate of $N_{spur}(\Delta LR)$. $N_{spur}(\Delta LR)$ is needed to construct the histogram of spurious associations. We list below the main steps of our method:

- We first consider the case of only one X-ray and one optical source of magnitude m in a field of view (FOV).
- The probability of observing the random X-ray source at any given point (α_x, δ_x) of the FOV is given by the distribution $g_X(\alpha_x, \delta_x)$ of the X-ray sources in the FOV.
- The chance for the optical source of magnitude m to be randomly associated with the X-ray source is given by the integral of the distribution of optical sources at

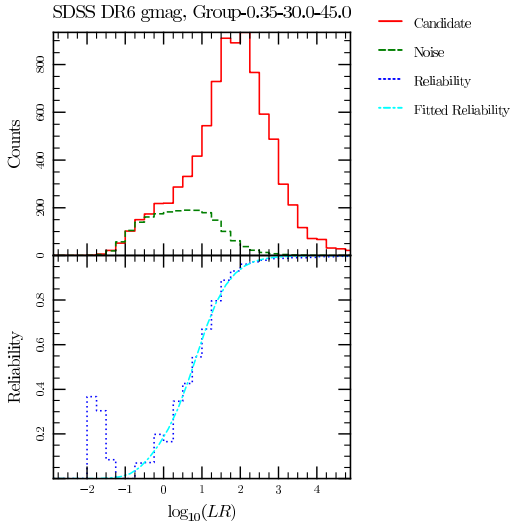


FIGURE 1. Top: Histograms of the number of associations and of the estimate of the number of spurious associations by bin of $\log_{10}(LR)$. Bottom: Reliability histogram by bin of $\log_{10}(LR)$ and its fitted curve. In the example, LRs have been computed according to the DR6 *g* magnitude for XMM sources having a systematic error of 0.35 *arcsec* with a galactic latitude $30° < |b| < 45°$.

least as bright as m over the surface of the convolution ellipse $S_{x,o} = k_\gamma^2 \pi \sigma_M \sigma_m$.

- The probability that the association has a LR in a bin ΔLR is then given by the integral of the distribution of optical sources at least as bright as m over the surface the bin occupies inside the convolution ellipse $S_{x,o}$: $S_{x,o}(\Delta LR)$.
- The sum of these probabilities for all X-ray and archival sources provides an estimate of the number of spurious associations by LR bin

Assuming that we have N X-ray and M optical sources in the FOV leads to the formula:

$$N_{spur}(\Delta LR) = \sum_{x=1}^{N} \int\int_{S_{FOV}} g_X(\alpha_x, \delta_x) \left(\sum_{o=1}^{M} \int\int_{S_{x,o}(\Delta LR)} g_o(\alpha_o, \delta_o, m) \mathrm{d}\alpha_o \mathrm{d}\delta_o \right) \mathrm{d}\alpha_x \mathrm{d}\delta_x \tag{1}$$

Where $S_{x,o}(\Delta LR)$ depends on the optical source, the X-ray source and its position: $S_{x,o}(\Delta LR, x, o, \alpha_x, \delta_x)$. To avoid a time consuming computation, one can assume that optical sources are uniformly distributed in the FOV and have as local density the mean value of their local densities. This less rigorous but acceptable approach leads to the simple formula:

$$N_{spur}(\Delta LR) = \sum_{x=1}^{N} \sum_{o=1}^{M} S_{x,o}(\Delta LR) \,/ S_{FOV} \tag{2}$$

SUPERVISED CLASSIFICATION

Learning sample

We built several samples of identified 2XMMi sources in order to investigate how the positions of the main classes of astrophysical objects are distributed into the 2XMMi-DR6 and the 2XMMi-DR6-2MASS parameter spaces. The main classes considered were stars, galaxies, QSOs, type II QSOs (QSO2), high redshift QSOs (HizQSO) and cataclysmic variables (CV).

The DR6 database, accessible via CasJob[2], provides optical spectra for a part of the SDSS photometric catalogue entries. Our QSO sample was readily extracted from the spectroscopic SDSS catalogue as was the galaxy sample. This latter sample was enriched by correlated entries in the Uppsala catalogue [3]. The cataclysmic variables sample has been extracted from the literature [4] and from the correlation of the 2XMMi, the DR6 and the Downes catalogue of CV [5].

Many of the stars detected in X-rays have SDSS magnitudes brighter than 15 mag and are thus saturated, precluding the use of the Tycho 2 catalogue. Furthermore, stars spectroscopically identified in the SDSS are so far limited to very late M types out of which only 20 are detected in X-rays. We thus enlarged our stellar sample by selecting in the colour – colour diagram unresolved objects lying close to the locus of main sequence stars. For that purpose, we used a kernel density classification. Local densities were estimated by a *balloon estimator* which is another type of variable bandwith kernel smoothing. Its main interest is to be faster than the *sample-point density estimator*.

Finally, we found in the literature [6] 5 QSO2 present in our sample.

PCA with Measurement Errors

The PCA is a well known and widely used technique to analyse high dimensional data. The classical algorithm is based on the eigendecomposition of the input parameters correlation matrix. Its main drawback is the sensitivity of the correlation coefficients to extreme values, i.e. to outliers since these coefficients are basically derived from the parameters estimated means and variances.

There are classical robust PCA methods, efficient for rejecting outliers, but they might reject small interesting populations in our case. We thus propose a method, estimating the means, variances and correlation coefficients of parameters taking into account individual measurement errors on each data.

The weighted mean squares ponderate each point i by a weight $w_i = 1/\sigma_{\varepsilon_i}^2$, defined as the inverse of the measurement errors (variances). However, in practice, a small number of points with low measurement error can account for a large part of the total weight, biasing the estimated mean of a parameter $\gamma = \sum_{i=1}^{n} w_i x_i / \sum_{j=1}^{n} w_j$. Because the total variance

[2] `http://casjobs.sdss.org/CasJobs/`

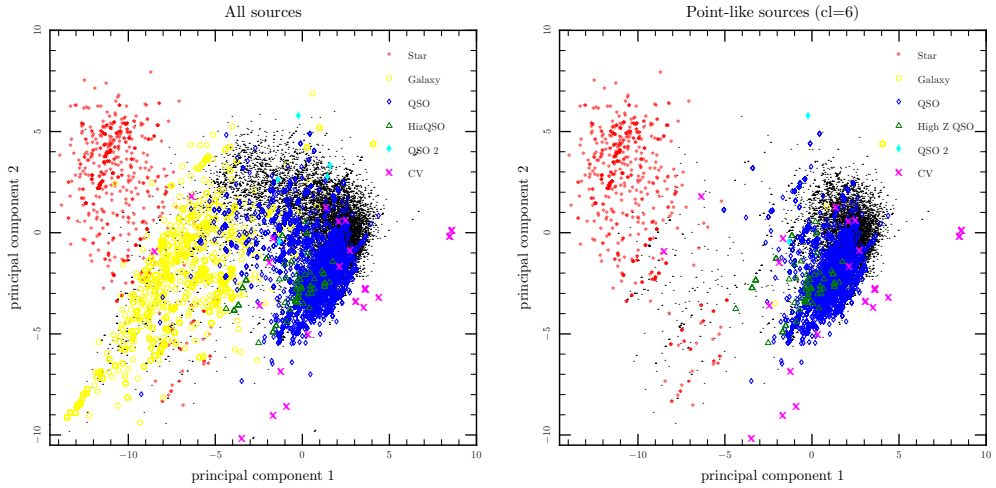


FIGURE 2. PCA on 24 parameters from the correlation of the 2XMMi with the DR6

for a parameter p depends on the unknown intrinsic scatter of this parameter $\sigma^2_{.,p}$ and the measurement errors, we use for each point i the total variance $\sigma^2_{i,p} = \sigma^2_{\varepsilon_{i,p}} + \sigma^2_{.,p}$.

We use the fact that the total variance of centred reduced data is equal to one to estimate $\sigma^2_{.,p}$ from: $\frac{1}{n}\sum_{i=1}^{n}\frac{(x_{i,p}-\gamma_{.,p})^2}{\sigma^2_{\varepsilon_{i,p}}+\sigma^2_{.,p}} = 1$. In practice, we look at the zero points of Eq. 3.

$$f(\sigma_{.,p}) = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_{i,p}-\gamma_{.,p})^2}{\sigma^2_{\varepsilon_{i,p}}+\sigma^2_{.,p}} - 1 \tag{3}$$

Where the mean is the weighted mean square estimated mean:

$$\gamma_{.,p} = \sum_{i=1}^{n}\frac{x_i}{\sigma^2_{\varepsilon_{i,p}}+\sigma^2_{.,p}} \Big/ \sum_{j=1}^{n}\frac{1}{\sigma^2_{\varepsilon_{i,p}}+\sigma^2_{.,p}} \tag{4}$$

The correlation coefficient between two parameters p_1 and p_2 is given by the covariance of their centre-reduced values:

$$\rho_{p_1,p_2} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_{i,p_1}-\gamma_{.,p_1})(x_{i,p_2}-\gamma_{.,p_2})}{\sqrt{(\sigma^2_{\varepsilon_{i,p_1}}+\sigma^2_{.,p_1})(\sigma^2_{\varepsilon_{i,p_2}}+\sigma^2_{.,p_2})}} \tag{5}$$

This method allows to use heteroscedastic data in the sense of:

- The error variance (measurement error) is different for each measured value
- We can use sets of data with very different typical errors: e.g. coming from different surveys
- The measurement errors are correlated to the measured values

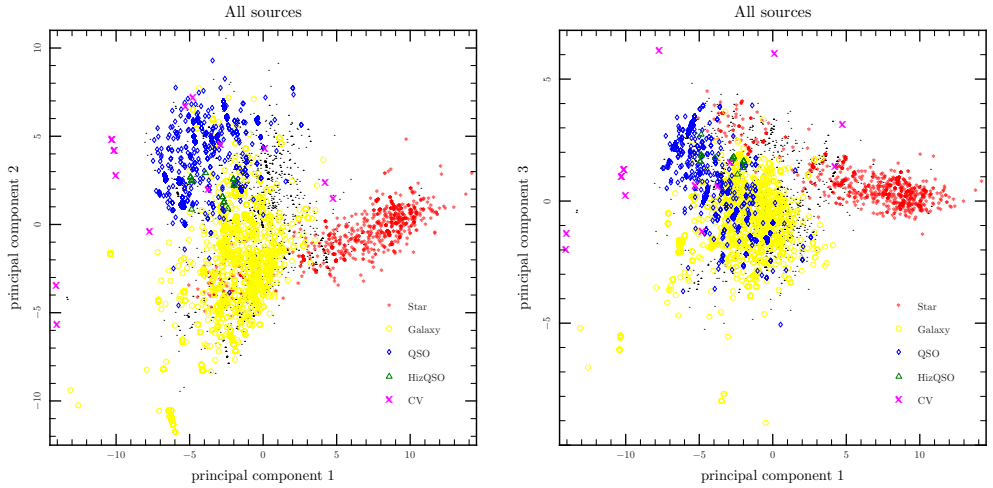


FIGURE 3. PCA on 40 parameters from the correlation of the 2XMMi with the DR6 and the 2MASS

In the last case, the estimator of the mean becomes biased. If the dependence between a value and its associated error is linear ($\sigma_x = ax + b$) we suggest to transform the parameter into a new parameter which is not correlated with its errors: $x' = \frac{\ln(ax+b)}{a}$.

Remarks:

- If individual errors are null, this reduces to the classical formulae of the mean and correlation.
- If the intrinsic scatters of the parameters are null, this reduces to the classical weighted mean square formula.
- If the distribution of the variances follows a Gaussian law of mean λ_{σ^2}, the total variance is $\sigma_p^2 \approx \lambda_{\sigma^2} + \sigma_{.,p}^2$

The main interest of our PCA is to provide a robust method which allows the use of data points with large errors without considering them as outliers.

Application of PCA, and classification

Fig. 2 shows the two first principal components of a PCA on 24 parameters from the correlation of the 2XMMi with the DR6. Using the qualitative SDSS parameter *cl* we can clearly separate the stars from the QSO and from the galaxies. The overlap between the QSO and the galaxies loci is not unexpected since the boundary between the two types is not always clear. Type II QSOs and CVs seem to form, in the three first principal components space (not shown here), a shell around the QSO locus with QSO2s in the top part and CVs in the bottom part of the locus.

Fig. 3, adding the 2MASS information allows, in the third first principal components space, the separation between QSOs, stars and galaxies without the use of the *cl* attribute.

TABLE 1. Assigned class percentage (*Class*, vertically) as a function of true class in the learning sample (*True*, horizontally), using three different methods: sum of all individual class probabilities obtained with KNN (left), with a KC (middle), and ratio of best scored objects for KC (right).

	KNN proba				KC proba				KC class			
True \ Class	Star	Gal	QSO	CV	Star	Gal	QSO	CV	Star	Gal	QSO	CV
Star	97.6	2.4	0.0	0.0	80.0	12.3	6.3	1.4	97.3	2.7	0.0	0.0
Galaxie	2.2	90.5	7.2	0.1	18.9	61.5	17.7	1.9	0.7	94.6	4.7	0.0
QSO	<0.1	15.4	83.7	0.8	1.0	26.7	60.3	3.0	0.0	14.5	85.5	0.0
CV	8.8	24.2	56.0	11.0	23.1	27.8	43.1	6.0	16.7	16.7	66.6	0.0

After reducing the dimension of the parameter space to the first three principal components, we have compared a knn classification with k=7 to a kernel density classification (KC, [7]). Table 1 lists the results of a self-check classification on the 2XMMi-DR6-2MASS learning samples.

CONCLUSION AND PROSPECTS

We have performed the cross-identification of 2XMMi with reference surveys. Using reference samples, we were able to test different classification methods after reducing the dimension of the parameter space from the original 24 or 40 dimensions using an original PCA method. We take into account both errors on positions for cross-identification, and individual measurement errors for the PCA. The first three principal components of the PCA offer a space where a good separation of stars, galaxies and QSOs is achieved.

Simple classification methods provide good results for these classes corresponding to well populated reference samples. For objects like CVs or HizQSOs, while their loci in the PCA space are reasonably well disjoint from other classes, the fact that they are scattered and the small size of the samples prevent a good classification with KNN or KC methods.

Other techniques, like kernel PCA or clustering might improve the results. But we already have detected some interesting outliers in the learning sample. And the method will be ultimately used to provide a statistical classification of other 2XMMi sources, with associated probabilities.

REFERENCES

1. H. R. De Ruiter, H. C. Arp, and A. G. Willis, *A&AS* **28**, 211–293 (1977).
2. S. Oyabu, M. S. Yun, T. Murayama, D. B. Sanders, K. Kawara, Y. Taniguchi, S. Veilleux, H. Okuda, H. Matsuhara, L. L. Cowie, Y. Sato, K. Wakamatsu, and Y. Sofue, *AJ* **130**, 2019–2042 (2005), arXiv:astro-ph/0507240.
3. A. Lauberts, *ESO/Uppsala survey of the ESO(B) atlas*, Garching: ESO, 1982.
4. M. Dillon, B. T. Gänsicke, A. Aungwerojwit, P. Rodríguez-Gil, T. R. Marsh, S. C. C. Barros, P. Szkody, S. Brady, T. Krajci, and A. Oksanen, *MNRAS* **386**, 1568–1578 (2008), arXiv:0803.0010.
5. R. A. Downes, R. F. Webbink, M. M. Shara, H. Ritter, U. Kolb, and H. W. Duerbeck, *PASP* **113**, 764–768 (2001), arXiv:astro-ph/0102302.
6. Zakamska, and al., *AJ* **126**, 2125–2144 (2003), arXiv:astro-ph/0309551.
7. Richards, and al., *ApJS* **155**, 257–269 (2004), arXiv:astro-ph/0408505.



Annexe G

Proceeding EURO-VO 2008

MULTI-WAVELENGTH OBSERVATIONS OF HIGH ENERGY SOURCES

C. Motch[1] **and F.-X. Pineau**[1]

[1]CNRS, Université de Strasbourg, Observatoire Astronomique, 11 rue de l'Université, 67000 Strasbourg, France

ABSTRACT

Gathering large sets of clean multi-wavelength measurements on specific classes of astrophysical objects or conversely, unveiling rare outliers, both first require a careful handling of the cross-identification process, taking into account positional errors and probabilities of spurious matches. Most high energy telescopes offer significantly lower spatial resolution than those operating at optical or infrared wavelengths. This can lead to confusion issues particularly in regions of high optical and infrared source densities such as in the Galactic plane. We review here the elementary steps leading to a statistically controlled identification and classification of X-ray sources and illustrate these methods and related problems with some examples extracted from the work carried out by the Survey Science Center of the XMM-Newton satellite. We then list some functionality and interoperability requirements on a few basic VO tools which would ease the proper identification and subsequent classification of high energy sources and, in general, facilitate the computation of the reliabilities of the cross-identifications of sources in pairs of catalogues collected at any wavelengths.

Key words: X-ray astronomy; Virtual Observatory.

1. INTRODUCTION

The astronomical observatories currently in operation collect every day large amounts of data over a wide range of the electromagnetic spectrum. Thanks to their good spatial resolutions, sensitivities and capabilities to cover increasingly larger regions of the sky, these observatories produce catalogues containing unprecedented numbers of sources. Future ground based or space born facilities such as ALMA, JWST, GAIA, the GSMT and the E-ELT, LSST, e-Rosita and IXO will return huge catalogues of sources from the radio to the hard X-ray domain. Multi-wavelength analysis will be a key approach to study these enormous volumes of data. In this respect, the Vitual Observatory will certainly play a very important role in offering transparent access to this large sets of heterogeneous data. However, the parallel increase of the spatial resolution and sensitivity of X-ray telescopes leading to larger source densities and, at the same time, fainter optical counterparts, does not alleviate the confusion issue. Constructing in an automatized manner the spectral energy distribution of a large number of sources requires efficient tools able to cross-identify catalogues entries detected at various wavelengths with varying spatial resolution and positioning accuracy and capable to sort out the associations according to their likelihood.

The identification of X-ray sources in the radio, IR, optical, etc.., has a high scientific value for understanding the emission mechanism, evolutionary status and how the high energy source influences its surroundings. For instance, most stars possess extended coronae that are believed to be the locus of X-ray emission. Optical spectroscopy is fundamental in assessing the influence of rotation, age and binarity on the strength of X-ray active coronae. The way accreting binaries transfer matter as well as the understanding of their evolution has greatly benefited from optical follow-up which revealed the presence of accretion discs and were instrumental in measuring the masses of the components and revealing the existence of stellar mass black-holes. In the extragalactic domain too, X-ray have proved to be the best tools to unveil the accretion power of central supermassive blackholes. Follow-up optical and infrared observations were important to better understand the viewing conditions leading to the high photo electric absorption occurring in many instances.

The goal of the present short review is to highlight the importance of having a statistically controlled cross-correlation mechanism pluggable in the VO. The first section sketches the performances of existing and future observatories putting emphasis on the X-ray domain and describes possible source confusion issues. The following sections outline the source identification and classification processes giving some illustrative examples extracted from our current efforts to identify XMM-Newton sources. The last part of this paper summarizes the facilities nowadays available in the VO and proposes some requirements on elementary VO tools which would help the cross-identification of current and future catalogues of sources.

2. CURRENT AND PLANNED FACILITIES

We list in Table 1 an incomplete, albeit probably representative, list of what are and will be in the near future the major providers of X-ray astronomical data. The exquisite spatial resolution of Chandra allows easy identification of counterparts in a large range of Galactic latitude and sensitivity regimes. However, even with a $\sim 0.6''$ radius error circle, identifying the right counterpart of a Chandra X-ray source can become challenging in the direction of the Galactic bulge and Galactic center where the number of optical and infrared sources rises to paramount densities (see e.g. Koenig et al., 2008; Bandyopadhyay et al., 2005). Likewise, the high sensitivity exposures obtained in the Chandra deep fields for instance (Szokoly et al., 2004) can reveal very low luminosity X-ray sources whose counterparts exhibit ambiguous signatures at other wavelengths. Compared to optical, X-ray images have in general much lower spatial resolution, often variable on the field of view. This can lead to severe difficulties to properly identify X-ray sources.

Future large X-ray observatories such as IXO will reach in the soft 0.5-2.0 keV fluxes of the order of a few 10^{-18}erg cm^{-2} s^{-1}. Extrapolating the source number counts observed in the Chandra deep fields down to these faint fluxes, Bauer et al. (2004) find that at these soft sensitivities, the density of active galactic nuclei will be of $\sim 2\times\ 10^4$ deg^{-2} and be comparable to that of X-ray emitting star forming galaxies. The density of faint optical candidates as bright as the actual X-ray source counterpart becomes high enough to render optical identification ambiguous even with a $1''$ radius error circle. In addition X-ray source confusion can severely limit the final X-ray sensitivity reached if the half-energy width is of the order of $\sim 5''$ (Hasinger et al., 2006).

3. THE PATH TO X-RAY SOURCE IDENTIFICATION AND CLASSIFICATION

In general, the determination of the nature of an X-ray source (or any other source detected in a given wavelength range) proceeds in two distinct steps:

- The first one, called 'identification' aims at finding the counterpart of the X-ray source in an image or catalogue obtained in the optical, infrared, etc. domain. This process relies on the coincidence in position (cross-correlation) and can be extended using a priori considerations on the nature of the astrophysical object in order to solve cases where several candidates are positionally consistent. It is highly advisable that the cross-correlation process provide probabilities of identification.
- The second step, named 'classification' consists in determining the true nature of the X-ray source (star, galaxy, AGN, etc.), once counterparts at various wavelengths are identified (first step). In optimal cases, X-ray or optical spectra of the source are available from which the object can unambiguously be classified. However, finding the nature of large amounts of X-ray sources requires the use of classification methods relying on less informative observational data. The strategy adopted by most approaches is:

  1. Gather a "learning sample" i.e. a group of sources for which the nature and the observable quantities are perfectly known. Such a sample can be made of theoretical spectral energy distributions for instance, or alternatively from subsets of X-ray sources, spectroscopically identified and properly classified and preferentially covering the range of parameters (flux, colours, galactic latitudes, etc.) of the catalogues used for cross-identification and classification.
  2. Find where in the parameter space sources of different nature cluster.
  3. Classify sources by measuring their "distances" (e.g. knn) to the position of the learning sample. Depending on the algorithm, the result of the classification can be either Boolean or associated to some probability.

4. CROSS-CORRELATION METHODS

Several "quick and dirty" tools (e.g. collecting all cross-matches within 5" from X-ray position) are readily available as web services or as standalone programs (see section 7). However, these facilities do not provide the probability that the source in one catalogue is truly the same as that present in the other catalogue, taking into account positional errors and the chance of spurious matches. Computing identification probabilities is important for several purposes:

- Assessing the relative merits of multiple candidates present in the X-ray error circle.
- Building homogeneous source samples with known "purity" and "completeness" in order to control the statistical properties of the selected sample.
- Determining the expected fraction of false outliers.
- Making subsequent proper source classification.

4.1. The likelihood ratio method

We recall below the rudiments of the method commonly used to compute the probabilities of cross-identification of two catalogue entries (here an X-ray source and an entry in an archival astronomical catalogue). This problem has already been addressed by many authors in various

Table 1. Main properties of currently operating and planned X-ray observatories

Project	Dates	Effective area (cm^2)	Energy range (keV)	Energy resolution	Pos accuracy (arcsec)
XMM-Newton	1999 -	3000	0.2 - 12	60 - 800	2
Chandra	1999-	600	0.1 - 10	60 - 1000	0.6
Integral	2002 -	3000	3 - 10 MeV	12	60
e-Rosita	2011 -	1000	0.2 - 12	50	5
Simbol-X	2014 -	1000	0.5 - 80	50	3
IXO	2018 -	50,000	0.5 - 40	50 - 3000	2

astrophysical contexts (see e.g. De Ruiter et al. (1977); Boller et al. (1998); Rutledge et al. (2000); Oyabu et al. (2005)).

Let us consider an X-ray source having only one candidate identification in the archival catalogue at a given distance. Two hypotheses can be considered: H_{cp} : the candidate is the counterpart of the X-ray source and H_{spur} : the candidate is a spurious match.

We assume that errors on both X-ray and archival positions are distributed as Gaussian and expressed in units normalized by the *rms* of the difference in position. Under the assumption that the archival source is the true counterpart of the X-ray source, the probability that the two positions fall between r and $r + \mathrm{d}r$ can be written as

$$P(r|H_{cp}) = re^{-\frac{1}{2}r^2} \tag{1}$$

The probability of finding a spurious object between r and $r + \mathrm{d}r$ is given by the Poisson law where λ is the estimated local density of archival sources:

$$P(r|H_{spur}) = 2\lambda r \tag{2}$$

The relative merit of the two hypotheses is usually described by the likelihood ratio (LR) (De Ruiter et al., 1977) which is the ratio of the two probability densities (1) and (2):

$$LR(r) = \frac{P(r|H_{cp})}{P(r|H_{spur})} = \frac{1}{2\lambda}e^{-\frac{1}{2}r^2} \tag{3}$$

What we actually need is the reliability of the identification which in Bayesian formalism can be written as $P(H_{cp}|r)$, namely the probability that the optical candidate is the true counterpart of the X-ray source if found at distance r. Using Bayes' theorem, one can express this probability as

$$P(H_{cp}|r) = \frac{P(r|H_{cp}) \times P(H_{cp})}{P(r)} \tag{4}$$

where $P(r)$, the probability to find a candidate at distance r from the X-ray source whatever is its nature, can be written as: $P(r) = P(r|H_{cp}) \times P(H_{cp}) + P(r|H_{spur}) \times P(H_{spur})$. The reliability of the identification can thus be expressed using the likelihood ratio:

$$R(LR) = P(H_{cp}|r) = \frac{1}{1 + \frac{P(H_{spur})}{P(H_{cp})}\frac{1}{LR}} \tag{5}$$

Following (Oyabu et al., 2005) we can also write the reliability as

$$R(LR) = \frac{N_{all}(LR) - N_{spur}(LR)}{N_{all}(LR)} \tag{6}$$

where $N_{all}(LR)$ is the total number of matches per bin of LR histogram and $N_{spur}(LR)$ the number of spurious matches in the same bin. Basically two methods are proposed to estimate $N_{spur}(LR)$. The first one (Monte Carlo) consists in moving randomly the positions of the X-ray sources and accumulating LR histograms representative of the distribution of spurious matches (Rutledge et al., 2000; Oyabu et al., 2005). Alternatively, one can apply the method developed by Pineau et al. (2008a) which is based on a direct computation of the sky area related to each LR bin. This approach has the great advantage of avoiding time-consuming Monte Carlo trials. The formalism above can be extended to the case in which several candidates are possible for a given X-ray source.

The resulting $R(LR)$ histogram can then be fitted by equation (5) to directly obtain the $P(H_{spur})/P(H_{cp})$ ratio. We show in Figure 1 an example of such a computation extracted from Pineau et al. (2008b). Hence, the reliability of a given association depends on LR and on the fraction of X-ray sources having an actual counterpart in the archival catalogue considered.

4.2. Complications

In general, the quantity $P(H_{spur})/P(H_{cp})$ depends on the relative fractions of the various X-ray populations detected and thus varies with the sensitivity and galactic latitude of the X-ray observation and on the properties (depth, wavelength range) of the archival catalogue used

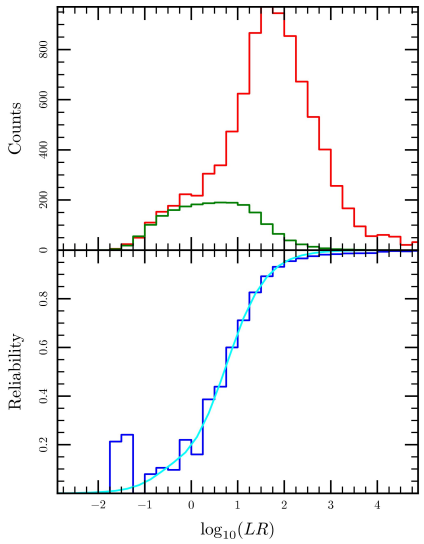


Figure 1. Upper panel; The upper red curve is the LR histogram for all associations, i.e. including both true and spurious cases. The lower green curve shows the histogram of spurious associations obtained using one of the methods described in the text. Lower panel; Histogram of the reliabilities computed from eq (6). The blue curve represents the best fit function of the theoretical $R(LR)$ relation assuming three populations.

in the cross-correlation. Two extreme cases can be considered for illustrative purposes. If most X-ray sources have properties such that they have a counterpart in the archival catalogue, then $P(H_{spur}) << P(H_{cp})$ and the reliability of any association remains high apart for very low LR. If on the contrary, the archival catalogue is not relevant to the X-ray population observed, i.e. only contains a small fraction of the counterparts of the X-ray sources, the reliability will remain low, even for very high LR. In practice, a single function of form (5) cannot properly fit the estimated $R(LR)$ relation. For instance, Pineau et al. (2008b) need to add three distinct populations i.e. three different $P(H_{spur})/P(H_{cp})$ values to obtain a good representation of the observed $R(LR)$.

So far, the method described above provides probabilities of identification using as sole criterion the coincidence in position. However, the spectral energy distribution or more specifically $fx/fopt$ ratios or X-ray and optical "colours" can provide additional criteria to judge of the reliability of a given association between an X-ray source and a candidate counterpart. This approach has been used for example in Sutherland & Saunders (1992) and Brusa et al. (2007). These authors weight the quantity $P(H_{spur})/P(H_{cp})$ with a factor $q(p)$ which describes the change of the fraction of X-ray sources having a counterpart in the archival catalogue with parameter p which can be for instance a magnitude, a flux ratio or a colour index.

Using this method introduces a Bayesian 'prior' which will automatically assign low identification probabilities to X-ray sources exhibiting extreme properties. Hence, if the criterion of positional coincidence can by itself identify a large enough fraction of the X-ray sources, it may not be advisable to use additional criteria in order to efficiently detect small groups of extreme objects and broadly speaking of outliers.

5. THE CLASSIFICATION PROCESS

As mentioned above, the classification process first requires building a "learning sample" which should preferentially cover as much as possible the parameter space spanned by the sources to classify. In many cases, the archival catalogues containing precise information on the nature of the sources also have flux limitations arising from the need to acquire optical spectra of good enough signal to noise ratio. Consequently, they usually cover only the brightest part of the sample, thus preventing from using them to classify the bulk of the fainter sources. Dedicated deep spectroscopic observations can be instrumental in gathering large enough groups of identified sources better covering the observed parameter space and providing a still reasonably large number of identified objects even for scarcely populated source classes.

5.1. Principal Component Analysis

The Principal Component Analysis (PCA) is widely used to analyse high dimensional data. PCA is a useful tool to reduce the dimensionality of the parameter space by finding redundancies. First principal components contain the largest part of the total variance of the dataset and thus highlight the most important combination of physical parameters for source classification. However, the classical PCA algorithm does not take into account measurement errors and when applied to astronomical data, develops instabilities due to its sensitivity to extreme points. In order to cope with data plagued by measurement uncertainties, robust PCA exist, which however in most cases only clip out outliers, still ignoring their statistical weight. In this respect, the new PCA algorithm developed by Pineau et al. (2008b) is an interesting step forward since this method accounts for both measurement errors and for the intrinsic scatter of the physical parameters.

5.2. Classification methods

The community uses a wide range of classification algorithms. Among the supervised methods one can mention decision trees (e.g. the OC1 algorithm, Murthy et al., 1994), neural networks, support vector machines (see e.g. Elting et al., 2008) and kernel density classification (Richards et al., 2004). Each method has its advantages and disadvantages but the outcome mainly depends on the quality and representativeness of the learning sample.

5.3. An example: the cross-correlation of the 2XMMi catalogue with SDSS R6

We illustrate the cross-correlation, identification and classification steps by showing below some results of the cross-correlation performed by Pineau et al. (2008a) of the 2XMMi catalogue with the 6^{th} release of the Sloan Digital Sky Survey (SDSS R6).

The incremental second catalogue of serendipitous X-ray sources from the European Space Agency's (ESA) XMM-Newton observatory results from a major reprocessing of the XMM-Newton archive carried out between 2006 and 2008 by the Survey Science Center (SSC) of the XMM-Newton satellite (Watson et al., 2009). A total of 289 083 X-ray detections (221 012 unique sources) extracted from 4 117 observations enter the 2XMMi catalogue. The total area covered is $\sim$ 560 deg^2 ($\sim$ 420 deg^2 taking into account overlaps, i.e. about 1% of the sky). Each X-ray source comes with a number of parameters such as hardness ratios and fluxes in various energy bands. The sensitivity of the catalogue is of a few 10^{-15} erg s^{-1} cm^{-2} in the soft (0.5 - 2.0 keV) band and $\sim 10^{-14}$ erg s^{-1} cm^{-2} in the hard (2.0 - 12.0 keV) band.

The SDSS covers $\sim$ 8500 deg^2, mostly at high Galactic latitude and contains $\sim$ 287 000 000 unique sources with astrometric errors below $0.1''$ rms (Adelman-McCarthy et al., 2008). Five photometric bands u, g, r, i and z are measured with limiting magnitudes in the range of 20.5 to 22.2 at the 3% error level.

Saturated and blended SDSS objects were discarded as were optical sources with errors on individual magnitudes larger than 0.03. Only unresolved X-ray sources were considered. A 'high reliability' sample of 20,050 associations having a probability of identification larger than a stringent value of 95% (corresponding to a sample purity of 99.3%) was extracted from the cross-correlation process. As a consequence of the small rate of wrong associations required, the corresponding sample completeness is only 58.4% (see Figure 2).

The spectroscopic SDSS catalogue offers a well suited learning sample for this study. However, the strategy applied to the selection of photometric targets for spectroscopic observations leads to a number of biases which at some stage must be compensated for. In particular, the sample of identified active stars, cataclysmic variables (CVs) and galaxies had to be enlarged using cross-correlation with additional catalogues (see Pineau et al. (2008b) for details). In addition, a group of type II QSO candidates was added (Zakamska et al., 2003).

We show in Figure 3, the distribution of the total 2XMMi/SDSS R6 cross-match (photometric sample) and learning samples in the $(g-i)$ / $log_{10}(f_x/f_r)$ physical parameter spaces and in the first two principal component axes derived from the PCA analysis. The PCA naturally discovers the best physical quantities to separate classes. Indeed, the first two PCA axes (those containing the highest data variance) are a mixture of optical

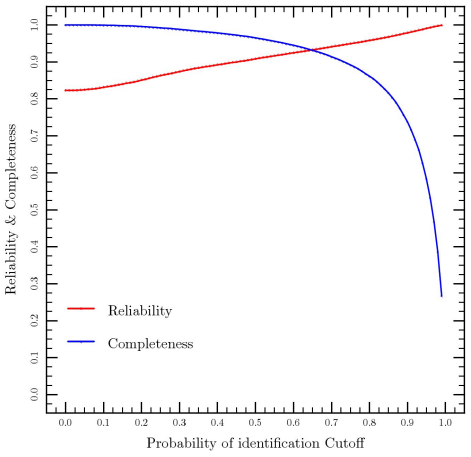


Figure 2. Integrated sample reliability and purity curves as a function of the threshold imposed on the individual source identification probability, applicable to the entire 2XMMi / SDSS R6 cross-correlation.

colours on one side and flux ratios and magnitude on the other side with an additional smaller contribution of the X-ray hardness ratios. Considering the third PCA axis allows to see that CVs form a kind of shell around the main AGN locus and can thus be separated. Not unexpectedly, the spectroscopic SDSS sample is too shallow to cover the range of (f_x/f_r) exhibited by the photometric sample. Spectroscopic identification of higher (f_x/f_r) objects (mainly fainter optical objects) are missing as are those located in the extension of the 'galaxy' branch toward larger (f_x/f_r) ratios. This parameter region is likely to contain many spatially resolved galaxies harboring a moderately luminous nucleus or more exotic objects such as type II QSOs. In both diagrams, a part of the X-ray active stars overlaps with X-ray emitting galaxies. Again, this is not unexpected since these two classes of objects share to some extent similar optical colours and flux ratios. They also exhibit comparable X-ray spectra at least when only few X-ray counts are collected. Taking into account the extension of the optical object obviously removes any ambiguity. Interestingly, Pineau et al. (2008b) have shown that introducing near infrared information from the 2MASS catalogue also allows to separate well stars from galaxies in the absence of information on the extent of the optical object.

A kernel density classification has been tested. It is a powerful tool to classify unidentified sources which are in the range of the learning sample. This Bayesian approach provides for each source the probabilities to belong to the various identified classes. Cross-checking with the learning sample leads to very good results for stars, galaxies and QSOs.

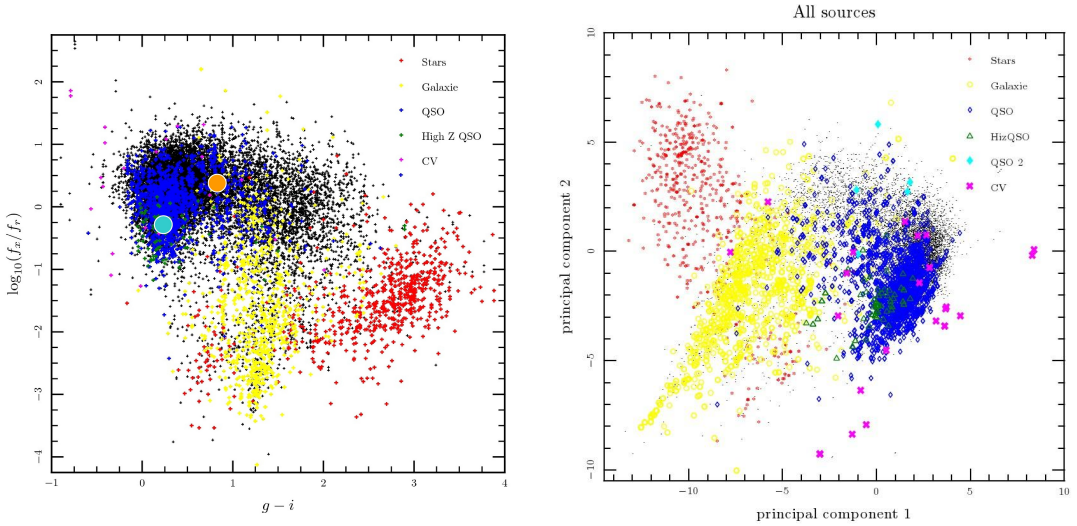


Figure 3. 2XMMi/SDSS R6 cross-correlation. Source distribution in two physical parameter spaces (left panel) and in the first and second main axes found by the PCA analysis (right panel). Black dots: 2XMMi sources having a photometric identification. Learning sample; blue - AGNs; red - stars; yellow - galaxies (extended objects); magenta Type 2 QSO candidates; purple - CVs. On the left panel, the blue and orange dots represent the position of the two SDSS candidates for the source discussed in section 6.

6. USING SPECTRAL ENERGY DISTRIBUTION AS FURTHER IDENTIFICATION CRITERION

Taking into account the expected energy distribution of the X-ray source and of candidate objects might substantially help to find the right counterpart, although as discussed in section 4, such an approach may have the drawback of rejecting interesting outliers. We illustrate this with two simple examples, one extracted from the work of Pineau et al. (2008a,b) briefly outlined in the previous section and a second case extracted from the identification campaign carried out by the SSC in the Galactic Plane.

We show in Figure 4 the SDSS finding chart of source 2XMM J155205.2+552351. Two SDSS objects are present in the 90% confidence radius error circle of the 2XMM source. The one closest to the X-ray position is extended, relatively faint (g = 20.23) and reddish ($g-i$ = 0.9; $u-r$ = 1.7). The object located close to the error circle is spatially unresolved, slightly brighter (g = 17.88) and much bluer ($g-i$ = 0.31; $u-r$ = 0.28). Although its probability of identification based on position (18%) is lower than that of the red object (80%), the blue object lies close to the main AGN locus (see Figure 3) while the red object is located in a parameter region containing significantly less photometric identifications of X-ray sources. SDSS R6 spectroscopy reveals that the blue object is a broad line AGN at z = 1.479 and is therefore the likely optical counterpart of the source in spite of its lower probability. Follow-up spectroscopy of the red galaxy would be needed to rule out any additional contribution from this object.

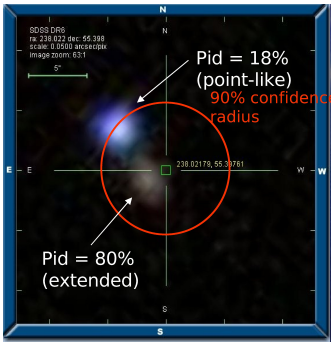


Figure 4. The SDSS R6 finding chart for source 2XMM J155205.2+552351

Figure 5 shows the position of source XGPS_3 (Motch & al, 2009) ($b \sim 0°$ $l \sim 20°$) overlaid on field images acquired in the V band (0.55μ) up to the Spitzer 5.8μ band. The X-ray source undergoes strong photoelectric absorption of mainly interstellar origin. While in the R band, the probability of identification with the candidate star closest to X-ray position is only 72% (and would be even lower in the V band), it reaches 99% in the 2MASS K band. At 5.6μ, the counterpart of the X-ray source is so bright that it saturates the detector. A spectral energy distribution was built from the GSC2.2 and 2MASS magnitudes (see Figure 6). It is well fitted by the Kurucz model atmosphere (Castelli & Kurucz, 2003) of a B0V star reddened with E(B-V) $\sim$ 7. Follow-up optical spectroscopy confirms that the counterpart of the X-ray source is a likely Be/X-ray binary undergoing strong interstellar

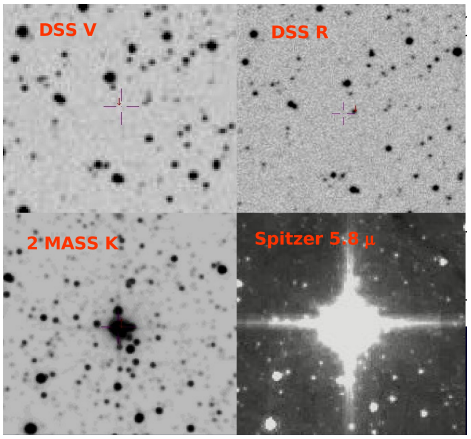


Figure 5. Finding chart for XGPS_3 in the Galactic plane. The X-ray position is marked by an arrow.

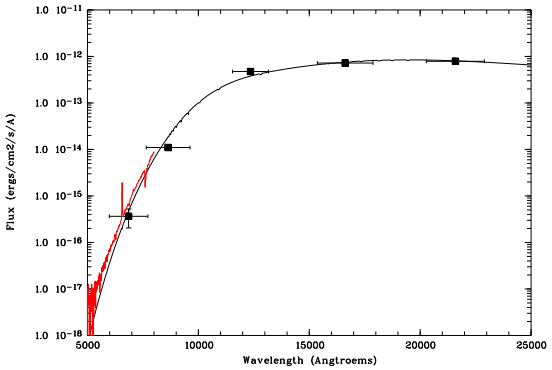


Figure 6. The energy distribution of the counterpart of XGPS_3. Dots and error bars represent broad band photometric data (GSC2.2, 2MASS). The continuous line is a best fit Kurucz model atmosphere of a B0V star with E(B-V) = 7. The observed optical spectrum is in red.

absorption.

7. EXISTING CROSS-MATCHING VO TOOLS

A number of VO services and VO-aware cross-match tools are available. We list below the facilities known to us, keeping in mind that we did not aim to be complete.

- The Russian VO offers cross-matches within a fixed radius for a preset list of eight of the major all-sky catalogues in addition to the 2XMMp and SDSS R5 and R6 catalogues[1]
- The Chinese VO has developed a cross-match tool which is realized by a web service. Cross-correlation can be done between a catalogue on the server side and one provided by the user. Further details can be found in Gao et al. (2008).
- The US National Virtual Observatory provides through the OpenSkyQuery facility both a simple and an advanced interface[2]. Several catalogues can be cross-matched together. This tool has two interesting features rarely offered, namely, the possibility to specify a cross-correlation radius expressed in terms of combined positional errors, and the possibility to enforce the absence of cross-correlation in a catalogue. It is also possible to put additional constraints on the values of any catalogue attribute. The facility uses a dedicated function XMATCH, an extension of the SQL language, which handles the computation of the combined positional errors.
- Several other cross-matching tools are listed on Astrogrid[3]. VO Clients such as TopCat and Aladin provide basic cross-match tools within a fixed radius. Aladin can also take into account error ellipses and/or non-matches. The Starlink Tables Infrastructure Library Tool Set (STILTS) task *tmatchn* allows multiple cross-matching with several possible criteria. Finally the ADQL language could also in principle allow cross-correlation taking into account complex relations involving constraints on positional errors and values of the catalogue attributes.

[1] http://vo.astronet.ru/cas/crossmatch.php
[2] http://openskyquery.net/Sky/SkySite/OSQForm/default.aspx
[3] http://www.astrogrid.org/wiki/Help/UsageExamples/CrossMatch

8. POSSIBLE REQUIREMENTS ON VO CROSS-IDENTIFICATION TOOLS

Source classification is a complex process which has to be tuned to the way the learning sample covers the properties of the group of sources to be classified. Furthermore, several algorithms may be applied according to the manner the various classes of sources cluster in the multi-dimensional physical parameter space or space derived from the PCA analysis. At present, there is probably no need to implement PCA and classification tools in the VO. Classification is not a main stream astronomical activity and for the time being is better done on the client side using data acquired from the VO. On the other hand, the source identification process would probably benefit from using VO-aware tools in order to make easier some of the elementary steps listed below:

1. Cross-match with distance limits expressed in terms of positional errors. This is already possible using the NVO advanced XMATCH tool for a restricted set of catalogues and could in principle be done in a generic manner using ADQL. However, the VO user should have an easy access to the meta-data describing the positional errors (given attribute or fixed

value), their statistical significance (1σ, 90% confidence level, etc.) and the existence of any systematic error to add. Even more problematic is the way ellispoidal errors are described in meta-data and properly handled in cross-correlation tools.

2. Estimating local densities requires access to the entire archival catalogue, possible repeatedly depending on the algorithm used and would thus be a good candidate for a VO development. For this task too, a tool providing the local density of any catalogue at any given point of the sky could be developed. One of the simplest algorithms consists in a *knn* averaging. Another possibility is a kernel smoothing, e.g. the sample-point density estimator (Sain, 2002).

3. Computing the likelihood ratio histograms of false matches using a Monte Carlo approach also requires heavy access to catalogues. Although this step might simply consist of repeated runs of tool [1], a VO wrapper with internal randomization of the input positions might be considered.

4. Using Bayesian 'priors' involving the expected spectral properties of classes of sources would be made easier if there were VO tools able to create on-the-fly the SEDs of many objects. They should be able to access broad band photometry, archival spectra or theoretical models and allow reddening and redshift corrections.

9. CONCLUSIONS

Finding the correct optical or infrared counterparts of the faint high energy sources which will be discovered by the future X-ray telescopes facilities will remain a challenging task due to the improved detection sensitivity of the high energy facilities which will reveal ever fainter optical counterparts. In many cases, the detailed spectroscopic observations required to properly identify the X-ray sources will not be available and the identification of most sources will rely on positional coincidence arguments combined with considerations on broad band colours and flux ratios. In this respect, we suggest that the VO consider developing some of the elementary tools listed above relieving the end user from struggling with large amounts of data and with the details of the processing required by a proper statistical calibration of the identification probabilities. We note that such tools will be useful well beyond the mere identification of X-ray sources and can be in principle applied to the cross-identification process at any pair of wavelength range.

REFERENCES

Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., et al. 2008, ApJS, 175, 297

Bandyopadhyay, R. M., Miller-Jones, J. C. A., Blundell, K. M., et al. 2005, MNRAS, 364, 1195

Bauer, F. E., Alexander, D. M., Brandt, W. N., et al. 2004, AJ, 128, 2048

Boller, T., Bertoldi, F., Dennefeld, M., & Voges, W. 1998, A&AS, 129, 87

Brusa, M., Zamorani, G., Comastri, A., et al. 2007, ApJS, 172, 353

Castelli, F. & Kurucz, R. L. 2003, in IAU Symposium, Vol. 210, Modelling of Stellar Atmospheres, ed. N. Piskunov, W. W. Weiss, & D. F. Gray, 20P

De Ruiter, H. R., Arp, H. C., & Willis, A. G. 1977, A&AS, 28, 211

Elting, C., Bailer-Jones, C. A. L., & Smith, K. W. 2008, in American Institute of Physics Conference Series, Vol. 1082, American Institute of Physics Conference Series, 9–14

Gao, D., Zhang, Y. X., & Zhao, Y. H. 2008, Acta Astronomica Sinica, 49, 348

Hasinger, G., Arnaud, M., Barcons, X., et al. 2006, in SPIE Conference Series, Vol. 6266, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series

Koenig, X., Grindlay, J. E., van den Berg, M., et al. 2008, ApJ, 685, 463

Motch, C. & al. 2009, in preparation

Murthy, S. K., Kasif, S., & Salzberg, S. 1994, ArXiv Computer Science e-prints

Oyabu, S., Yun, M. S., Murayama, T., et al. 2005, AJ, 130, 2019

Pineau, F.-X., Derriere, S., Michel, L., & Motch, C. 2008a, in Astronomical Society of the Pacific Conference Series, Vol. 394, Astronomical Data Analysis Software and Systems XVII, ed. R. W. Argyle, P. S. Bunclark, & J. R. Lewis, 369–+

Pineau, F.-X., Derriere, S., Michel, L., & Motch, C. 2008b, in American Institute of Physics Conference Series, Vol. 1082, American Institute of Physics Conference Series, 15–21

Richards, G. T., Nichol, R. C., Gray, A. G., et al. 2004, ApJS, 155, 257

Rutledge, R. E., Brunner, R. J., Prince, T. A., & Lonsdale, C. 2000, ApJS, 131, 335

Sain, S. R. 2002, Comput. Stat. Data Anal., 39, 165

Sutherland, W. & Saunders, W. 1992, MNRAS, 259, 413

Szokoly, G. P., Bergeron, J., Hasinger, G., et al. 2004, ApJS, 155, 271

Watson, M. G., Schröder, A. C., Fyfe, D., et al. 2009, A&A, 493, 339

Zakamska, N. L., Strauss, M. A., Krolik, J. H., et al. 2003, AJ, 126, 2125



Annexe H

Article en préparation sur la corrélation 2XMMi/DR7

Astronomy & Astrophysics manuscript no. XCorrelation2XMMiSDSS˙WorkingDraft˙V3
September 11, 2009
© ESO 2009

Cross-Correlation of the 2XMMi catalogue with Data Release 7 of the Sloan Digital Sky Survey

F.-X. Pineau[1] and al[2,⋆]

[1] CNRS, Université de Strasbourg, Observatoire Astronomique, 11 rue de l'Université, 67000 Strasbourg, France
e-mail: pineau at astro.u-strasbg.fr
[2] Others, ...
e-mail: other@email.com ⋆⋆

Received, accepted

ABSTRACT

The Survey Science Center of the XMM-Newton satellite recently released the first incremental version of the 2XMM catalogue. Containing more than 220,000 X-ray sources the 2XMMi is the largest catalogue of X-ray sources ever published and thus constitutes an unprecedented resource to study the high energy properties of various classes of X-ray emitters such as AGNs and stars. Thanks to the high throughput of the EPIC cameras on board XMM-Newton accurate positions, fluxes and hardness ratios are available for a large fraction of the X-ray detections. The advent of the 7^{th} release of the Sloan Digital Sky Survey offers the opportunity to cross-match two major surveys and extend the spectral energy distribution of many 2XMMi sources toward the optical bands. This implies building large homogeneous samples having a statistically controlled rate of spurious matches and completeness. We present here a cross-matching algorithm based on the classical likelihood ratio estimator. The method developped has the advantage of providing true probabilities of identifications without resorting to heavy Monte-Carlo simulations. Over 30,000 2XMMi sources have SDSS counterparts with individual probabilities of identification larger than 90%. At this threshold, the sample has only 2% spurious matches and is 77% complete. Using spectroscopic identifications from the SDSS DR7 catalogue supplemented by extraction from other catalogues, we build a learning sample from which the way the various classes of X-ray emitters gather in the multi dimensional parameter space can be analysed and later used to design a source classification scheme. We illustrate the interest of this clean source sample by investigating some particular science cases.

Key words. To be filled later

1. Introduction

The growing collecting area and sensitivity of modern astronomical detectors combined with the increasing storage and processing capabilities offered by current computer facilities has made possible the gathering on comparatively short time scales of very large sky surveys that were beyond reach only a few years ago. Most parts of the electromagnetic spectrum benefit from this evolution. Among recently completed or ongoing projects are the Two Micron All Sky Survey (2MASS) (Cutri et al. 2003) and the Sloan Digital Sky Survey (Adelman-McCarthy et al. 2008) for instance. Space born missions currently in operation such as the Spitzer Space Telescope (Werner et al. 2004) observing in the infra-red or the Chandra (Weisskopf et al. 2000) and XMM-Newton (Jansen et al. 2001) X-ray observatories are collecting at a high rate a wealth of measurements on an unprecedented number of objects in their energy range. In the relatively near future, ground based automated very large telescopes such as the Large Synoptic Survey Telescope (Tyson 2002) will collect detailed photometric information on a breathtaking amount of faint galaxies.

Merging measurements arising from several instruments allows to build spectral energy distributions over a range of wavelengths extending over a large part of the electromagnetic spectrum. The recent availability of wide angle surveys with high detection sensitivities allows to measure with comparable accuracies and in several scientifically important wavelength ranges the spectral energy density of the main classes of X-ray emitting astrophysical sources. Building large homogeneous samples provides valuable insight on the emission mechanisms and evolutionary processes and may allow the detection of rare objects or outliers which would be otherwise hard to unveil in smaller samples. In this respect, a good estimate of the true rate of false cross-identification is important to assess the relevance of any group of outliers.

However, the gathering of large groups of sources with well characterized multi-wavelength properties first requires a proper handling of the cross-matching process between two or more catalogues. Although spatial resolution at high energy steadily increased during the last years and may go on improving in the future, source density also grows as a result of the improved sensitivity and the risk of confusion between unrelated objects detected at different wavelengths does not necessarily vanish. The confusion problem can be particularly arduous when comparing catalogues with very different spatial resolutions and densities, a problem often encountered in the identification process of high energy sources which in several cases lack the superb spatial resolution affordable in the optical domain for instance.

The XMM-Newton satellite (Jansen et al. 2001) was launched by European Space Agency late 1999. XMM-Newton is currently the X-ray (0.2-12 keV) telescope in operation having the largest effective area. Three co-aligned telescopes feed two EPIC MOS (Turner et al. 2001) and one EPIC pn (Strüder

⋆ Nothing to declare
⋆⋆ The university of heaven temporarily does not accept e-mails

et al. 2001) cameras. Two refection grating arrays deviate about half of the X-ray photons from the EPIC MOS camera toward two Reflection Grating Spectrometers (RGS; den Herder et al. 2001). An Optical Monitor (OM; Mason et al. 2001) providing UV and optical images of a fraction of the field of view covered by the EPIC cameras down to the 21^{th} magnitude complements the X-ray instrumentation. One of the remarkable properties offered by the X-ray telescopes on-board XMM-Newton is to provide a large field of view of 30 $'$ diameter with weakly degraded image point spread function and low vignetting even at large off-axis angles. Accordingly, a large number of sources may be serendipitously discovered around the main target of the observation which build up to make an X-ray survey with an unprecedented combination of sensitivity and area covered. Starting from the beginning of the project, ESA recognized the high scientific interest of exploiting the XMM-Newton survey and appointed the present Survey Science Centre (SSC) on a competitive basis. Lead by the University of Leicester, the SSC is a consortium of ten European institutes conducting its activity on behalf of ESA. The SSC responsibilities have been presented in Watson et al. (2001). One of the most demanding task devoted to the consortium is the compilation of a catalogue of all serendipitous sources discovered in the field of view of the X-ray instruments and of their characterisation and identification at least in a statistical way.

Several spectroscopic identification campaigns and multi-wavelength studies have been recently performed by the SSC on samples of thousands of EPIC sources using follow-up observations at 4-m and 8-m class telescopes. The availabilities of the recently published SDSS Data Release 7 (DR7) and of the incremental version of the 2XMM catalogue (2XMMi) offer a unique opportunities to extend to a much larger sky area the identification work. With its spectroscopic and photometric limiting magnitude about 2 magnitudes brighter than that typically reached for the SSC source samples, SDSS identifications of XMM-Newton sources conveniently extend the identified sample toward brighter magnitudes and at the same time provide access to a rich group of accurately quantified photometric and spectroscopic data.

As part of its scientific activities, the Survey Science Centre of the XMM-Newton satellite has developed a specific cross-correlation algorithm yielding actual probabilities of identification based on positional coincidence and applied this algorithm to the cross-identification of the 2XMMi and SDSS DR7 catalogues, thus creating one of the largest set of optically identified X-ray source available so far. The result of the cross-correlation is made available as a separate fits file and is also available through the XCAT-DB[1] (Motch et al. 2007).

The first sections of this paper present the details of the algorithm used to identify 2XMMi X-ray sources with SDSS DR7 optical objects. We apply the commonly used likelihood ratio to quantify the chance that a SDSS object is the counterpart of the X-ray source. Identification probabilities are computed using an original method which does not rely on Monte Carlo simulations and thus offers a better efficiency when cross-correlating large sets of data. The last part of the paper describes the range of optical and X-ray parameters occupied by the main astrophysical classes of X-ray emitters and discuss some scientific results arising from this work.

[1] http://amwdb.u-strasbg.fr/2xmmi/home

2. Description of the cross correlated catalogues

2.1. 2XMMi catalogue

The incremental Second XMM-Newton Serendipitous Source Catalogue (2XMMi) is an extended version of the 2XMM Catalogue (Watson et al. 2009). It has been built from 4117 individual pointed observations performed by the XMM-Newton Observatory and contains 289 083 heterogeneous detections for a total of 221 012 unique X-ray sources. The catalogue covers $\sim 1\%$ of the sky over a large range of Galactic latitudes and longitudes. EPIC cameras encompass a field (FOV) of $\sim 30'$ dia and are sensitive in the energy range of $\sim 0.2 - 12$ keV. Source positions have a typical accuracy of $\sim 2''$. In this paper, we limit our analysis to point-like sources having a positional error smaller or equal to $5''$. A source is defined as point-like if its extent maximum likelihood parameter (*pn_ext_ml*) is ≤ 4. The resulting 2XMMi source sample consists of 264,361 detections and 200,067 unique 2XMM sources.

2.2. SDSS Data Release 7

The Seventh Data Release of the Sloan Digital Sky Survey (Abazajian et al. 2009), covers 11663deg^2 , mostly in the northern Galactic cap. A total of 357 million objects have 5 band photometry among which 1.6 million galaxies, quasars and stars were spectroscopically observed. Most of the $\sim$ 2000deg^2 increment over data release 6 are located at low galactic latitude. Astrometric errors are $< 0.1''$ rms. At the 3% error level, the catalogue reaches magnetude limits in the range of 20.5 to 22.2 in the five photometric bands – u, g, r, i and z –. In this paper we only consider the so-called primary sources of the SDSS DR7 Photometric Catalog as available from the VizieR data server. Source lists have been extracted using the VO ConeSearch protocol. The central point of each query is the center of the FOV of the XMM-Newton observation considered and the search radius is the distance from the center to the farthest X-ray source to which we add $3'$ for completeness.

3. Counterpart identification procedure

We discuss in (3.1) how we select optical candidates, taking into account arbitrary error ellipses on the source's spherical coordinates. We compute a likelihood ratio (LR) for each target-candidate pair (3.2). This LR involves a measure of the local density using a kernel smoothing method (3.3). Estimating the true LR distribution for spurious associations (3.4) then allows us to compute for each target-candidate pair the probability of association only based on positional coincidence (3.5).

3.1. Selection of optical candidates

3.1.1. Selection criterion

We consider a target X-ray source and a candidate optical source, with: α_X, δ_X the equatorial coordinates of the X-ray source ; σ_{α_X}, σ_{δ_X} and $\rho_{X_{\alpha,\delta}}$ the error on $\alpha_X \cos\delta_X$ and on δ_X and the correlation between σ_{α_X} and σ_{δ_X} respectively; α_o, δ_o the equatorial coordinates of an optical source σ_{α_o}, σ_{δ_o} and $\rho_{o_{\alpha,\delta}}$ the error on $\alpha_o \cos\delta_o$ and on δ_o and the correlation between σ_{α_o} and σ_{δ_o}, respectively.

As everybody implicitly does – except Budavari & Szalay (2007) – we convert the spherical problem into a plane one and positional errors are interpreted as usual 2D Gaussians. We have chosen a projection on a 2D plane with a frame centered on the position of the X-ray source and having for *x*-

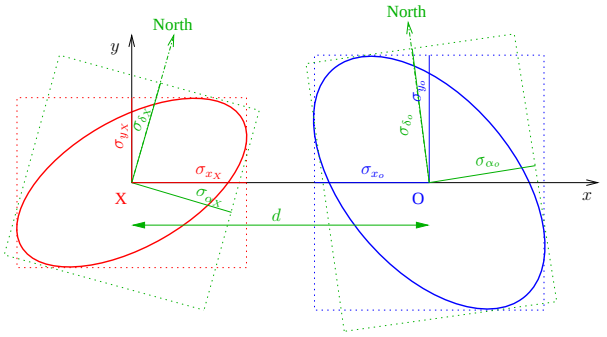


Fig. 1. The chosen projection plane: the xy frame is centered on the X-ray source position X ; the x-axis is the direction toward the optical candidate, located at point O. d is the angular distance between the two sources. This frame is useful at high declinations when we can't consider any longer the meridians – the directions of the north pole in X and in O – to be parallel. It allows to naturally deal with the poles.

axis the direction of the optical candidate (Fig. 1). Errors on positions become Gaussians: $\mathcal{N}_X(x, y; \sigma^2_{x_X}, \sigma^2_{y_X}, \rho_X \sigma_{x_X} \sigma_{y_X})$ and $\mathcal{N}_o(x - d, y; \sigma^2_{x_o}, \sigma^2_{y_o}, \rho_o \sigma_{x_o} \sigma_{y_o})$ with d the angular distance between the X-ray and the optical source. As suggested by Sinnott (1984), d is computed using the Haversine function.

The density of probability that the two sources are at the same location, and thus are the same object, is given by the convolution product of these two distributions. It leads to a new Gaussian:

$$P(x, y) = \mathcal{N}_c(x, y; \sigma_{x_c}, \sigma_{y_c}, \rho_c \sigma_{x_c} \sigma_{y_c}) \tag{1}$$

With $\sigma^2_{x_c} = \sigma^2_{x_X} + \sigma^2_{x_o}$, $\sigma^2_{y_c} = \sigma^2_{y_X} + \sigma^2_{y_o}$ and $\rho_c \sigma_{x_c} \sigma_{y_c} = \rho_X \sigma_{x_X} \sigma_{y_X} + \rho_o \sigma_{x_o} \sigma_{y_o}$.

If the optical source is the counterpart of the X-ray source, it falls with a probability γ inside the ellipse defined by the equation:

$$\begin{pmatrix} x \\ y \end{pmatrix}^t \begin{pmatrix} \sigma^2_{x_c} & \rho_c \sigma_{x_c} \sigma_{y_c} \\ \rho_c \sigma_{x_c} \sigma_{y_c} & \sigma^2_{y_c} \end{pmatrix}^{-1} \begin{pmatrix} x \\ y \end{pmatrix} = k^2_\gamma \tag{2}$$

The completeness we have chosen is a 3σ criterion, often used as a compromise between the total number of associations and the number of counterparts missed (0.3%). This completeness, $\gamma = 99.7\%$, leads in 2D to $k_\gamma = 3.43935$. In the frame we have chosen, the coordinates of the optical source are $x = d$ and $y = 0$. The selection criterion we adopt will retain all candidates satisfying:

$$\frac{d}{\sigma_{x_c} \sqrt{1 - (\rho_c \sigma_{x_c} \sigma_{y_c})^2}} \leq k_\gamma \tag{3}$$

We make the additional following hypotheses. First, we neglect any systematic offset between the positions of the two catalogues. Second, we assume that all positions and associated errors have been computed at the same epoch and therefore corrected for proper motions.

3.1.2. Application to XMM-SDSS DR7 data

The 2XMMi catalogue provides for each source a circular error on position (*radec_err*) and a systematic error (*syserr*). The error on positions of each X-ray source is the quadratic sum of these two values:

$$\sigma_{\alpha_X} = \sigma_{\delta_X} = \sqrt{radec_err^2 + syserr^2} \tag{4}$$

Since it is symmetric we have $\rho_{X_{\alpha,\delta}} = 0$, $\rho_X = 0$ and thus $\rho_c \sigma_{x_c} \sigma_{y_c}$ is directly equal to $\rho_o \sigma_{x_o} \sigma_{y_o}$.

Positional errors are elliptical in the SDSS DR7 catalogue: $\sigma_{\alpha_o} = raErr$, $\sigma_{\delta_o} = decErr$ and $\rho_{o_{\alpha,\delta}} = raDecCorr$. The definition of the different parameters are summarized in Table 1

3.2. Likelihood ratio

We will compute a likelihood ratio (*LR*) for each target-candidate pair meeting the criterion of eq. 3: the probability of finding the optical counterpart at a normalized distance r divided by the probability of having a spurious object at that distance.

We are in the common case where we correlate a sparse catalogue (the 2XMMi) with a dense one (the SDSS-DR7).

The density of probability that the two sources are at the same location knowing x and y corresponds to the density of probability of having the counterpart in x and y assuming that it is the same astrophysical object as the X-ray emitting one. The Gaussian $\mathcal{N}_c$ (eq. 1) can be written, in its canonical form , $\frac{1}{2\pi\sigma_M\sigma_m} \exp -\frac{1}{2}(\frac{x_1^2}{\sigma_M^2} + \frac{y_1^2}{\sigma_m^2})$. Where σ_M and σ_m are the semi-major and semi-minor axis, in the eigenvector frame (x_1, y_1), given by the eigendecomposition of the variance-covariance matrix of $\mathcal{N}_c$,

$$\sigma_{M,m} = \frac{1}{2}(\sigma^2_{x_c} + \sigma^2_{y_c} \pm \sqrt{(\sigma^2_{x_c} - \sigma^2_{y_c})^2 + 4(\rho\sigma_{x_c}\sigma_{y_c})^2}) \tag{5}$$

We change the scale and switch to polar coordinates, leading to the dimensionless angular distance:

$$r = \sqrt{\frac{x_1^2}{\sigma_M^2} + \frac{y_1^2}{\sigma_m^2}} \tag{6}$$

Therefore, the new elementary surface becomes $\pi\sigma_M\sigma_m$, the surface of the 1σ (or $r = 1$) ellipse.

The *LR* we use is inspired by the one described in De Ruiter et al. (1977). As Wolstencroft et al. (1986), we do not only consider the first candidate but all sources satisfying eq. 3. We thus replace the probability "*of finding the first confusing object at a distance lying between r and r* + dr" by the one of finding *a* confusing object between r and $r + \mathrm{d}r$.

The probability of finding the optical counterpart (cp) at a distance lying between r and $r + \mathrm{d}r$ is:

$$\mathrm{d}p(r|cp) = re^{-\frac{1}{2}r^2}\mathrm{d}r \tag{7}$$

And the probability of finding a spurious object (spur) between r an $r + \mathrm{d}r$ is given by the Poisson law:

$$\mathrm{d}p(r|spur) = 2\lambda r \mathrm{d}r \tag{8}$$

We adopt the local surface density of sources at least as bright as m_o, the magnitude of the candidate. In the 1σ ellipse (eq. 6), which is the elementary surface, the Poissonian density of sources is $\lambda = \pi\sigma_M\sigma_m f(\alpha_x, \delta_x, m_o)$.

Since more sources are available in a same given area, the densities computed with this method are more local – or more accurate – than densities computed in arbitrary bins of magnitudes. It is equivalent to computing local densities using more and more sensitive instruments.

The likelihood ratio is the ratio of the two probability densities (7) and (8):

$$LR(r) = \frac{\mathrm{d}p(r|cp)}{\mathrm{d}p(r|spur)} = \frac{1}{2\lambda} e^{-\frac{1}{2}r^2} \tag{9}$$

Table 1. Summary of astrometric parameters for the 2XMM catalogue and for the SDSS DR7.

	σ_α	σ_δ	$\rho_{\alpha,\delta}$
2XMMi	$\sqrt{radec_err^2 + syserr^2}$	$\sqrt{radec_err^2 + syserr^2}$	0
SDSS DR7	*raErr*	*decErr*	*raDecCorr*

The formalism we apply here aims at providing probabilities of identification based on positional coincidences only. We do not use other information on sources such as the spectral energy distribution. Hence, we do not add an extra term $q(m)$ to the LR as is done for example in Wolstencroft et al. (1986), Sutherland & Saunders (1992) and Brusa et al. (2007a). The quantity $q(m)$ corresponds to the probability of having, among the real counterparts, a source of magnitude m, or in a bin Δm around m (see formula (B.8) of the appendix). In this case, $q(m)$ should be local but becomes then hard to estimate. In general the estimate of $q(m)$ is plagued with large errors which, besides the error on the local density estimation, affect dramatically the error on LR.

We will see in section 3.4 that the $q(m)$ factor is somehow taken into account in our reliability function.

3.3. Estimation of the local density

The accuracy of the likelihood ratio (eq: 9) depends sensitively on the local density estimation $f(\alpha_x, \delta_x, m_o)$. A simple manner to estimate the local density of the SDSS-DR7 catalogue is to use a k nearest neighbour (knn) averaging. In this case, the estimated densities have a well known error: $\sqrt{k}/\pi d_k^2$, with d_k the distance of the k^{th} neighbour from the X-ray source. However, knn averaging has the impeding property of being discontinuous. In fact, two candidates with similar magnitudes – or two nearby sources with the same magnitude – could have quite different LRs. The Voronoi tessellation and the use of wavelets also seems too complicated and time consuming for our simple purpose.

Considering these drawbacks, we preferred to use the sample-point density estimator, which is a kernel smoothing, using the Epanechnikov profile as kernel function and the distance to the 100^{th} nearest neighbours as bandwidth.

Kernel smoothing – also called Parzen window technique or kernel density estimation – can be performed with a fixed or a variable bandwidth. Zhang et al. (2006) provide Markov Chain Monte Carlo (MCMC) algorithms for estimating optimal data-driven fixed bandwidths. These algorithms involve the creation of a density map. In our case we consider for each candidate the density of sources at least as bright as the candidate. We thus should compute roughly as many density maps as the number of candidates and apply several times the algorithms. Since the MCMC technique in our case is time consuming, it has been discarded.

We have thus investigated variable bandwidth kernel smoothing techniques. The simplest one is the *balloon* estimator (see Sain 2002). Unfortunately it usually fails to integrate to one. We therefore preferred a *sample-point density estimator* in which the estimate of the density at a point $\boldsymbol{x}$ of a FOV containing N sources at positions $\boldsymbol{x_i}$, $i \in [[1, N]]$ is given by:

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{N} \frac{1}{h(\boldsymbol{x_i})^2} K_E\left(\frac{\|\boldsymbol{x} - \boldsymbol{x_i}\|}{h(\boldsymbol{x_i})}\right) \tag{10}$$

where K is a kernel function and $h(\boldsymbol{x_i})$ a bandwidth depending on the position. According to Comaniciu et al. (2001), this estimator *"is proved to be almost all the time much better than the fixed bandwidth estimator"*. The article gives an usual form of the variable bandwidth $h(\boldsymbol{x_i})$:

$$h(\boldsymbol{x_i}) = h_o \left[\frac{\lambda}{f(\boldsymbol{x_i})}\right]^{1/2} \tag{11}$$

where h_o represents a fixed bandwidth, λ a proportionality constant impacting on the smoothness and f a pilot function. It is also mentioned that *"the method is insensitive to the fine detail of the pilot estimate"*, and *"a good initial choice is to take λ as the geometric mean"* of the pilot function.

The selected pilot function stems from a knn averaging: $f(\boldsymbol{x_i}) = \frac{k}{\pi d_k^2}$ where d_k is the distance to the k^{th} nearest neighbour at least as bright as the candidate. It is fast to compute and the discontinuities can be considered as negligible *fine details*. As suggested in (Comaniciu et al. 2001), we define λ as the mean value of the pilot function: $\lambda = \frac{N}{\pi R_{fov}^2}$, where N is the total number of sources in the field of view having a magnitude smaller or equal to that of the candidate, and where R_{fov} is the radius of the FOV. We have used a fixed bandwidth which corresponds to the area in which we find, on the average, k sources: $h_o = \sqrt{\frac{k}{N}} R_{fov}$. This leads to $h(\boldsymbol{x_i}) = d_k$. In order to complete our estimate of the local density we only need to define parameter k, the number of nearest neighbours being at least as bright as the candidate considered.

The smaller is k, the more local is the density, but the larger is the error on the density estimation. Using $k = 100$ leads to a relative error on LR of about 15%, which corresponds to 6.5% for $\log_{10} LR$. In addition, we implemented an algorithm handling border effects.

The selected kernel profile is the 2D Epanechnikov profile K_E:

$$K_E(\tau) = \begin{cases} \frac{2}{\pi}(1 - \tau^2) & \text{if} \quad \tau < 1 \\ 0 & \text{else} \end{cases} \tag{12}$$

This truncated parabola allows computational optimizations thanks to its finite extent. Moreover, since it minimizes the mean integrated square error (Comaniciu & Meer 2002), this kernel maximizes the quality of the density reconstruction.

We have tested the relative merits of a knn averaging (knn) and a kernel smoothing (ks) on simulated data. A density map of size 400x400 was computed on a generated FOV containing 20,000 sources uniformly spread on the top of which we added 500 sources distributed according to a 2D Gaussian density law of means 3 and 2 in x and y respectively and with a common standard deviation of 1.5. The radius of the FOV is 15. Using $k = 64$ for knn and $k = 100$ for ks lead approximately to the same relative error – respectively 13% and 15% – on the local densities estimates (see Fig. 2). Fitting a 2D Gaussian centered on (3,2) to the resulting density profile inside a 3σ radius ($r = 4.5$) yields a standard deviation of $\sigma_f = 1.476$ and $\sigma_f = 1.492$ for the knn and ks methods respectively, fully consistent with the 1.5 input value.

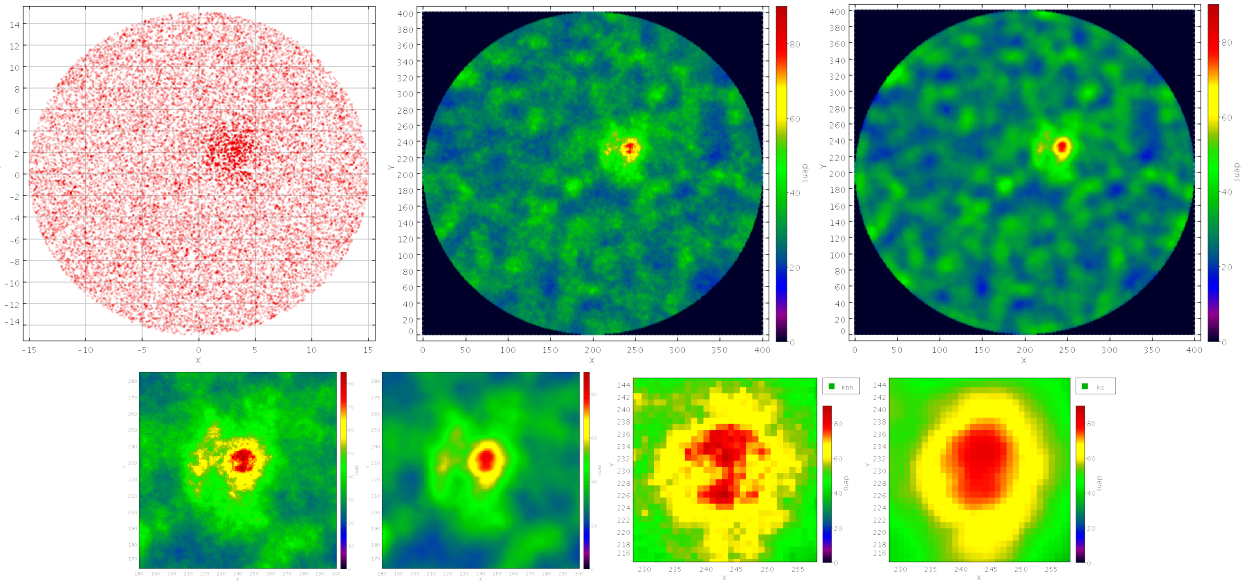


Fig. 2. Application of a knn averaging and of a kernel smoothing on simulated data (see text in 3.3). Top left: the generated distribution of sources (uniform + Gaussian). Top center: knn averaging with k=64. Top right: kernel smoothing with k=100. Bottom: zoom on the Gaussian part of the distribution.

3.4. Computing reliabilities

Although we use a different LR definition, a different estimator of the rate of spurious associations and a different function to fit the reliability histogram, we more or less follow the work presented in part 3 of Oyabu et al. (2005). The method originates in Rutledge et al. (2000).

We define the reliability of an association in a given bin of LR as:

$$R(LR) = \frac{N_{real}(LR)}{N_{real}(LR) + N_{spur}(LR)} = \frac{N_{cand}(LR) - N_{spur}(LR)}{N_{cand}(LR)} \quad (13)$$

where N_{real} and N_{spur} are the unknown number of candidates which are respectively real and spurious counterparts in a given bin of LR ; N_{cand} is the number of candidates in a given bin of LR.

We therefore have to estimate N_{spur}. An often used method consists in correlating X-ray sources with artificial samples of optical sources. The generated samples have the same characteristics as the real sources: same density, same positional errors distribution, etc. Positions are randomly distributed. The sum of the results of such Monte-Carlo samples provides an estimate of the number of spurious associations as function of the distances, of the LRs, etc. This approach is used by Oyabu et al. (2005) among others. In Stephen et al. (2005) the random sample consists in a list of *"anti [...] sources"* which are *"mirrored in Galactic longitude and latitude"*.

We propose here a new method to estimate the number of spurious associations, not based on Monte-Carlo simulations, but instead computing directly their expected results. This scheme offers a better computing efficiency when cross-correlating very large sets of data. The basic idea of estimating the surface of an association related to the total available area can be found in Boller et al. (1998).

Let us first consider the case of one X-ray and one optical source of magnitude m in a FOV. The probability of observing the random X-ray source at a given point (α_X, δ_X) of the FOV is given by the distribution $g_X(\alpha_X, \delta_X)$ of the X-ray sources in the FOV. The chance for the optical source of magnitude m to be randomly associated with the X-ray source is given by the integral of the distribution of optical sources at least as bright as m over the surface of the convolution ellipse $S_{x,o} = k_\gamma^2 \pi \sigma_M \sigma_m$. The probability that the association has a LR in a bin ΔLR is then given by the integral of the distribution of optical sources at least as bright as m over the surface the bin occupies inside the convolution ellipse $S_{x,o}$: $S_{x,o}(\Delta LR)$. Finally, the sum of these probabilities for all X-ray and archival sources provides an estimate of the number of spurious associations by LR bin. Considering now the case of N X-ray and M optical sources in the FOV leads to the formula of $N_{spur}(\Delta LR)$:

$$\sum_{x=1}^{N} \iint_{S_{FOV}} g_X(\alpha_X, \delta_X) \left(\sum_{o=1}^{M} \iint_{S_{x,o}(\Delta LR)} g_o(\alpha_o, \delta_o, m) \mathrm{d}\alpha_o \mathrm{d}\delta_o \right) \mathrm{d}\alpha_X \mathrm{d}\delta_x \quad (14)$$

where $S_{x,o}(\Delta LR)$ depends on the X-ray source, the optical source, and the position of the X-ray source and can thus be written as $S_{x,o}(\Delta LR, X, o, \alpha_X, \delta_X)$.

A possible manner to make the problem simpler is to assume that $g_o(\alpha_o, \delta_o, m)$ is roughly constant inside the surface $S_{x,o}(\Delta LR)$. Formula (14) can then be written as:

$$\iint_{S_{FOV}} g_X(\alpha_X, \delta_X) \left(\sum_{x=1}^{N} \sum_{o=1}^{M} S_{x,o}(\Delta LR) g_o(\alpha_X, \delta_X, m) \right) \mathrm{d}\alpha_X \mathrm{d}\delta_x \quad (15)$$

In order to estimate $N_{spur}(\Delta LR)$ we have to compute a complete distribution map of the X-ray sources and as many distribution maps as the number of distinct optical source magnitudes. Since such a computation is potentially time consuming, we may suppose that optical sources are uniformly distributed in the FOV

and have as local density the mean value of their local densities. This less rigorous approach leads to the simple formula:

$$N_{spur}(\Delta LR) = \sum_{x=1}^{N}\sum_{o=1}^{M} S_{x,o}(\Delta LR)/S_{FOV} \quad (16)$$

The quantity $S_{x,o}(\Delta LR)$ can be computed as follows: the radius corresponding to a given LR is easily derived from equation (9):

$$r^2(LR) = \begin{cases} 0 & \text{if } LR \geq LR_{max} = \frac{1}{2\lambda} \\ \ln\frac{1}{(2\lambda LR)^2} & \text{if } LR_{min} < LR < LR_{max} \\ k_\gamma^2 & \text{if } LR \leq LR_{min} = \frac{1}{2\lambda}e^{\frac{1}{2}k_\gamma^2} \end{cases} \quad (17)$$

Thus for a bin $\Delta LR = LR_2 - LR_1$, $LR_2 > LR_1$, if we have $LR_1 \leq LR_{max}$ and $LR_2 \geq LR_{min}$, the surface of the bin of LR in the ellipse is expressed by:

$$S_{x,o}(\Delta LR) = 2\pi\sigma_M\sigma_m \ln\frac{min(max(LR_2, LR_{min}), LR_{max})}{max(min(LR_{max}, LR_1), LR_{min})} \quad (18)$$

In principle, we should remove the optical sources which are the real counterparts since they are not randomly distributed. However, their fraction can be neglected at low LR and at very high LR their influence on the final probability of identification is insignificant. In the case of real data, the source distribution is not completely Poissonian since two sources cannot be infinitely close due to the PSF, the resolution of the instrument and the extraction algorithm. There is therefore a small surface around each counterpart which cannot contain another source. Neglecting this can lead to slightly overestimate the rate of spurious associations.

3.4.1. Practical computation

In order to avoid computing too many local densities for estimating the rate of spurious associations, we divide the magnitude range into bins and we associate all sources in a same magnitude bin with the mean value of their local density. The width of the bins depends on the magnitude accuracy of the catalogue. We then compute for all optical and X-ray sources the $\sigma_M\sigma_m$ factor, the LR_{min} and LR_{max}. It is thus possible to compute the histogram of the expected number of spurious associations according to LR values. To increase the computing efficiency, we can bin the $\sigma_M\sigma_m$ values. However, this approach involves another loss of accuracy for a meager reduction of computing time.

As shown in Fig. 3, the histograms used in the computation of the reliability are the number of candidates and the number of spurious associations grouped in bin of $\log_{10} LR$.

3.4.2. Fitting the reliability function

In order to estimate the number of spurious associations we take the relatively realistic example in which the X-ray source has at most one candidate. The reliability of an association (not to be mixed with the integrated reliability for all association having a $LR \geq l$) is directly given by (see formula (B.6) of the appendix):

$$R(LR) = p(id|r) = \frac{1}{1 + \frac{p(spur)}{p(cp)}\frac{1}{LR}} \quad (19)$$

The term $p(cp)/p(spur)$ – the probability that the optical source is a counterpart divided by the probability that it is spurious – must be independent of the dimensionless distance r (see formula 6). It is similar to the term $(1-\theta)/\theta$ used in De Ruiter et al. (1977). However, $p(cp)/p(spur)$ may depend on the nature of the underlying X-ray source population (e.g. stars, AGN) and may thus vary with source properties such as magnitude, optical colour or flux ratios. In order to obtain a LR similar to that used in Brusa et al. (2007a) for instance, we would need to consider an additional parameter $q(m)$ describing the variation of $p(cp)/p(spur)$ with the magnitude (or any other relevant property) of the candidate counterpart. Alternatively, $q(m)$ may be replaced by an other term such as that playing the role of B_{phot} in Budavari & Szalay (2007). A $R(LR)$ histogram can be built from the $N_{cand}(LR)$ and $N_{spur}(LR)$ histograms made using the method explained in the previous paragraph. If the ratio $p(cp)/p(spur)$ were independent of source properties, $R(LR)$ could be fitted with function (19) using only one free parameter $a = p(cp)/p(spur)$. Including a term $q(m)$ in LR with $N_{\Delta m}$ bins of magnitude, requires to build $N_{\Delta m}$ $R(LR, \Delta m)$ histograms and fit each of them with functions (19) having different a parameters. However, in general the lack of statistics does not allow to do so.

The $R(LR)$ histogram can then be seen as the sum of $N_{\Delta m}$ $R(LR, \Delta m)$ histograms and consequently can be modelled by the following function:

$$R_f(LR) = \sum_{i=1}^{N_{\Delta m}} \frac{b_i}{1 + a_i\frac{1}{LR}}, \text{ with } b_i = \frac{N_{cand_{\Delta m_i}}}{\sum_{j=1}^{N_{\Delta m}} N_{cand_{\Delta m_j}}} \quad (20)$$

where $N_{cand_{\Delta m_i}}$ is the total number of entries in histogram number i.

In practice, the histograms are not binned according to LR but to $\log_{10}(LR)$. Best fits were obtained using the function:

$$R_f(x = \log_{10}(LR)) = \frac{\sum_{i=1}^{N_{\Delta m}} b'_i 10^{(1-i)x}}{1 + \sum_{i=1}^{N_{\Delta m}} a'_i 10^{-ix}} \quad (21)$$

with $N_{\Delta m} = 3$, i.e. 6 free parameters.

The fit is performed using a Levenberg-Marquard algorithm. We compute as many LR and we construct and fit as many LR histograms as the number of magnitude bands in the SDSS.

3.5. Computing probabilities of identification in the general case

We now extend the Bayesian approach to X-ray sources having N_{cand} candidates. We assume that at most one association is real. Although this assumption may be wrong, especially in the case of SDSS extended objects or close optical sources not resolved in X-ray, it makes the problem more tractable and can easily be linked to the case of a unique candidate. Let us consider $N_{cand}+1$ hypotheses:

- H_{cp_i}: the ith optical source is the counterpart
- $H_{spur_{all}}$: there is no counterpart

Then the Bayesian probability that the ith source is the counterpart knowing r_j, $j \in [1, N_{cand}]$ is:

$$P_{id,i} = R'_i = p(H_{cp_i}|r_1 \cap \ldots \cap r_{N_{cand}}) = \frac{LR_i\frac{p(H_{cp_i})}{p(H_{spur_{all}})}}{1 + \sum_{j=1}^{N_{cand}} LR_j\frac{p(H_{cp_j})}{p(H_{spur_{all}})}} \quad (22)$$

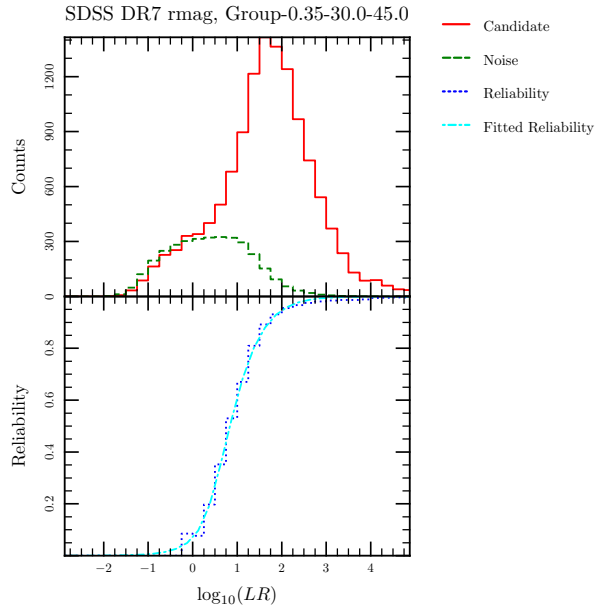


Fig. 3. Top: Histograms of the number of associations and of the estimate of the number of spurious associations by bin of $\log_{10}(LR)$. Bottom: Reliability histogram by bin of $\log_{10}(LR)$ and its fitted curve. In the example, LRs have been computed according to the SDSS DR7 *r* magnitude for XMM sources having a systematic error of 0.35 *arcsec* with a galactic latitude $30° < |b| < 45°$.

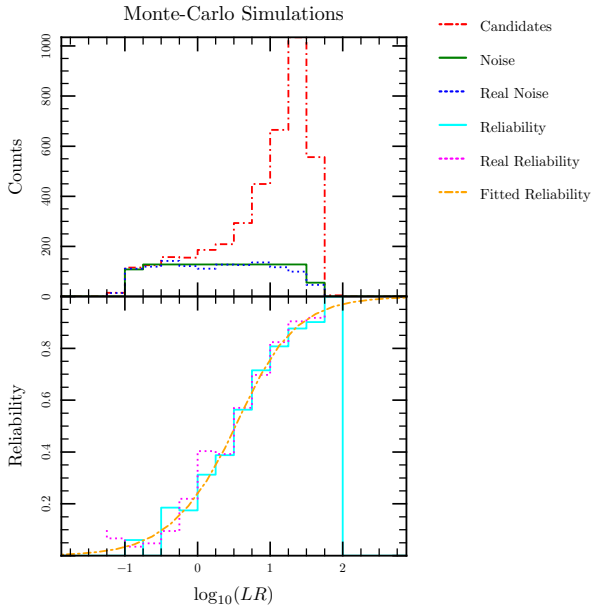


Fig. 4. Comparison of real and estimated reliabilities on simulated data (see text in section 4.2). Top panel: the dot-dashed line represents the total number of associations ; the solid and the dotted lines are respectively the number of spurious associations estimated by our algorithm and the real number of spurious associations. Bottom panel: the solid and the dotted lines represent the estimated reliability and the the real reliability respectively ; the dot-dashed line is the fit of the estimated reliability histogram.

If $\frac{p(H_{cp_i})}{p(H_{spur_{all}})} = \frac{p(cp_i)}{p(spur_i)}$, equation (22) leads to the formula below, obtained following Rutledge et al. (2000) prescription:

$$P_{id,i} = R'_i = \frac{\frac{R_i}{1-R_i}}{1 + \sum_{i=1}^{M} \frac{R_i}{1-R_i}} \quad (23)$$

Using equation (19), we easily show that $R/(1-R) = LRp(cp)/p(spur)$. Computing $p(H_{cp_i})$ and $p(H_{spur_{all}})$ normalizing the terms $p(cp_i)$, $p(spur_i)$ as Rutledge et al. (2000) do to construct $P_{id,i}$ from R_i, we obtain the equality $\frac{p(H_{cp_i})}{p(H_{spur_{all}})} = \frac{p(cp_i)}{p(spur_i)}$. We thus apply formula (23) to compute final probabilities of identification.

Each candidate possesses as many reliabilities as the number of magnitude bands in the SDSS. The R_i we consider in the final probability of identification formula are for each source the best one among all photometric bands.

It should be stressed that the identification probabilities given here are only based on positional coincidences. We plan to compute more complex likelihood ratios taking into account various sources parameters and using the object type found by the source classification algorithm.

4. Test and implementation

4.1. Observation grouping

XMM-Newton EPIC sources are correlated FOV by FOV. In order to tail off count-rate noise on FOV LR histogram bins without sacrificing resolution, we have to increase count statistics. We therefore stacked data coming from similar FOV:

- We split in two groups XMM-Newton FOV having different systematic errors on position: 0.35" or 1.0"
- Observations of the LMC and SMC regions are set apart
- Since they presumably share objects of same nature and same pattern of logN-logS law, observations are grouped according to their galactic latitude

4.2. Tests on simulated data

We applied our algorithm to simple samples of simulated data. We generated 90 FOV, each containing 100 X-ray sources located inside a 15′ radius FOV and 2000 optical sources inside a 18′ radius FOV. The 3 additional arcminutes are required to avoid boundary effects on local density estimates (see above). A random offset error of 2.5 and 1 *arcsec* was applied to X-ray and optical sources positions respectively using the Marsaglia polar method (MPM). Among the 100 and the 2000 sources, 30 were common: i.e. had exactly the same location before we applied the MPM. All optical sources had the same magnitude hence, we do not have generate logN-logS curves. A comparison of the rate of spurious associations and reliability derived from our method with those entered in the simulation is exemplified in Fig. 4. The estimated reliabilities are in good agreement with those of the simulation.

4.3. Implementation

All software was written in Java. 2XMMi sources were retrieved using direct SQL queries – through a JDBC connection – on the XCAT-DB. SDSS DR7 sources were collected from the VO thanks to the ConeSearch protocol, the correlation has thus been performed independently for each XMM observation. The re-

sults of the correlation process have been added, via Saada, to the XCAT-DB and are available as plain FITS files.

Local density estimations required a lot of k nearest neighbours search. Moreover, for each candidate we are only interested in sources having a magnitude smaller or equal to that of the candidate. In order to quickly perform such knn queries, SDSS DR7 sources have been stored in quadtrees especially designed for that purpose: each leaf of the tree stores the greatest magnitude among those of the sources held in its sub-tree ; knn queries contain an extra parameter which is the magnitude of the candidate ; in addition to the distance criterion, we thus easily add a magnitude criterion to stop the search in a sub-tree.

A quadtree scheme was preferred to a kdtree since it is fastest to build and easier to update. It is then possible to build the quadtree during the parsing of the VOTable returned by the ConeSearch query.

5. Results of the 2XMMi-SDSS DR7 cross-correlation

A total of 1337 XMM-Newton FOV hold at least one source having a SDSS counterpart candidate. These 1337 FOV contain 73 636 unique 2XMMi sources and a total of 95 452 detections. The cross-correlation of the 2XMMi catalogue with the SDSS DR7 leads to 72 169 "associations" involving 45 727 and 55 726 unique 2XMMi and SDSS DR7 sources respectively. This first number represents 20% and 62% of the unique sources available in the entire 2XMMi catalogue and in the 1337 FOV respectively. The distribution of the number of SDSS DR7 candidates by unique 2XMMi sources is given in Table 2 and the main properties of the distribution of the probabilities of identification and of their cumulative values are given in Table 3. A total of 7 740 unique 2XMMi sources have several SDSS DR7 candidates. In this sub-sample, there are 896 and 2 672 2XMMi sources for which the candidate with the largest identification probability is not the nearest and the brightest SDSS DR7 candidate respectively. The left panel of Fig. 5 shows the distribution of the individual SDSS source identification probabilities. Most SDSS entries found within the combined 3σ search radius from the 2XMMI source have a high likelihood to be the true optical counterpart. The small tail of very low identification probability objects reflects the expected rising contribution of SDSS entries unrelated to the X-ray source at large matching distances. Most SDSS entries with identification probability larger than $\sim$ 90% are found less than 3 arcsec from the X-ray position (Fig. 5, center panel). The rather large spread of the 2XMMi positional errors accounts for the scatter affecting the distances at which high probability SDSS sources are found from the X-ray position. Expressed in terms of combined 2XMMi + SDSS errors, the distance distribution shown in the right panel of Fig. 5 follows the usual shape of a Rayleigh distribution. Fitting this histogram with a Rayleigh function plus a linear component, we obtain $\sigma_r = 0.865$ for the Rayleigh curve parameter, 0.178 for the linear slope and $R = 0.675$ for the ratio between the total number of real associations and the total number of spurious ones within the search radius. Fitting separately distance histograms of the sources whose positions were corrected by *eposcorr* and uncorrected ones leads to $\sigma_r = 0.856$, $R = 0.698$ and $\sigma_r = 1.013$, $R = 0.50$ respectively. All errors on σ_r, on the slope and on the R values are about 0.003. Keeping only the best candidate for each unique XMM sources, we obtain R=0.81 which is consistent with the value of 79.4% given in Table 3. The origin of this small apparent overestimate ($\sim$ 14%) of the positional errors of *eposcorr* corrected sources is so far unclear. In any cases, the effect of this slight change on the identification probabilities is not large.

About 40% of all unique 2XMMi sources located in the 1 337 FOV are identified with a probability larger than 90% with a SDSS DR7 entry. Among all unique 2XMMi sources having at least one candidate, 66% have a probability of identification greater than 90%. This value rises to 72% and 90% if we consider optical sources at a distance smaller than 5″and having at least one and all magnitudes less than 22 mag respectively.

Table 2. Number of unique 2XMMi sources (Nx) having Nc SDSS DR7 candidates.

Nc	1	2	3	4	5	6	7	8	>8
Nx	37 988	6059	1196	317	96	47	12	4	8

Table 3. Number of associations with all XMM detection (*det_id*) and all unique sources (*src_id*) having a probability of identification greater than given reliability cutoffs. If an X-ray source has several candidates, we only keep the one with the best probability of identification. Reliability (R) and completeness (C): we only consider the best match for each unique XMM sources having at least one counterpart.

	Probability of identification cutoffs					
id	0.0	0.5	0.7	0.8	0.9	0.95
# det_id	60 567	53 347	49 527	46 387	40 193	32 610
# src_id	45 727	39 839	36 943	34 605	30 055	24 327
R	79.4%	91.7%	94.7%	96.2%	98.0%	99.0%
C	100.0%	96.8%	92.1%	87.5%	77.2%	63.2%
Frac X	62.1%	54.1%	50.2%	47.0%	40.8%	33.0%

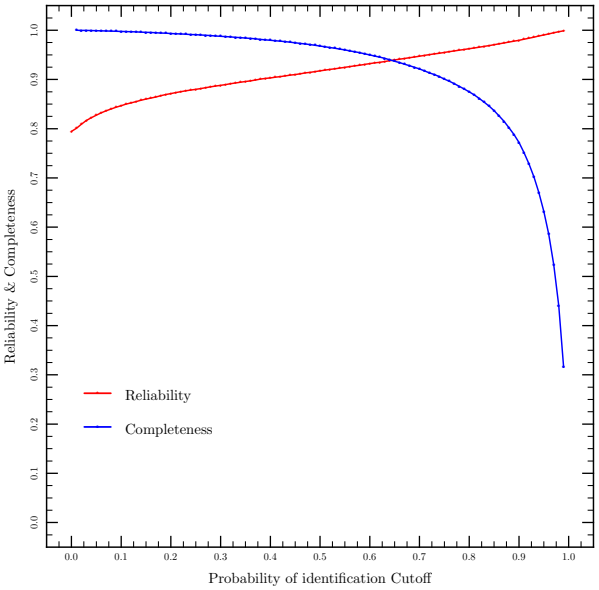


Fig. 6. Sample reliability and completeness percentages as functions of the probability of identification cutoff.

6. Building a learning sample

One of the important goal of the cross-correlation process is to analyse the parameter locii occupied by the different classes of X-ray emitters having a reliable optical counterpart in the SDSS. This analysis constitutes the first mandatory step toward build-

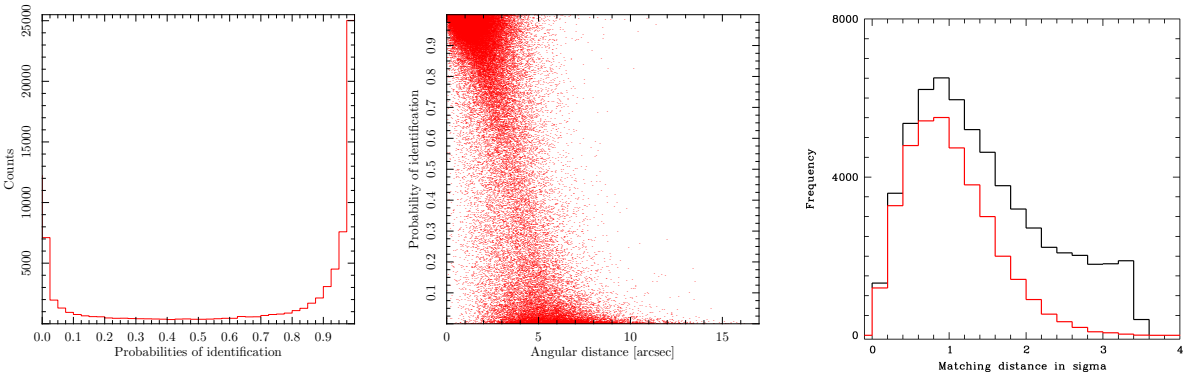


Fig. 5. Left: Histogram of the individual probabilities of identification. Center: probabilities of identification versus matching distances. Right: distribution of the distance of SDSS candidates to 2XMMi sources expressed in units of the combined 2XMMi + SDSS positional error; black = all matches, red = identification probabilities ≥ 90%

ing a source classification scheme which may eventually use advanced statistical methods such as principal component analysis, knn, decision trees, etc. We thus tried to gather a sample of 2XMMi sources having reliable SDSS DR7 counterparts and as much as possible a source nature derived from a true spectroscopic identification. In order to do so we mostly relied on entries in the SDSS spectroscopic catalog accessible via CasJob and following SDSS spectral class scheme, define the groups: Stars, Galaxies, QSOs to which we add classes AGN (see below) and X-ray binaries. Further identifications of various members of these classes were extracted from the cross-correlation with Simbad and other archival catalogues performed during the pipeline processing of XMM-Newton observations and available from the XCat-DB. We also added 5 candidate type II QSOs taken from Zakamska et al. (2003) which have a reliable match in the 2XMMi catalogue.

All objects found in a selection of archival catalogues and lying less than 3″ from an X-ray source associated with a SDSS DR7 entry have entered the learning sample.

The list of catalogues used is the following:

- Stars : 2XMMi/SDSS associations with a probability of identification > 0.80 and having the specClass attribute set to 1 or 6 ; Simbad entries from which we only have retrieved objects flagged as 'TTau*', 'pMS*', 'Pulsar', 'WR*', 'Em*', 'Be*' or 'WD*' and to which we add the sample of stars coming from the kernel density classification (see below).
- X-ray binaries : the Downes catalogue of cataclysmic variables (Downes et al. 2001) ; the Ritter catalogue of catactlysmic variables (Ritter & Kolb 2003) and the Ritter catalogue of LMXRBs (Ritter & Kolb 2003).
- Galaxies : the Uppsala catalogue (Lauberts 1982). We selected here all 2XMMi sources located within the optical extent of the galaxy. However, in order to avoid too many spurious matches we only considered galaxies having a major axis of less than 5 arcmin; 2XMMi/SDSS associations with a probability of identification > 0.80 and having the specClass attribute set to 2.
- AGNs : sources from the Véron catalogue (Véron-Cetty & Véron 2006) having the spectral type sp='S' or sp='Sx' or sp='Sxx' or empty but having the class cl='A'. Consequently, the sample also contains the 'S3' types which correspond to LINERs.
- QSOs : entries from the Véron catalogue having the spectral type sp empty or set to 'HP' with the class cl='Q' ; objects from SDSS DR7 associated to a 2XMMi source with a probability of identification > 0.80 and having the specClass attribute set to 3 or 4

Table 4. Distribution of objects of each types coming from the XMS and the XBS in the learning sample for "All" 2XMMi/SDSS associations and for the "Final" sample only containing associations with a probability of identification > 0.9, with SDSS sources unsaturated, not blended, ..., and having all their magnitudes < 22.2 mag.

Sample	Star		Galaxy		AGN		QSO	
	XMS	XBS	XMS	XBS	XMS	XBS	XMS	XBS
All	3	19	1	0	7	21	49	95
Final	1	3	1	0	5	14	37	85

An X-ray source associated with both a star and a binary was flagged as binary. We applied the same rules for star–agn, star–qso, binary–agn, binary–qso and qso–agn pairs of apparently conflicting nature. We have also added sources identified in the XMM-Newton Medium Sensitivity Survey (XMS, Barcons et al. 2002; Carrera et al. 2007; Barcons et al. 2007) and in the Bright Sources Survey (XBS or BSS Della Ceca et al. 2004) two surveys conducted by the SSC. For the XMS, we considered NELG as AGN and BLAGN and BLLac as QSO. For the XBS, AGN2 and elusive AGN were assigned the general AGN type and AGN1 and BLLac the QSO type.

6.1. The stellar learning sample

Building a clean stellar sample turned out to be more difficult since most stellar sources detected in X-rays have optical SDSS magnitudes smaller than 15 mag and are flagged as saturated. In particular, this prevents from using the Tycho 2 catalogue as reference stellar identifications. Furthermore, the SDSS R7 spectroscopic database provided only 26 cross matches with acceptable properties (i.e. non staturated and probabilities of identification larger than 90%). Therefore, in order to enlarge the stellar sample, we applied a classification method allowing to identify stars on the basis of their multi-colour properties.

We performed a kernel density classification (KDC Richards et al. 2002) on all unresolved (type=6) SDSS candidates having none of the flags BLENDED, DEBLENDED_AS_MOVING, EDGE or PSF_FLUX_INTERP set and with magnitudes in all bands brighter than 22.2 mag. This selection returns 10 533 SDSS sources having a correlation in the 2XMMi catalogue.

Table 5. Origins of the objects of each types in the learning sample for "All" 2XMMi/SDSS associations and for the "Final" sample only containing association having a probability of identification > 0.9 with SDSS sources unsaturated, not blended, ..., and having all their magnitudes < 22.2 mag. Ver. stand to Véron, R1 for Ritter catalogue of CV, R2 for Ritter catalogue of LMXB, Sim. for Simbad, Tyc. for the Tycho 2 catalogue and its first supplement, Upp. for Uppssala and Dow. for the Downes catalogue of CVs.

Sample	Star			Binary			Galaxy		AGN	QSO	
	DR7	Sim.	KDC	Dow.	R1	R2	Upp.	DR7	Ver.	Ver.	DR7
All	43	39	531	49	40	2	3	965	704	1 231	1 455
Final	26	8	517	29	22	2	2	757	497	1 027	1 264

Table 6. Results of the classification method applied to the learning sample

org\assign	Star	QSO
Star	96.89%	3.11%
QSO	1.62%	98.38%

The classification only uses the four colours $u-g$, $g-r$, $r-i$ and $i-z$ as parameters. The learning sample used for this classification consists of 2 classes: star and qso since we only consider point-like objects in the optical. It has been built from all unresolved SDSS sources, independently of their association with a 2XMMi entry. We only retained good quality detections (i.e. no flag SATURATED, BLENDED, DEBLENDED_AS_MOVING, INTERP_CENTER, EDGE, SATUR_CENTER or PSF_FLUX_INTERP set) spectroscopically identified in the DR7. The data have been retrieved from the DR7 database with CasJob. The stellar sample contains 67 269 sources flagged by the SDSS specClass attribute as star (STAR or STAR_LATE) and therefore it the also contains CVs and WDs. The non–stellar sample has 75 248 sources flagged by the SDSS specClass attribute as QSO (QSO) or high redshift QSO (HIZ_QSO) plus 253 sources flagged by specClass as galaxy (GALAXY). For simplicity we hereinafter name the non–star sample as QSO sample.

Estimates of the probability densities have been computed using a fixed bandwidth kernel smoothing. The kernel applied uses the Epanechnikov profile and the bandwith has been chosen equal to 0.2 mag. Table 6 list the results of the self-check of the learning sample, i.e., the results of the classification method applied to the learning sample only.

The prior probability p(star) has been set to 0.25 and so p(qso) to 0.75 as a result of iterative kernel density classifications converging to this relative number of stars and qsos in the SDSS/2XMMi learning sample. In order to obtain the cleanest stellar SDSS/2XMMi sample we only retained objects having a probability of being a star larger than 99% as given by the kernel density classification. So as to select SDSS/2XMMi identification with the best chance to be normal stars, we removed 13% of all SDSS entries classified as stars, but falling in low density regions (i.e. far from the center of the stellar multi-colour locus) and thus prone to be doubtfull cases such as binaries, unidentified cataclysmic variables or even mis-identified AGN. The corresponding fraction of 2XMMi/SDSS stellar identification removed is $\sim 32\%$ indicating that these stellar photometric outliers are more likely to be X-ray active.

The star/qso classification has been made according to the optical properties of the spectroscopically identified SDSS objects. However, by construction, the density distribution in colours of the SDSS/2XMMi sample is not likely to follow that of the non X-ray emitting SDSS objects and can thus lead to some biases. For instance, there may be a large over density of non-X-ray emitting stars in some part of the 4-d colour diagram in which most objects classified as qso appear to be strong X-ray sources. This problem actually occurs in the region covered by the branch of high redshift QSOs where there is some overlap with A stars. Although some A stars do emit X-rays for debated reasons, not all do. We thus manually removed from the SDSS/2XMMi learning sample all classified stars having a $u-g$ colour less than 1.2 (values taken from Covey et al. 2007). The stellar sample arising from the KDC contains 842 entries with a high classification probability larger than 99%, with $u-g>1.2$ and excluding possible outliers.

We also checked that the stellar SDSS/2XMMi sample was following in the 4-d colour space the stellar locus derived by (Covey et al. 2007) using synthetic photometry. The agreement in good, apart for the reddest stars of spectral type later than $\sim$ M5.

Table 7. Distribution of number of unique 2XMMi entries with types in the learning sample. "All" refers to entire set of 2XMMi/SDSS associations considered and "Final" refers to the final sample considered which only contains associations with a probability of identification > 0.9 and SDSS sources unsaturated, not blended, ..., and having all their magnitudes < 22.2 mag.

Sample	Star	Binary	Galaxy	AGN	QSO
All	842(tbc)	47	860	797	1704
Sub	536	26	673	572	1425

6.2. The final learning sample

In the following, we will only consider optical identifications with a probability larger than 90%, for both the spectroscopically identified learning sample and for the general photometric sample corresponding to a sample purity of 98%. We also removed all SDSS entries having one of the following flag set: BLENDED, DEBLENDED_AS_MOVING, SATURATED , INTERP_CENTER, EDGE, SATUR_CENTER, PSF_FLUX_INTERP. The distribution of the various classes is listed in Table 7.

Finally, for all AGNs having a spectroscopic redshift we compute the X-ray luminosity using:

$$L_{abs} = 4\pi F \frac{(1+z)^2}{(1+z)^{2-\gamma}} \left(\frac{c}{H_0} \int_0^Z \frac{1}{\sqrt{\Omega_m (1+z)^3 + \Omega_\Lambda}} dz \right)^2 \qquad (24)$$

Where F is the 0.2-12 keV X-ray flux in erg.s^{-1}cm^{-2}, $H_0 = 73$, $\Omega_m = 0.3$ and $\Omega_\Lambda = 0.7$. The photon index γ has been taken equal to 1.9.

7. Discussion

The cross-correlation of two large catalogues such as the 2XMMi and the DR7 of the SDSS pathes the way to a wide range of possible investigations. In addition, the availability through

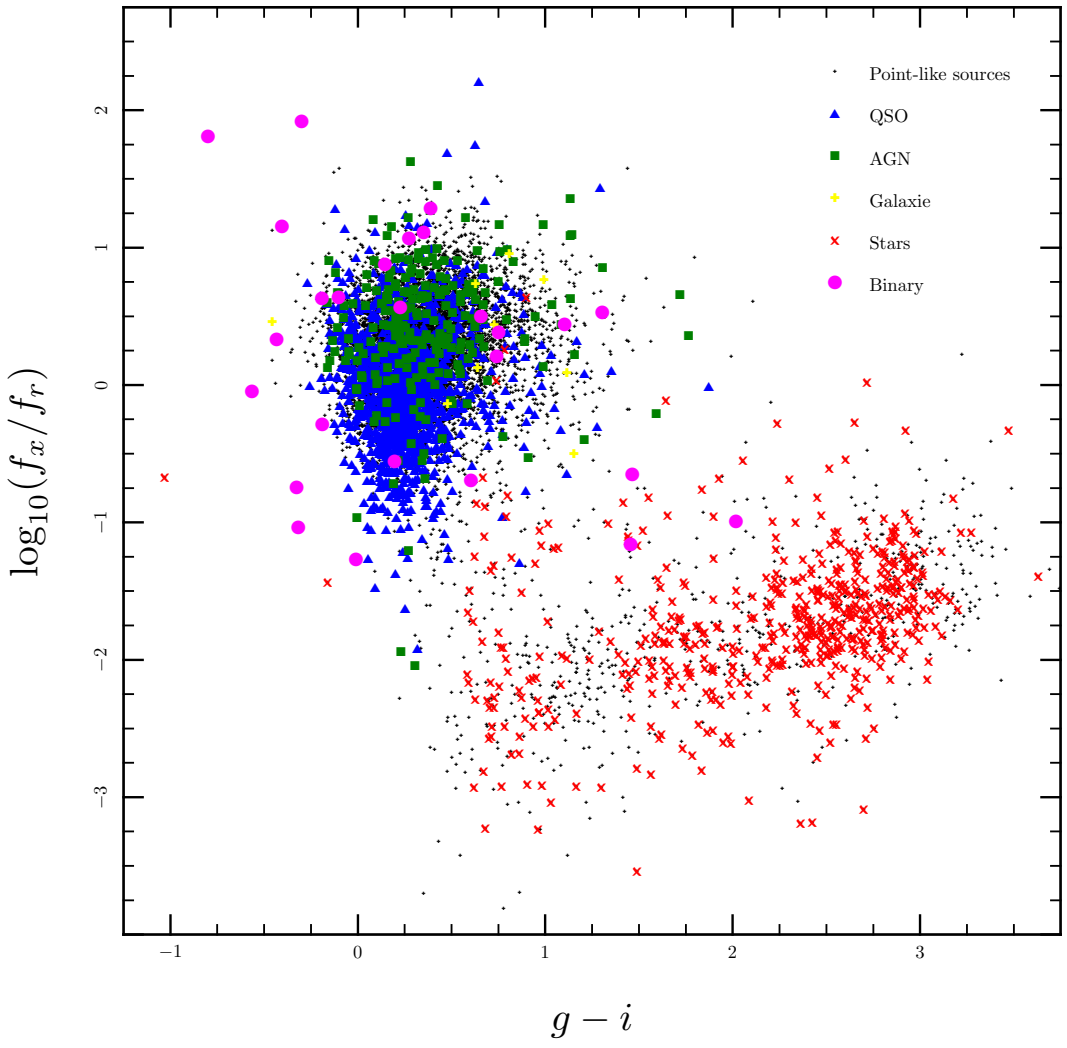


Fig. 7. The distribution of the identified sample (coloured large symbols) in the $log(f_x/f_r)$ versus $g-i$ diagram superimposed on the entire cross-match sample (points) for unresolved optical objects. We only show SDSS entries with g, r, or i magnitudes brighter than 22.2 in order to keep photometric errors below ~ 3%

the SDSS database of line fluxes for over a million of spectroscopic targets offers unprecedented opportunities to study how the micro-physics of the optically emitting regions depends on X-ray properties. For active stars, it might be possible to study the physical links and heat transport mechanisms between the optical line emitting chromosphere and the X-ray active corona. An even wider range of line based studies are probably possible in active galaxies. In spite of their sound scientific interest, such studies are beyond the scope of the present work.

Keeping in mind the long term objective endorsed by the SSC to perform an automatic classifications of all X-ray sources discovered serendipitously in the field of view of the EPIC cameras of XMM-Newton, we will focus in the next section on the analysis of the main locii occupied by the different classes of X-ray sources and will give some general rules to separate as much as possible the different groups of X-ray emitters. In the rest of the discussion we will investigate in some details how members of each class are distributed in their domain according to their intrinsic properties.

7.1. Identifications of the main classes of X-ray sources

We show in Figs 7 and 8 the positions of the various learning samples and of all cross-identifications having a probability larger the 90% in the $g-i$ / $log(f_x/f_r)$ diagram, separately for spatially unresolved and extended objects.

7.1.1. Separating stellar from extragalactic sources

As expected, the f_x/f_r ratio is very powerful to separate the late type stellar X-ray population in which the high energy emission arises in a magnetic active corona from X-ray luminous sources

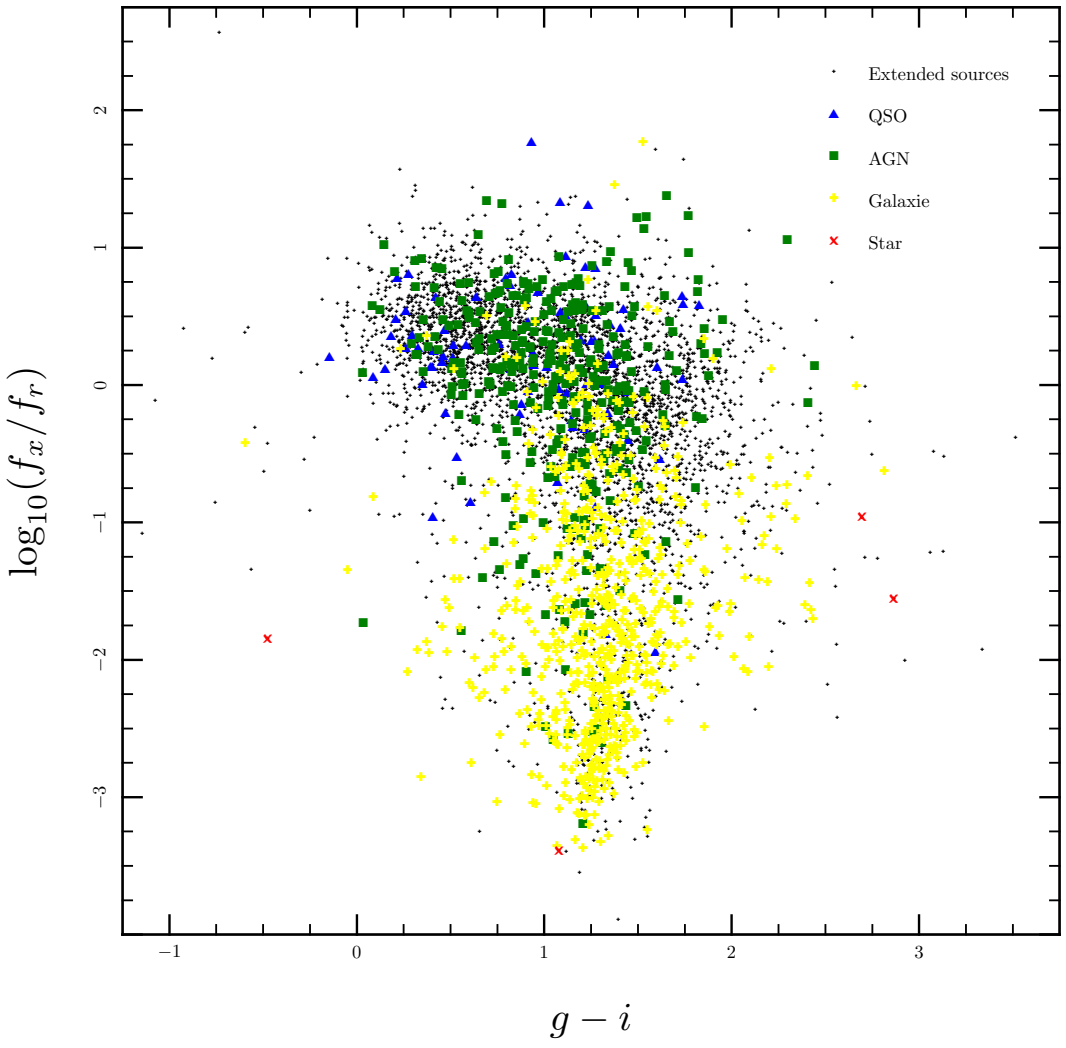


Fig. 8. The distribution of the identified sample (coloured large symbols) in the $log(f_x/f_r)$ versus $g-i$ diagram superimposed on the entire cross-match sample (points) for extended optical objects only. We only show SDSS entries with g, r, or i magnitudes brighter than 22.2 in order to keep photometric errors below ~ 3%

powered by accretion such as active galactic nuclei or cataclysmic variables. However, the distribution of low L_X galaxies in $log(f_x/f_r)$ clearly overlaps with that of active coronae. Introducing the $g-i$ colour index allows to separate the bulk of the stars, especially the reddest M stars from most galaxies. Unfortunately, many galaxies, in particular of the early type exhibit optical energy distributions similar to those of G type stars, show comparable (f_x/f_r) and consequently cannot be distinguished from stars in the diagram. However, taking into account the spatial extension of the optical source allows to efficiently separate them from stars (see Figs. 7 and 8).

Although cataclysmic variables occupy a locus in the $g-i$ / $log(f_x/f_r)$ comparable to that of most quasars, their distribution exhibits a larger spread than that of AGNs. This large scatter can be used to provide a high likelihood identification of their class, at least for part of them. For instance, very blue objects, typically with $g-i$ below -0.2 as well as those with extreme f_x/f_r have a high probability to be cataclysmic variables. A number of new CV candidates can be extracted from the present work.

7.1.2. Distinguishing between the various classes of extragalactic sources

It can be readily seen from Fig. 7 that many of the 2XMMi sources having a counterpart in the DR7 of the SDSS cluster in a rather narrow range of blueish $g-i$ colours in the interval of -0.2 to 0.8. They have $log(f_x/f_r) \sim 0$ and appear as point-like sources in the optical. Their positions in this diagram overlaps with that of the vast majority of the spectroscopic SDSS QSOs found in our learning sample which for most of them are UV excess optically selected quasars. The mean $log(f_x/f_r)$ of the learning sample appears slightly shifted to smaller values compared to that of

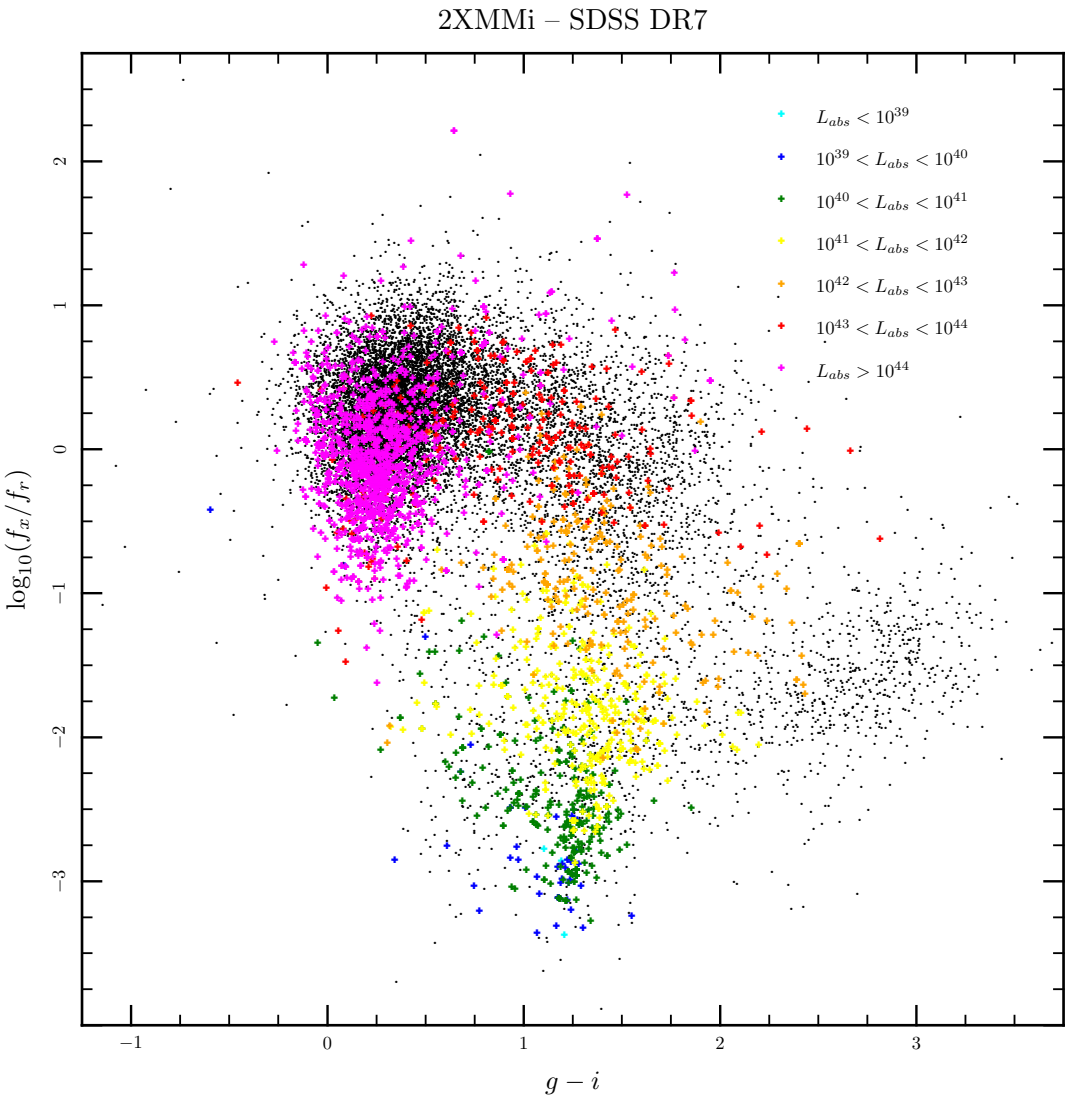


Fig. 9. The distribution of the quasar + galaxy identified sample in the $log(f_x/f_r)$ versus $g - i$ diagram (large symbols with colours changing with L_X) superimposed on the entire cross-match sample (points) for both resolved and unresolved optical objects. We only show SDSS entries with g, r, or i magnitudes brighter than 22.2 in order to keep photometric errors below ~ 3%

the fully identified sample. This is not surprising considering the brighter limiting magnitude of the spectroscopic sample compared to that of the SDSS photometry. It can be seen on Fig. 9 that these UVX spectroscopically identified quasars are the most energetic with X-ray luminosities in excess of 10^{44}ergs/s.

For $g - i$ values in the range of 0.8 to 2.0, a population of X-ray sources with $log(f_x/f_r) \sim 0.2$ appears, albeit in smaller numbers than UVX quasars. Most of the optical counterparts are spatially resolved. The SDSS spectroscopic entries occupying this region have attributed galaxy or quasar spectral classes. Their derived X-ray luminosities are in the range of 10^{43-44}ergs/s. This suggests that the vast majority of these reddish objects are likely Seyfert galaxies. Interestingly, only few spectroscopic targets are found in the reddest $g-i \geq 2$ part of the $log(f_x/f_r) \sim 0.2$ branch. We discuss in section 7.3.3 the possible nature of these AGNs.

The rest of the extragalactic X-ray sources are clearly extended in the optical have $g - i$ in the range of 0.8 to 1.3 and $log(f_x/f_r)$ in the range of 0 to −3. Their X-ray luminosity ranges from 10^{42-43}ergs/s, clearly in the AGN domain down to less than 10^{40}ergs/s a luminosity explainable in terms of ULXs, starbursts or of a collection of low-mass X-ray binaries in elliptical galaxies.

7.2. X-ray active stars

At the high galactic latitudes covered here by the cross-correlation ($|b| \gtrsim 20°$with a mean $|b|$ of 58°), most of the X-ray sources are of extragalactic origin. For instance, the Extended Medium Sensitivity Survey which constitutes the largest optically identified sample of serendipitous Einstein X-ray sources at high galactic latitude contains only 25% of active stars (Stocke

et al. 1991). Optical identification campains of ROSAT all-sky survey (RASS) sources yielded similar results (e.g. 35% of active stars in Zickgraf et al. (1997)) while the ROSAT deep survey of the Lockman field (Schmidt et al. 1998) being ∼ 50 times more sensitive that the RASS collected less than 10% of stellar sources. Stars constitute a bounded population of comparatively soft X-ray sources. Consequently, their relative contributions to the high galactic number source number count is expected to decrease very significantly with increasing sensitivity in Chandra and XMM-Newton observations which both offer a lower flux limit and harder energy response (see for instance Fig. 1 in (Hérent et al. 2006)). Active coronae indeed constitute less than 10% of all sources identified in the XMM-Newton serendipitous survey of Barcons et al. (2007). The deepest Chandra surveys such as the Chandra Deep Field North (Alexander et al. 2003) counts only ∼ 3% of stars.

This is at variance with the low latitude situation where most of the RASS X-ray sources were found to be associated with stars (Motch et al. 1997). XMM-Newton and Chandra have extended to lower fluxes the predominance of active coronae on the soft X-ray low latitude source population (Motch et al. 2003; Rogel et al. 2006). At higher energies, a longitude dependent population of galactic hard X-ray sources appears on the top of a usually dominant background of extragalactic sources (Hands et al. 2004; Motch 2006).

Recently, Covey et al. (2008) reported the results of the cross-correlation of the Extended Chandra Wultiwavelength Project (ChaMP) with release 6 of the SDSS. They identify a total of 384 Chandra sources with active stars on the basis of positional coincidence with an unresolved optical source, colours considerations and for a fraction of the sources, confirmation from optical spectroscopy.

As stated above, the number of X-ray emitting and spectroscopically identified stars available in the DR7 is relatively small and because of the scientific goals put forward at the time of the selection of targets for spectroscopic follow-up concentrates towards the reddest M type stars. The availability of the SEGUE archive in DR7 has somewhat increased the number of spectroscopically identified stars but its effect on the cross-correlation statistics remains small. As mentionned in section 6, in order to increase the stellar sample towards earlier types, we used the stellar colour tracks defined by (Covey et al. 2007) and applied a kernel density classification to identify the SDSS R7 / 2XMMi matches having multicolour properties consistent with those expected from reddening free stars of main sequence class. At $g-r$ colour indexes below ∼ 1.5, many active galaxies exhibit multicolour properties which overlap to some extent with those of stars. However the fact that stars show bluer $u-g$ indexes and of course do not appear as extended optical sources allow to efficiently separate them from galaxies up to $g-r \sim 0.2$.

At the mean absolute galactic latitude of 58°, and typical distances of a few hundred parsecs, interstellar absorption remains negligible compared to other uncertainties and has little effect on the observed stellar colours. We computed the total galactic reddening in the directions of each of the X-ray emitting stars following Schlegel et al. (1998). The average E(B-V) is 0.029 with a rms of 0.025. Assuming the absorption coefficients computed by (Girardi et al. 2004) ($T_{\rm eff}$ = 4500K; log g = 4.5), the maximum reddening applicable on average to our sample of stars would be of 0.028 and 0.018 in the $g-r$ and $r-i$ colours respectively. Only very few spatially unresolved SDSS sources matching 2XMMi entries have combined $u-g$, $g-r$ and $r-i$ colours compatible with those expected from giant class III stars, not to mention supergiants. In a recent paper, Guillout (2009) estimate that their sample of $|b| \lesssim 30°$ RASS sources identified with bright Tycho star has a mean contamination of 35% by evolved stars with a peak at 60% for K stars. However, a smaller fraction of X-ray emitting evolved stars of ∼ 10% was present in the sample of Covey et al. (2008) which being at higher galactic latitude and fainter flux is more representative of what we should expect in our case. In the following, we will assume that all stars belong to the main sequence, keeping in mind that a fraction of the class III and class IV stars, in particular in short period binaries such as RS CVn systems could contribute to some extent and being considered as single dwarfs would have computed X-ray luminosities below that actually emitted.

Our sample comprises ∼ 820 active coronae candidates with individual probability of identification above 90%, corresponding to a total sample reliability of 98%. The interval of $g-r$ colour corresponds to late F to late M stars. Neglecting reddening effects, we computed the distances using the absolute magnitude calibration listed in Covey et al. (2007) and X-ray luminosity using the broad band (0.2-12 keV) flux listed in the 2XMMi catalogue for the EPIC pn camera. The mean photometric distance is 570 pc and 90% of the stars have distances in the range of 80 pc to 2300 pc. The overall Log(Lx) distribution peaks at 29.04 and ranges from ∼ 27.7 to ∼ 31.9. This interval of X-ray luminosity covers that exhibited by old stars such as the Sun or even less active, up to that emitted by the most active T Tauri or RS CVn stars. The mean X-ray luminosity does not vary with galactic latitude in the interval covered by our sample.

We show in Fig. 10 the distribution of X-ray active stars in the $g-r$ / $r-i$ diagram for two ranges of X-ray luminosities. It can be readily seen that the locus of X-ray bright sources is shifted by ∼ 0.1 mag in colour above that occupied by low L_X active coronae. Active stars appear to be bluer in $g-r$ for a given $r-i$. This shift cannot be due to the lack of reddening correction since the most X-ray luminous stars which are expected to be the most remote and absorbed ones should appear redder than the closest low L_X stars, a trend opposite to what is observed. We also checked with the isochrones of Girardi et al. (2004) designed for the SDSS band passes that age or metallicity effects were unable to explain the different colour/colour tracks of the low and high L_X stars. Similarly, an enhanced Hα emission expected to be especially important for late M stars ($g-r > 1.3$) can be excluded since it would yield larger $g-r$. Interestingly, Covey et al. (2008) report that the counterparts of their extended ChaMP stellar X-ray survey do exhibit a ∼ 0.1 bluer $u-g$ index than average low mass stars, although they do not report any significant change in the $g-r$ colour index. Comparing with their work is however not straightforward. While the colour/colour track followed on average by our active coronae is consistent with that of Covey et al. (2007) for the entire SDSS and therefore does not indeed contradicts the results of (Covey et al. 2008), the difference we find rather arises among two groups of active stars. The SDSS M stars templates compiled by Bochanski et al. (2007) might indicate a similar trend in the $u-g$ colour index between active and inactive stars with however a large intrinsic scatter, while no such effect occurs in $g-r$. Unfortunately, we have too few good u band measurements to be able to confirm the $u-g$ trend seen by Covey et al. (2008). We agree with these authors that low-level optical flaring might be responsible for the bluer colours seen in X-ray active stars.

Active coronae emit essentially thin thermal spectra dominated by a series of narrow emission lines superposed on a weak continuum. In most cases, two thermal components are required to satisfactorily represent the observed energy distribution (see a recent review in Guedel & Naze 2009). X-ray studies of open

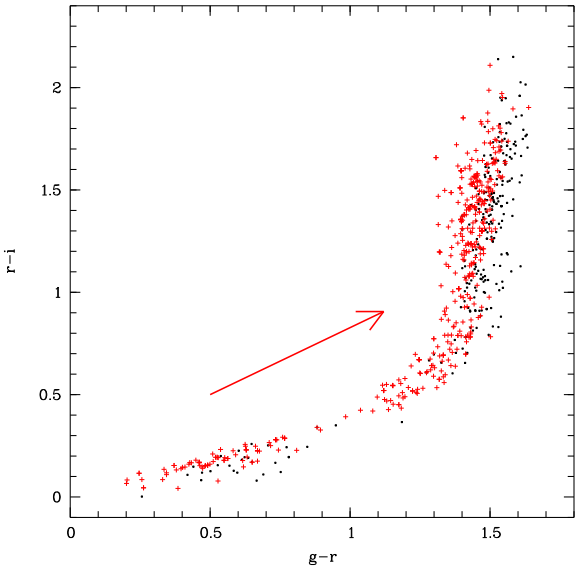


Fig. 10. the distribution of X-ray active stars in the $g-r$ / $r-i$ diagram for two ranges of X-ray luminosities. Black dots Log(Lx)<29. Red crosses Log(Lx)>29. The red arrow show the direction of interstellar reddening. Only sources with photometric errors below 0.1 mag in g are shown.

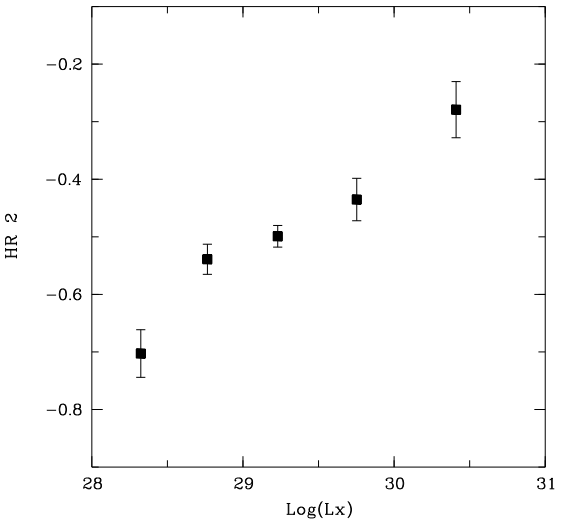


Fig. 11. Variation of the EPIC pn Hardness ration 2 with mean X-ray luminosity for the 305 sources with errors on HR2 below 0.2. We plot here the median and associated errors.

clusters and field stars of different ages led to a relatively coherent picture linking stellar rotation rates, overall X-ray luminosity and X-ray temperatures. Whereas the young (age < 1 Myr) stars in Orion exhibit X-ray spectra with kT_1 ~ 0.8 keV and kT_2 ~ 2.9 keV, the analysis of ~ 115 Myr old Pleiades stars yields kT_1 ~ 0.4 keV and kT_2 ~ 1.1 keV while the X-ray corona of our Sun can also be characterized by 2T spectrum with $kT1$ ~ 0.2 keV and $kT2$ ~ 0.6 keV (see Sung et al. 2008, and references therein). Using ROSAT and ASCA observations of a dozen of carefuly selected stars, Guedel et al. (1997) established that the overall X-ray luminosity, the temperature of the two components and the emission measurement ratio of the hot to the cool plasma were all decreasing with age and rotation rate. The range of X-ray luminosity observed in our survey suggests ages younger than ~ 2 Gyr.

We thus investigated whether the X-ray properties of our identified active coronae were depending on luminosity. For that purpose we selected the 305 identified stellar X-ray sources located at distances between 80 and 2300 pc and having errors on 2XMMi hardness ratio 2 of less than 0.2. This sample was then split into five range of luminosities for which the mean HR2 was computed. Hardness ratio 2 measures the relative count rates in the energy ranges 0.5-1.0 keV and 1.0-2.0 keV. It is therefore weakly dependent on the EPIC filter used and varies little within the range of N_H applicable to the present survey. HR2 can thus be considered as an indicator of the intrinsic shape of the X-ray energy distribution well suited to the range of temperatures exhibited by stellar coronae. We show in Fig. 11 the variation of the median HR2 with X-ray luminosity. A clear spectral hardening accompanies the luminosity increase. Assuming a single temperature plasma undergoing $N_H \sim 1.7 \times 10^{20}$ cm^{-2} would imply a thin thermal temperature of ~ 0.45 for the lowest L_X bin and 0.72 keV for the largest L_X. However, the mean third hardness ratio which compares count rates in the 1.0-2.0 keV and 2.0-4.5 keV ranges has a mean value of ~ -0.75, significantly harder than would be expected from the HR2 value assuming a one temperature model. The existence of a second hotter thin thermal component naturally accounts for this discrepancy. Unfortunately, using only two hardness ratios, it is impossible to fit both kT_1, kT_2 and the ratio of the emission measurements of the hot and cool components. In spite of this shortcoming, our data which benefit from the large throughput of XMM-Newton and of its capability to obtain detailed spectral information for an unprecedented number of sources confirm for the first time the spectral hardening with X-ray luminosity in groups of mixed age field objects.

7.3. Active galaxies and quasars

A number of optical identification campaigns of XMM-Newton X-ray sources are currently underway and have already led to several important scientific results. As for galactic studies, the high throughput of the EPIC cameras provides unprecedented X-ray spectral information on relatively faint sources over a wide energy range extending up to ~ 12 keV. The Survey Science Center itself drives several distinct follow-up identification programmes aimed at studying the nature of the extragalactic X-ray sources detected at high galactic latitudes in several ranges of X-ray flux. The first survey is the Bright Sources Survey (BSS, Della Ceca et al. 2004) which focuses on 400 sources brighter than 7×10^{-14}erg cm^{-2} s^{-1} in the 0.5-4.5 keV band and in the harder 4.5-7.5 keV band. Caccianiga et al. (2007) show that the fraction of obscured objects varies significantly with the energy range in which source selection is done. Carried out tt intermediate X-ray fluxes of a few 10^{-14}erg cm^{-2} s^{-1} in the 0.5-4.5 keV band the XMM-Newton Medium Sensitivity Survey (XMS, Barcons et al. 2002; Carrera et al. 2007; Barcons et al. 2007) contains 318 sources with a high rate of identification of ~ 90%. Several surveys exploring the nature of sources as faint as 10^{-15}erg cm^{-2} s^{-1} up to a factor 100 brighter are reported in the literature. They necessarily cover a small field of view usually restricted to a handful of XMM-Newton pointings. Among these

are these are the 695 sources of the COSMOS field (Brusa et al. 2007b) and the 328 sources of the Marano field (Krumpe et al. 2007). The slightly brighter HELLAS2XMM survey (Cocchia et al. 2007) contains 220 XMM-Newton sources.

The main issues dealt with by these studies concern the definition of log N(>S)-log S curves at faint fluxes and in the hard X-ray bands where most AGN responsible for the diffuse X-ray background are expected to be found. The search for type II QSOs remains a topical activity as is the search for associated X-ray and optical signatures. An unexpectedly large population of X-ray bright optically normal galaxies (XBONGs) is emerging and the reasons why nuclear activity does not show up at optical wavelengths remains challenging.

At the time we write this article, only a handful of papers report studies based on a cross-correlation of X-ray data with the SDSS. Recently, Young et al. (2008) cross-matched the third release of the SDSS with the XMM-Newton archive. They selected a sample of 17 red quasars at moderate redshift and suggest that low accretion rates could explain the intrinsically red optical continua. The X-ray properties of the most optically luminous quasars were investigated by Just et al. (2007) with the help of dedicated Chandra observations and the DR3 of the SDSS. The authors are able to constrain the possible evolution of the spectral slope with X-ray luminosity and redshift.

Most 2XMMi sources match with SDSS-DR7 entries close to the limiting magnitude of the SDSS survey (mag $\sim$ 22) and only a relatively small fraction is bright enough to have been selected for spectroscopic observations. For instance, over a grand total of 60567 2XMMi detections matching a SDSS DR7 entry, 87% have an error on their r magnitude below 0.2 but only 12% are SDSS spectroscopic targets as well. In addition, the repartition of the spectroscopic targets are far from covering uniformly the parameter space spun by the optical counterparts of the serendipituous XMM-Newton sources. According to Abazajian et al. (2009) SDSS DR7 spectroscopic targets arise from the need to explore particular classes of objects such as bright galaxies, luminous red ellepticals, counterparts of ROSAT X-ray sources, etc. For our purpose, the two most important groups of spectroscopic targets are the sample of quasar candidates defined by Richards et al. (2002) and all the stars belonging to the SEGUE programme (Yanny et al. 2009) which aims at studying the galactic structure. The quasar sample is essentially a classically UV excess (UVX) selected sample to which is added a small number of redder targets appearing as likely high redshift QSOs. The distribution in the $g - i$ / f_x/f_r diagram of the SDSS DR7 photometric entries extends much further to the red than covered by the spectra of UVX quasars and only few spectroscopi targets classified as either quasars or galaxies do sample the red part of the area covered by the photometric sample (see Figs. 7, 8 and 9). Furthermore, SDSS entries with optical spectra are necessarily optically brighter that the bulk of the matches with those without associated spectroscopy and cannot bring information on remote high f_x/f_r objects. It should ths be stressed that although SDSS spectroscopy can shed light on the nature of the most frequent optically bright X-ray emitters encountered at high galactic latitude and that, because of the large area covered, has the capability to discover rare outliers, it cannot alone be used to make a complete census of the high $|b|$ X-ray population, nor to study evolutionary effects affecting the nature of the X-ray sources optically fainter than the SDSS limit. The huge merit of the dedicated identification programmes such as the ones carried out by the SSC is to extend spectroscopic identifications to very low optical fluxes and thus offer a unique opportunity to unveil the different populations of extragalactic sources which may appear at fainter optical fluxes and in general at higher X-ray to optical flux ratios. The two strategies, wide and shallow on one hand and narrow and deep on the other hand are in fact quite complementary and suited to best characterize and scientifically mine the entire serendipitous XMM-Newton catalogues.

However, the good photometric quality of the Sloan Digital Sky Survey and the availability of numerous parameters on emission and absorption lines detected in the spectroscopic SDSS paths the way to a large number of studies based on correlations between X-ray and optical properties. It is obviously far beyond the scope of this paper to address all possible study avenues and we will only explore in a shallow manner some of the most obvious correlations unveiled by this crosscorrelation.

7.3.1. Type II QSO candidates

Low luminosity obscured AGNs (Seyfert II galaxies) are detected in large numbers in the local universe. According to the standard unification model, absorption by the torus seen at high inclination blocks the direct view to the central engine, preventing the detection of the broad line region at optical wavelengths and suppressing the soft part of the X-ray spectrum by photoelectric absorption. Soft X-rays may however leak out from the AGN through multiple scattering. The intensity of the hard X-ray background requires the existence of a sizable population of obscured AGNs. Type II quasars, the high luminosity analogs of Seyfert II galaxies could indeed account for a large fraction of this background.

Taking advantage of the large SDSS spectroscopic database Zakamska et al. (2003) have identified a sample of 291 type II AGN candidates in the redshift range from 0.3 to 0.83. The selection criterion was the existence of narrow emission lines with high equivalent widths, high ionization line ratios and strong OIII 5007Å emission used as a proxy of the nuclear luminosity. A small fraction of these candidates has been observed in X-rays using dedicated or serendipitous Chandra and XMM-Newton observations (Ptak et al. 2006) and the corresponding 2XMMi entries are naturally recovered in our cross-correlation. We list in Table 8 the main properties of the seven type II candidates matching a serendipitous EPIC source. The X-ray luminosities listed here being computed assuming an average shape for the large band 0.2 to 12 keV energy distribution are necessarily less accurate than those derived by Ptak et al. (2006) from a detailed spectral analysis. The four candidates in our list common with the eight discussed in Ptak et al. (2006) have luminosities consistent within a factor of two.

It can be seen that, ignoring the two low L_X candidates which are likely normal Seyfert II galaxies, hardness ratio 2 of high L_X type II candidates appears systematically hard, while the segregation in terms of optical colours is much less evident.

7.3.2. Variations of X-ray hardness ratios with luminosity, galaxy type and redshift

We show in Fig. 14 the behaviour of three EPIC pn hardness ratios with X-ray luminosity for all spectroscopic SDSS targets. The bulk of the SDSS QSOs with X-ray luminosities larger than 10^{44}erg s^{-1} (0.2-12 keV) cluster around hardness ratios (<HR2> = -0.12; <HR3> = -0.38 and <HR4> = -0.28 in excellent agreement with the values expected from a canonical Γ = 1.9 power law X-ray spectrum undergoing negligible local absorption and a mean galactic absorption of 1.16×10^{21}cm^{-2} (the average over all directions of galaxy and QSO targets). For this group

Table 8. Type II quasar candidates from Zakamska et al. (2003) matching 2XMMi entries

2XMMi name	id prob	z	u-g	g-i	HR 2 (pn)	Log(L_X)
2XMM J005621.6+003235	+0.964	+0.484	1.891 ±0.643	+1.800 ±0.061	-0.56 ±0.28	+42.78
2XMMi J011522.2+001518	+0.997	+0.390	6.023 ±1.239	+2.437 ±0.046	+0.73 ±0.04	+44.26
2XMM J015716.9-005305	+0.988	+0.540	+.565 ±0.167	+1.836 ±0.050	-0.18 ±0.46	+42.78
2XMM J021047.0-100152	+0.999	+0.540	+.348 ±0.136	+1.771 ±0.047	+0.66 ±0.12	+44.40
2XMM J103951.5+643005	+0.991	+0.402	+.490 ±0.075	+1.209 ±0.027	-0.68 ±0.33	+43.02
2XMM J122656.4+013124	+0.998	+0.732	+.970 ±0.256	+1.134 ±0.056	+0.76 ±0.06	+44.83
2XMM J164131.6+385841	+0.995	+0.596	+.550 ±0.112	+1.768 ±0.028	+0.63 ±0.05	+44.96

of QSOs, there is no evidence of strong evolution of the powerlaw index with L_X. There is however a small number of QSOs exhibiting a considerably harder X-ray spectrum. Another interesting point is that some very luminous X-ray sources seem associated with otherwise normal or at least narrow line galaxies and are thus good type II QSO and BL Lacs candidates or cases of extreme XBONGs.

As one enters the AGN regime at X-ray luminosities below 10^{44}erg s^{-1}, the number of extend sources with galaxy like spectra rises considerably and the shape of the X-ray energy distribution shows a much larger scatter. However, the figure involving HR3 reveals the existence of a rather well defined group of active galaxies with $L_X$as low as a few 10^{41}erg s^{-1}clustering around the same hardness ratio as orders of magnitude more luminous quasars. It should be kept in mind that the larger HR dispersion seen for low L_X AGN is likely due to the strong bias resting in the selection of the spectroscopic SDSS sample since quasar candidates were mostly selected on the basis of an UV excess (Richards et al. 2002). Consequently, the spectroscopic QSO catalogue will likely not sample well the overall population of X-ray detected AGN. As shown in Fig. **??** a large number of unresolved optical sources displaying $g - i$ indexes redder than ~ 1 have f_x/f_r ratio in the same range of the bluer spectroscopic QSO sample and are thus likely "red" quasars.

A marked spectral transition occurs at X-ray luminosities below $\sim 10^{41-42}$erg s^{-1}, i.e. when X-ray emission from the central black hole becomes less likely to dominate the high energy output of the entire galaxy. The combined emission of star forming regions, super novae remnants and both massive and low-mass X-ray binaries appears in many cases significantly softer than those of classical AGN. Because of their softer spectrum and lower count rates, low L_X galaxies have no well enough measured HR4 and are therefore missing in the bottom panel of Fig.14.

The four high X-ray luminosity quasars extracted from the sample of Zakamska et al. (2003) all exhibit extreme hard HR2 whereas their distribution in HR4 appears much more alike that of type I QSOs. The very positive HR2 probably reflects the relatively large intrinsic photoelectric absorption present in many type II quasars and responsible for their preferential discovery in hard X-ray surveys (see e.g. Caccianiga et al. 2004). X-ray spectral fits do reveal indeed considerable N_H of the order of 10^{22}cm^{-2} or more with a powerlaw index,and hence an HR4, similar to that of type I quasars.

We also investigated whether the mean spectral shape measured by the hardness ratios would show evidences for a possible evolution with redshift. Fig. 12 shows that no detectable spectral evolution occurs with look back time, at least till $z \sim 3$. The absence of evolution of the powerlaw spectral index with both X-ray luminosity and redshift has been extensively studied by ...

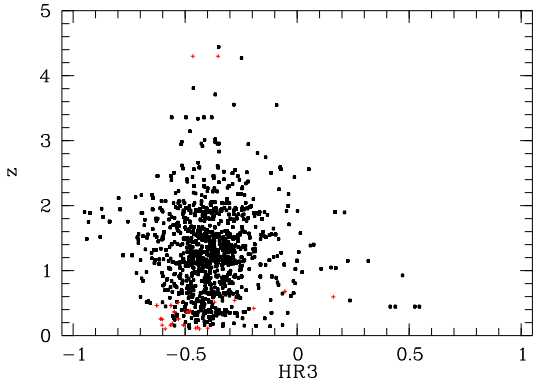


Fig. 12. Variation of the EPIC pn Hardness ratios with the redshift for both spectroscopic SDSS quasars (black squares) and galaxies (red crosses). We only keep here sources with errors on HR3 less than 0.2. We only consider QSOs with $L_X$in the range of 10^{44} to 10^{46} erg s^{-1}

7.3.3. Characterizing specific AGN populations

Type II quasar candidates: A number of spectroscopic SDSS entries occupy the same region of the L_X HR2 diagram as Zakamska's candidates (see Fig. 14). We explored the nature of these candidates by selecting objects with log(L_X) larger than 44, EPIC pn HR2 larger than 0.5 with an error of less than 0.2 on the hardness ratio. Twelve objects match these conditions among which only 2XMM J134507.9-001901 / SDSS J134507.93-001900.9 at $z = 0.419$, $L_X \sim 3 \times 10^{44}$erg s^{-1} could rank as a likely new type II quasar. The rest of the selected sources are described in the literature as type I AGN at high redshift quasars or showing broad absorption lines or reddening by dust or no specific comment. It is thus clear that although type II quasars do separate well in the L_X / HR2 diagram other rare kinds of object occupy the same locus.

Red objects with high f_x/f_r : As seen on Fig.**??**, a large numbers of matches fill in a region of relatively high $Log(f_x/f_r)$ ratio around 0 while exhibiting red $g - i$ colours. Although they share similar X-ray to optical flux ratios as the bulk of the UVX selected quasars, only few of these sources were selected for SDSS spectroscopic follow-up. Selecting spectroscopic targets with $Log(f_x/f_r) > -0.2$, $g - i > 1.5$ with errors of less than 0.5 on the colour index returns 30 entries among which only 6 are flagged as being spatially unresolved. Half of them are classified as Seyfert 1 and only 10% apear as Seyfert 2, the rest being made of BL Lac candidates, high redshift QSOs and unclassified objects. In several cases, the underlying red spectrum of the Galaxy is clearly detected in SDSS spectra. A total of 3817 2XMMi/DR7 photometric matches occupies this parameter space, representing about 10% of the total number of matches assuming as usual a probability of identification larger than 0.9.

Among these 85% of the optical counterparts are flagged as being extended. It is thus very likely that for the majority of these high f_x/f_r their red colours arises from the probable contamination of moderately luminous nucleus light by the reddish emission from the host galaxy. A fraction of these sources and especially those appearing as point-like in the optical are likely to be high redshift quasars or BL LACs.

High X-ray luminosity AGN We tried to explore the nature of the high L_X AGN following two distinct paths of investigation.

In a first step, we only considered SDSS spectroscopic entries classified as having a "galaxy" spectrum and associated to X-ray sources with luminosities above 10^{44}erg s^{-1}. A total of 25 distinct SDSS entries fulfill these constraints. In two very bright cases, at least, the high L_X arises from an erroneous spectroscopic redshift. However, the vast majority of these X-ray sources are either known BL Lacs or very likely new BL Lac candidates, based on their rather featureless optical spectrum and association with a luminous radio source. We also recover the type II quasar candidate 2XMM J134507.9-001901 found in the L_X/ HR2 selection above.

The second step was then to investigate the nature of the reddish ($g-i>1$; err$(g-i)<0.5$) and high luminosity ($L_X > 10^{45}$erg s^{-1}) AGNs independently of their SDSS determined spectral class (galaxy or quasar). Among the 27 sources returned by this selection, 20 have a redshift larger than 3, while the seven remaining targets all have a redshift of less than 2. This surprising dichotomy probably reflects the different strategy applied to select the spectroscopic targets. At least three of the lower redshift objects were investigated by Young et al. (2008) who conclude that their red colours are due to absorption by dust.

BAL QSO candidates

We show in Fig.13 the distribution in the $u-g$ / $g-i$ diagram of the SDSS spectroscopic identification of 2XMMi sources for various ranges of X-ray luminosity together with the position of the photometric matches. The diagram highlights again the existence of a numerous and reddish in $g-i$ population of intermediate luminosity AGNs which is likely to consist of Seyfert galaxies for which the long wavelength $g-i$ colour index is strongly modified by the contribution of the host galaxy. In contrast, the $u-g$ colour index is more consistent with those of the "classical" UVX selected quasars. Galaxies with X-ray luminosities below 10^{42}erg s^{-1}follow the colour/colour relation exhibited by "normal" galaxies (Strateva et al. 2001) while galaxies somewhat brighter in X-ray are located slightly offset from the main colour colour track. The group of photometric entries located in the reddest part of the diagram is mostly made of stellar identifications.

A distinct branch of spectroscopic targets of high X-ray luminosity appears in the $u-g$ range from 1 to 4 and in the $g-i$ range from 0 to 1. These quasars share the $g-i$ colour indexes of UVX AGNS but exhibit extremely red $u-g$ colours. Most of these AGN turn out to be known or new BAL QSOs. A sizable number of photometric entries fall in the same region and are therefore good candidates for strongly absorbed BAL quasars.

Unfortunately, there is no firm evidence showing that QSOs with hard X-ray spectra preferentially display red colours. We show in Fig. 15 the best evidence in favour of a possible correlation between the $u-g$ index and a hard HR3 since the X-ray hardest QSOs turn out to be very red. However, many red spectroscopic QSOs and QSOs candidates have red $u-g$ colours while displaying otherwise normal spectra. Other combinations of colours and hardness ratio yield even less convincing evidences.

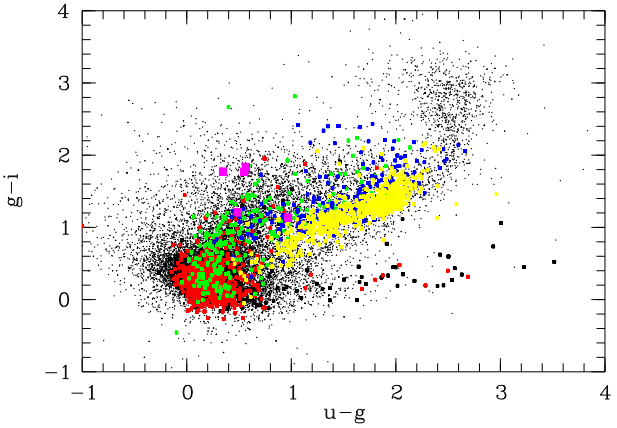


Fig. 13. The distribution of the various 2XMMi/DR7 cross matches in the $u-g/g-i$ diagram. We plot here only sources with errors on colour indexes below 0.5. For spectroscopic SDSS entries (squares), the colour codes the range of X-ray luminosity. Black; $>10^{45}$, red 10^{44} to 10^{45}, green 10^{43} to 10^{44}, blue 10^{42} to 10^{43}, yellow 10^{40} to 10^{42}. Type II QSO candidates from Zakamska et al. (2003) are shown in magenta. Black points show the position of the 2XMMi matches with SDSS photometric catalogue

8. Conclusions

We present the results of the cross-identification of the 2XMMi catalogue containing over 220,000 XMM-Newton EPIC serendipitous with the Data Release 7 of the Sloan Digital Sky Survey consisting of 357 million unique objects and over 1.6 million spectra. In order to have the best statistical control on the quality of the cross-correlation process, we use a likelihood ratio scheme only based on the probability of spatial coincidence of the X-ray source with the optical candidate. Using an original method which does not resort to heavy Monte Carlo simulations, we are able to compute true probabilities of identifications taking into account the varying fraction of X-ray sources expected to have a SDSS match as function of galactic latitude. We discuss the details of the statistical method used and present reliability and completeness curves for the entire set of matches. A total of 30,000 unique X-ray sources have a match in the DR7 with a probability of identification above 90%. At this threshold, the reliability of the total sample is of 98%, i.e. we expect only 2% of spurious cross-identifications while the completeness is 77% meaning that we miss about a quarter of the true macthes which therefore appear with individual probabilities of identifications below the 90% threshold.

One of the most ambitious task attributed to the Survey Science Center of the XMM-Newton satellite is the statistical identification of all serendipitous X-ray sources discovered in the large field of view of the EPIC cameras. The 2XMMi/DR7 cross-identified sample offers an interesting opportunity to test various strategies and methods which would eventually led to the classification of the XMM catalogues. Pineau et al. (2008) have presented a first attempt to reduce the parameter space dimension using Principal Component Analysis tools adapted to data affected by rather large instrumental errors and taking into account the intrinsic scatter of the measured quantities. These authors also test the relative merits of various classification methods.

We use the SDSS DR7 spectroscopic catalogue to build a learning sample made of objects of known nature to which we

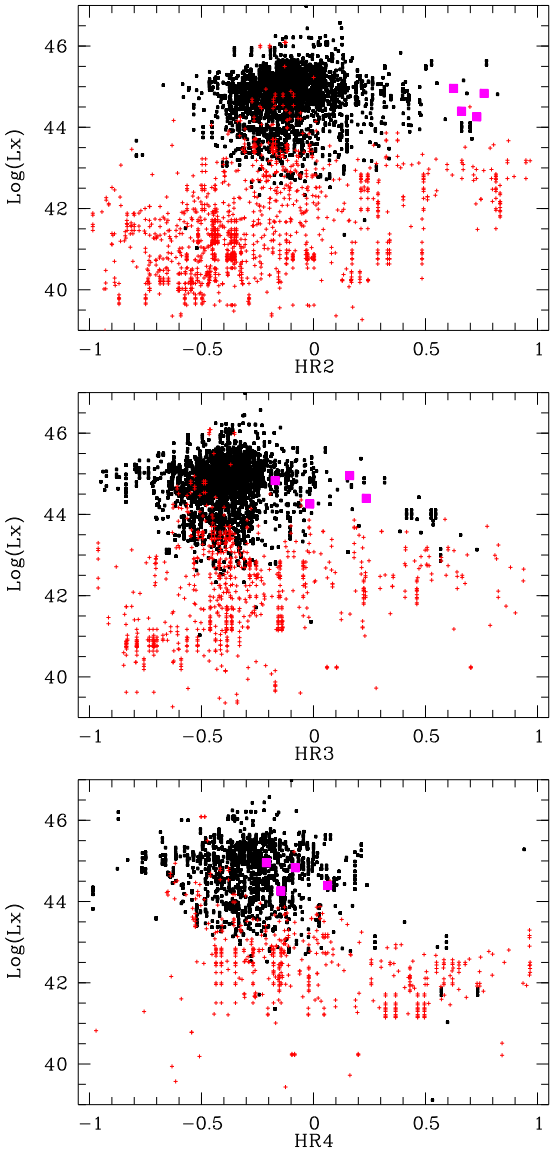


Fig. 14. Variation of the EPIC pn Hardness ratios with X-ray luminosity for spectroscopic SDSS targets classified as quasars (black dots) and galaxies (red crosses). M agenta filled squares are Zakamska et al. type II QSO candidates. Errors on hardness ratios are below 0.2 in all cases.

add a few entries extracted from specific additional catalogues (e.g. Cataclysmic Variables). We find that the most discriminating diagrams involve the f_x/f_r flux ratio and various other parameters such as the $u-g$ or $g-i$ colour indexes and the extent of the optical source. Active galactic nuclei more luminous than 10^{42}erg s^{-1} have f_x/f_r ratios well above those of most galactic stars in their range of $g-i$. Galaxies of lower X-ray luminosity overlap the region of the diagram covered by active stars but can be easily separated from them on the basis of the optical extent.

X-ray bright stars appear slightly bluer in $g-r$ than less active coronae. The origin of this effect is not clear, but could be related to the UV and blue micro flaring occurring in many of the active late type K and M stars. Assuming that most active stars are indeed main sequence, we find a strong dependency

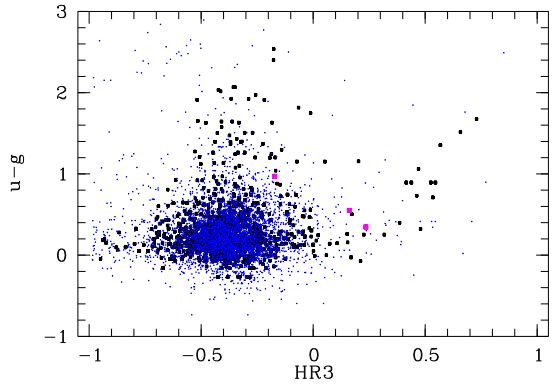


Fig. 15. Variation of the EPIC pn Hardness ratios with the $u-g$ colour index for both spectroscopic SDSS quasars (black squares), Zakamska et al. type II QSO candidates (magenta squares) and unresolved objects from the photometric catalogue. We only keep here sources with errors on HR3 less than 0.2 and errors on $u-g$ below 0.3. We also ignored sources with $\log(f_x/f_r)$ less than -1 in order to exclude the stellar contribution.

of the hardness ratio 2 with the inferred X-ray luminosity, the most luminous coronae being those emitting with the highest temperature.

Not unexpectedly, active galactic nuclei constitute by far the most numerous class of X-ray emitters present in the cross-correlation. In general, optical and X-ray AGN properties mainly depend on the X-ray luminosity (or on the f_x/f_r ratio) and on optical colours. UV excess quasars are the most frequent X-ray emitters and cluster in a well defined area in the f_x/f_r / $g-i$ diagram. As L_X decreases, the $g-i$ colour index becomes redder and f_x/f_r decreases. A rather large group of optically extended X-ray sources are found at f_x/f_r comparable to those of UVX quasars, but much redder $g-i$ colours.

Search for type II QSOs, predominance of BL Lacs at high f_x/f_r, position of high z QSOs, BAL QSOs, etc...

References

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543
Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., et al. 2008, ApJS, 175, 297
Alexander, D. M., Bauer, F. E., Brandt, W. N., et al. 2003, AJ, 126, 539
Barcons, X., Carrera, F. J., Ceballos, M. T., et al. 2007, A&A, 476, 1191
Barcons, X., Carrera, F. J., Watson, M. G., et al. 2002, A&A, 382, 522
Bochanski, J. J., West, A. A., Hawley, S. L., & Covey, K. R. 2007, AJ, 133, 531
Boller, T., Bertoldi, F., Dennefeld, M., & Voges, W. 1998, A&AS, 129, 87
Brusa, M., Zamorani, G., Comastri, A., et al. 2007a, ApJS, 172, 353
Brusa, M., Zamorani, G., Comastri, A., et al. 2007b, ApJS, 172, 353
Budavari, T. & Szalay, A. S. 2007, ArXiv e-prints, 707
Caccianiga, A., Severgnini, P., Braito, V., et al. 2004, A&A, 416, 901
Caccianiga, A., Severgnini, P., Della Ceca, R., et al. 2007, A&A, 470, 557
Carrera, F. J., Ebrero, J., Mateos, S., et al. 2007, A&A, 469, 27
Cocchia, F., Fiore, F., Vignali, C., et al. 2007, A&A, 466, 31
Comaniciu, D. & Meer, P. 2002, IEEE Trans. Pattern Anal. Mach. Intell., 24, 603
Comaniciu, D., Ramesh, V., & Meer, P. 2001, Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, 1, 438
Covey, K. R., Agüeros, M. A., Green, P. J., et al. 2008, ApJS, 178, 339
Covey, K. R., Ivezić, Ž., Schlegel, D., et al. 2007, AJ, 134, 2398
Cutri, R. M., Skrutskie, M. F., van Dyk, S., et al. 2003, 2MASS All Sky Catalog of point sources. (The IRSA 2MASS All-Sky Point Source Catalog, NASA/IPAC Infrared Science Archive. http://irsa.ipac.caltech.edu/applications/Gator/)
De Ruiter, H. R., Arp, H. C., & Willis, A. G. 1977, A&AS, 28, 211
Della Ceca, R., Maccacaro, T., Caccianiga, A., et al. 2004, A&A, 428, 383
den Herder, J. W., Brinkman, A. C., Kahn, S. M., et al. 2001, A&A, 365, L7

Downes, R. A., Webbink, R. F., Shara, M. M., et al. 2001, PASP, 113, 764
Girardi, L., Grebel, E. K., Odenkirchen, M., & Chiosi, C. 2004, A&A, 422, 205
Guedel, M., Guinan, E. F., & Skinner, S. L. 1997, ApJ, 483, 947
Guedel, M. & Naze, Y. 2009, ArXiv e-prints
Guillout, P. 2009, A&A, 0, xxx
Hands, A. D. P., Warwick, R. S., Watson, M. G., & Helfand, D. J. 2004, MNRAS, 351, 31
Hérent, O., Motch, C., & Guillout, P. 2006, in ESA Special Publication, Vol. 604, The X-ray Universe 2005, ed. A. Wilson, 91–+
Jansen, F., Lumb, D., Altieri, B., et al. 2001, A&A, 365, L1
Just, D. W., Brandt, W. N., Shemmer, O., et al. 2007, ApJ, 665, 1004
Krumpe, M., Lamer, G., Schwope, A. D., et al. 2007, A&A, 466, 41
Lauberts, A. 1982, ESO/Uppsala survey of the ESO(B) atlas, ed. A. Lauberts
Mason, K. O., Breeveld, A., Much, R., et al. 2001, A&A, 365, L36
Motch, C. 2006, in ESA Special Publication, Vol. 604, The X-ray Universe 2005, ed. A. Wilson, 383–+
Motch, C., Guillout, P., Haberl, F., et al. 1997, A&A, 318, 111
Motch, C., Herent, O., & Guillout, P. 2003, Astronomische Nachrichten, 324, 61
Motch, C., Michel, L., & Pineau, F. X. 2007, in Astronomical Society of the Pacific Conference Series, Vol. 376, Astronomical Data Analysis Software and Systems XVI, ed. R. A. Shaw, F. Hill, & D. J. Bell, 699–+
Oyabu, S., Yun, M. S., Murayama, T., et al. 2005, AJ, 130, 2019
Pineau, F.-X., Derriere, S., Michel, L., & Motch, C. 2008, in American Institute of Physics Conference Series, Vol. 1082, American Institute of Physics Conference Series, 15–21
Ptak, A., Zakamska, N. L., Strauss, M. A., et al. 2006, ApJ, 637, 147
Richards, G. T., Fan, X., Newberg, H. J., et al. 2002, AJ, 123, 2945
Ritter, H. & Kolb, U. 2003, A&A, 404, 301
Rogel, A. B., Lugger, P. M., Cohn, H. N., et al. 2006, ApJS, 163, 160
Rutledge, R. E., Brunner, R. J., Prince, T. A., & Lonsdale, C. 2000, ApJS, 131, 335
Sain, S. R. 2002, Comput. Stat. Data Anal., 39, 165
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Schmidt, M., Hasinger, G., Gunn, J., et al. 1998, A&A, 329, 495
Sinnott, R. W. 1984, S&T, 68, 158
Stephen, J. B., Bassani, L., Molina, M., et al. 2005, A&A, 432, L49
Stocke, J. T., Morris, S. L., Gioia, I. M., et al. 1991, ApJS, 76, 813
Strateva, I., Ivezić, Ž., Knapp, G. R., et al. 2001, AJ, 122, 1861
Strüder, L., Briel, U., Dennerl, K., et al. 2001, A&A, 365, L18
Sung, H., Bessell, M. S., & Sana, H. 2008, Journal of Korean Astronomical Society, 41, 1
Sutherland, W. & Saunders, W. 1992, MNRAS, 259, 413
Turner, M. J. L., Abbey, A., Arnaud, M., et al. 2001, A&A, 365, L27
Tyson, J. A. 2002, in Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, Vol. 4836, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, ed. J. A. Tyson & S. Wolff, 10–20
Véron-Cetty, M.-P. & Véron, P. 2006, A&A, 455, 773
Watson, M. G., Augueres, J.-L., Ballet, J., et al. 2001, A&A, 365, L51
Watson, M. G., Schröder, A. C., Fyfe, D., et al. 2009, A&A, 493, 339
Weisskopf, M. C., Tananbaum, H. D., Van Speybroeck, L. P., & O'Dell, S. L. 2000, in Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, Vol. 4012, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, ed. J. E. Truemper & B. Aschenbach, 2–16
Werner, M. W., Roellig, T. L., Low, F. J., et al. 2004, ApJS, 154, 1
Wolstencroft, R. D., Savage, A., Clowes, R. G., et al. 1986, MNRAS, 223, 279
Yanny, B., Rockosi, C., Newberg, H. J., et al. 2009, AJ, 137, 4377
Young, M., Elvis, M., & Risaliti, G. 2008, ApJ, 688, 128
Zakamska, N. L., Strauss, M. A., Krolik, J. H., et al. 2003, AJ, 126, 2125
Zhang, X., King, M. L., & Hyndman, R. J. 2006, Computational Statistics & Data Analysis, 50, 3009
Zickgraf, F.-J., Thiering, I., Krautter, J., et al. 1997, A&AS, 123, 103

Acknowledgements. Part of this work was supported by the *CNRS* and the *Region Alsace*.

Appendix A: Transformation of error ellipses in the new frame: For SSC members information (will not be published)

A.1. Spherical trigonometry

On the unit sphere of center C, we consider the spherical triangle NXO where N is the north pole, X and O are the positions of the X-ray and the optical sources respectively. Let us write:

- x, n, o the angular distances between N and O, O and X, N and X respectively.
- $\hat{N}$, $\hat{X}$ and $\hat{O}$ the dihedral angles between planes NCX and NCO, XCN and XCO, OCN and OCX respectively. Those angles in the range $]-180\,^\circ, 180\,^\circ]$ and are define as positive in the East direction.

We have immediately $\hat{N} = \alpha_O - \alpha_X$, $x = 90\,^\circ - \delta_O$, $o = 90\,^\circ - \delta_X$ and $n = d$ the angular distance between the two sources which is given by the Haversine formula:

$$d = 2 \cdot \arcsin \sqrt{\sin^2 \frac{\delta_o - \delta_X}{2} + \sin^2 \frac{\alpha_o - \alpha_X}{2} \cos\delta_X \cos\delta_o} \quad \text{(A.1)}$$

The spherical trigonometry gives the formula:

$$\frac{\sin\hat{N}}{n} = \frac{\sin\hat{X}}{x} = \frac{\sin\hat{O}}{o} \quad \text{(A.2)}$$

As n, x and o are positive, we notice that $\hat{N}$, $\hat{X}$ and $\hat{O}$ have the same sign, which depends on $\alpha_O - \alpha_X$. As the function *arcsin* returns values $\in [-90, +90]$, we deduce two possible values for both angle $\hat{X}$ and $\hat{O}$. The cosine of half-angles $\hat{X}/2$ and $\hat{O}/2$ enable to know if $|\hat{X}| > 90\,^\circ$ and if $|\hat{O}| > 90\,^\circ$ respectively. From the spherical trigonometry formulae:

$$s = \frac{1}{2}(n + x + o) \quad \text{(A.3)}$$

$$\cos\frac{\hat{X}}{2} = \sqrt{\frac{\sin s \sin(s-x)}{\sin n \sin o}} \quad \text{(A.4)}$$

$$\cos\frac{\hat{O}}{2} = \sqrt{\frac{\sin s \sin(s-o)}{\sin n \sin x}} \quad \text{(A.5)}$$

We derive the values of angles $\hat{X}$ and $\hat{O}$:

$$\hat{X} = \begin{cases} 180 - \arcsin\frac{x\cdot\sin\hat{N}}{n} & \text{if} \quad \cos\frac{\hat{X}}{2} > \frac{\sqrt{2}}{2} \quad \text{and} \quad \arcsin\frac{x\cdot\sin\hat{N}}{n} \geq 0 \\ -180 - \arcsin\frac{x\cdot\sin\hat{N}}{n} & \text{if} \quad \cos\frac{\hat{X}}{2} > \frac{\sqrt{2}}{2} \quad \text{and} \quad \arcsin\frac{x\cdot\sin\hat{N}}{n} < 0 \\ \arcsin\frac{x\cdot\sin\hat{N}}{n} & \text{else} \end{cases} \quad \text{(A.6}$$

$$\hat{O} = \begin{cases} 180 - \arcsin\frac{o\cdot\sin\hat{N}}{n} & \text{if} \quad \cos\frac{\hat{O}}{2} > \frac{\sqrt{2}}{2} \quad \text{and} \quad \arcsin\frac{o\cdot\sin\hat{N}}{n} \geq 0 \\ -180 - \arcsin\frac{o\cdot\sin\hat{N}}{n} & \text{if} \quad \cos\frac{\hat{O}}{2} > \frac{\sqrt{2}}{2} \quad \text{and} \quad \arcsin\frac{o\cdot\sin\hat{N}}{n} < 0 \\ \arcsin\frac{o\cdot\sin\hat{N}}{n} & \text{else} \end{cases} \quad \text{(A.7}$$

A.2. Transformation of an error ellipse into a variance-covariance matrix

Our starting frame F_s is the plane centered in X and perpendicular in X to (CX). It has for x-axis the East direction and for y-axis the North direction. Errors on positions in astronomy are almost always given in the form of an ellipse of major axis a, minor axis b and of angle θ between the North pole and the major axis part oriented in the East direction. So $\theta \in [0, 180[$. We can transform the ellipse into a variance-covariance matrix in the frame F_s. The frame in which the ellipse is in its canonical form is obtain by rotating the frame F_s of an angle $\beta = 90 - \theta$. Applying a rotation of angle $-\beta$ on the canonical variance-covariance matrix give us the expression of the variance-covariance matrix in F_s:

$$V_{F_s} = \begin{pmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \begin{pmatrix} \cos\beta & -\sin\beta \\ \sin\beta & \cos\beta \end{pmatrix} \quad \text{(A.8)}$$

It leads to:

$$\sigma_x^2 = a\sin^2\theta + b\cos^2\theta \quad \text{(A.9)}$$

$$\sigma_y^2 = a\cos^2\theta + b\sin^2\theta \quad \text{(A.10)}$$

$$\rho\sigma_x\sigma_y = \cos\theta\sin\theta(a-b) \quad \text{(A.11)}$$

A.3. Variance-covariance matrices in our new frame

As mention in paragraph 3.1, our new frame F_n is centered in X, has for x-axis the direction of the perpendicular in X of (CO) in the plane XCO oriented in the East part and for y-axis the perpendicular in X of the plane XCO oriented in the North part. Given the results of the previous section, the new frame is obtain from the starting one (F_s) by a rotation of angle $\Omega = 90 - \hat{X}$ if $\hat{X}$ is positive and $\Omega = -90 - \hat{X}$ if $\hat{X}$ is negative. The variance-covariance matrix of the errors on positions can be expressed in the new frame F_n by a matrix V_n:

$$V_n = \begin{pmatrix} \cos\Omega & \sin\Omega \\ -\sin\Omega & \cos\Omega \end{pmatrix} \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y 0 \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \begin{pmatrix} \cos\Omega & -\sin\Omega \\ \sin\Omega & \cos\Omega \end{pmatrix} \quad \text{(A.12)}$$

Leading to:

$$\sigma_{x_n}^2 = \sigma_x^2\cos^2\Omega + \sigma_y^2\sin^2\Omega + 2\cos\Omega\sin\Omega\rho\sigma_x\sigma_y \quad \text{(A.13)}$$

$$\sigma_{y_n}^2 = \sigma_x^2\sin^2\Omega + \sigma_y^2\cos^2\Omega - 2\cos\Omega\sin\Omega\rho\sigma_x\sigma_y \quad \text{(A.14)}$$

$$\rho_n\sigma_{x_n}\sigma_{y_n} = \cos\Omega\sin\Omega(\sigma_y^2 - \sigma_x^2) + (\cos^2\Omega - \sin^2\Omega)\rho\sigma_x\sigma_y \quad \text{(A.15)}$$

A.4. Convolution product of two independent 2D Gaussians

The errors on the X-ray and the optical source are 2 Gaussian random variables Z_X and Z_o with variance-covariance matrices:

$$V_{Z_X} = \begin{pmatrix} \sigma_{x_X}^2 & \rho_X\sigma_{x_X}\sigma_{y_X} \\ \rho_X\sigma_{x_X}\sigma_{y_X} & \sigma_{y_X}^2 \end{pmatrix}, V_{Z_o} = \begin{pmatrix} \sigma_{x_o}^2 & \rho_o\sigma_{x_o}\sigma_{y_o} \\ \rho_o\sigma_{x_o}\sigma_{y_o} & \sigma_{y_o}^2 \end{pmatrix} \quad \text{(A.16)}$$

The random variable $Z = Z_x + Z_o$ is define by the convolution product of Z_X by Z_o. We know that the sum of 2 Gaussians is a Gaussian with a variance-covariance matrix:

$$V_Z = \begin{pmatrix} \sigma_{x_X+x_o}^2 & \rho\sigma_{x_X+x_o}\sigma_{y_X+y_o} \\ \rho\sigma_{x_X+x_o}\sigma_{y_X+y_o} & \sigma_{y_X+y_o}^2 \end{pmatrix} \quad \text{(A.17)}$$

As Z_X and Z_o are independent, x_X and x_o are independent. Idem for y_X and y_o. We thus have $Cov(x_X, x_o) = Cov(y_X, y_o) = 0$. The variance formula $Var(a + b) = Var(a) + Var(b) + 2Cov(a, b)$ leads to:

$$\sigma_{x_X+x_o}^2 = \sigma_{x_X}^2 + \sigma_{x_X}^2 \quad \text{(A.18)}$$

$$\sigma_{y_X+y_o}^2 = \sigma_{y_X}^2 + \sigma_{y_X}^2 \quad \text{(A.19)}$$

The covariance is given by $\rho\sigma_{x_X+x_o}\sigma_{y_X+y_o} = E\{(x_X + x_o)(y_X + y_o)\} - E\{(x_X+x_o)\}E\{(y_X+y_o)\}$. But as our four 1D Gaussian x_X, x_o, y_X and y_o are centered, their mean is null and thus $E\{(x_X+x_o)\} = E\{x_X\}+E\{x_o\} = 0$ and $E\{(y_X+y_o)\} = E\{y_X\}+E\{y_o\} = 0$. It leads to $\rho\sigma_{x_X+x_o}\sigma_{y_X+y_o} = E\{(x_Xy_X)\} + E\{(y_Xy_o)\} + E\{(x_Xy_o)\} + E\{(y_Xx_o)\}$ which is a sum of covariance since our four distributions are centered. We have already mentioned that the covariance of x_X and x_o is null. Idem for y_X and y_o. We finally obtain:

$$\rho\sigma_{x_X+x_o}\sigma_{y_X+y_o} = \rho_X\sigma_{x_X}\sigma_{y_X} + \rho_o\sigma_{x_o}\sigma_{y_o} \quad \text{(A.20)}$$

Appendix B: Some details on the Likelihood ratio: For SSC members information (will not be published)

We consider the case where an X-ray source has only one candidate and we formulate two hypotheses:

H_{cp}: the candidate is the counterpart, it is not random
H_{spur}: the candidate is a random source belonging to a Poissonian distribution of sources of density λ

The union of these two hypotheses gives the set of all possibilities: $P(H_{cp}) + P(H_{spur}) = 1$, .i.e $H_{cp} = \bar{H}_{spur}$. The probability of finding a candidate at a distance r of the X-ray source can thus be written:

$$P(r) = P(r|H_{cp}) \cdot P(H_{cp}) + P(r|H_{spur}) \cdot P(H_{spur}) \quad \text{(B.1)}$$

From conditional probabilities:

$$P(H_{cp}|r) = \frac{P(H_{cp} \cap r)}{P(r)} \quad \text{(B.2)}$$

$$P(r|H_{cp}) = \frac{P(H_{cp} \cap r)}{P(H_{cp})} \quad \text{(B.3)}$$

We deduce:

$$P(H_{cp}|r) = \frac{P(r|H_{cp}) \cdot P(H_{cp})}{P(r)} \quad \text{(B.4)}$$

Which can be written (formula (B.1)):

$$P(H_{cp}|r) = \frac{P(r|H_{cp}) \cdot P(H_{cp})}{P(r|H_{cp}) \cdot P(H_{cp}) + P(r|H_{spur}) \cdot P(H_{spur})} \quad \text{(B.5)}$$

We transform that equation in:

$$P(H_{cp}|r) = \frac{1}{1 + \frac{1}{\frac{P(H_{cp})}{P(H_{spur})} LR}} \quad \text{(B.6)}$$

To exhibit the likelihood ratio:

$$LR(r) = \frac{P(r|H_{cp})}{P(r|H_{spur})} \quad \text{(B.7)}$$

The remaining term $\frac{P(H_{cp})}{P(H_{spur})}$ has to be estimated to compute $P(H_{cp}|r)$, the probability that the candidate is a real counterpart. Considering the estimate of the number of real counterparts divided by the number of spurious associations, it leads to the $\theta/(1-\theta)$ factor of De Ruiter et al. (1977).

If we add a prior knowledge on the magnitude, the formula (B.6) is unchanged except that it now expresses $P(H_{cp}|r \cap m)$ and the LR becomes:

$$LR(r \cap m) = \frac{P(r \cap m|H_{cp})}{P(r \cap m|H_{spur})} = \frac{P(r|H_{cp})P(m|H_{cp})}{P(r|H_{spur})P(m|H_{spur})} \quad \text{(B.8)}$$

The term $P(r \cap m|H_{spur}) = P(r|H_{spur})P(m|H_{cp})$ expresses the probability of having a spurious source of magnitude m. It is the Poissonian local distribution of sources of magnitude m. The term $P(r|H_{cp})P(m|H_{cp})$ is the classical 2D Gaussian distribution times the probability of having, among all real counterparts, a counterpart of magnitude m. It can be seen as the distribution of counterpart sources according to their magnitudes and correspond to the $q(m)$ factor in Brusa et al. (2007a) and of $q(m, c)$ in Sutherland & Saunders (1992) if we ignore object types.



Annexe I

Demandes de temps télescope

L'utilisation de l'ACP sur les corrélations croisées entre le GSC2.2.1, le 2MASS et le 2XMMi d'une part et le GSC2.2, le 2MASS et le ROSAT All Sky Survey d'autre part a permis de sélectionner des candidats étoiles dans le cadre de recherches effectuées par Patrick Guillout et ses associés sur les *moving groups* d'étoiles jeunes. Cette sélection a été utilisée pour deux demandes de temps qui ont été acceptées : l'une porte sur l'Observatoire de Haute Provence (OHP), l'autre sur le Calar Alto en Espagne. Une partie des sources ont été observées sur le Calar Alto. Au moins sept nouvelles étoiles jeunes ont été découverte.

1 Demande de temps sur le Calar Alto

Time Allocation Committee for
Centro Astronómico Hispano Alemán (CAHA)
C/ Jesús Durbán Remón
E-04004 Almería / Spain

APPLICATION FOR OBSERVING TIME

from [ ] Germany [X] Spain [ ] other

Application No. H09-2.2-038
Observing period H2009
Received 15.03.09 18:23
Service

[X] Autumn period beginning of July - end of December, acceptance till March 15.

[ ] Spring period beginning of January - end of June, acceptance till September 15.

1. Telescope: 2.2-m [X] 3.5-m [ ]

2.1 Applicant: Prof. David **Montes** (Name); Universidad Complutense de Madrid (Institute)
Fac. Físicas, Dpt. Astrofísica, UCM (street); E-28040 Madrid (ZIP code - city)
SPAIN (country); dmg@astrax.fis.ucm.es (e-mail)

2.2 Collaborators: P. Guillout, F.X. Pineaut / A. Klutsch (name(s)); OA Strasbourg (OAS) / UCM (institute(s))
A. Frasca, E. Marilli / J. López-Santiago / R. Freire Ferrero, N. Grosso (name(s)); INAF-OA Catania / UCM / OAS (institute(s))

2.3 Observers: D. Montes, A. Klutsch (name); J. López-Santiago (name)

CAHA points out that by specifying the names under item 2.3 it is obligatory to also send out these observers to Calar Alto. Correspondence on the rating of this application will be sent to the applicant (P.I.) as quoted under 2.1 above.

3. Observing programme and method: Category: [E]

Title : **Searching for co-moving young stars around post-T Tauri candidates**

Abstract : Once mixed in the ambient galactic plane stellar population, young stars are virtually indiscernible because neither their global photometric properties nor the presence of nearby gas can help to disentangle them from older ones. Nevertheless, couples of very young field stars sharing the same space motion have recently been detected in various regions of the sky. Very young nearby stars in the field are of great importance to better understand the recent local star formation history and to give new insight into the process of star formation outside standard formation regions. Moreover, they offer an unique opportunity to study the evolution of circumstellar disks, planet formation and migration process. Finding such a population, for the first time in the northern hemisphere, is the main goal of this proposal.

4. Instrument: FOCES Method: Spectroscopy

5. Brightness range of objects to be observed: from 10 to 12 V-mag

6. Number of nights:

applied for			already awarded	still needed
3			none	none
no restriction	grey	dark		

7. Optimum date range for the observations: ..... 01.08.09 – 31.10.09
Usable range in local sidereal time LST: ..... 0:00h – 2:00h

8a. Description of the observing programme

Astrophysical context

It is known from statistical studies that star formation derived from optically visible clusters accounts for about 50 % of the total galactic one, the remainder occurring in OB associations. Although at the end of their parent cloud collapse newly-formed stars lye in the same region of space, many events tend to carry them away from each other. Finally, once mixed in the ambient galactic plane stellar population, young stars are virtually indiscernible because neither their global photometric properties nor the presence of nearby gas can help to disentangle them from older stars.

During the last decade, a fascinating picture of the recent star formation history in the solar neighbourhood has emerged. About fifty Myr ago molecular clouds were forming stars near the present position of the Sun. Forty Myr ago the first generation of massive stars exploded as supernovae triggering star formation in an expanding ring-like structure (i.e. the Gould Belt) and modeled the gas cavity in the solar vicinity (i.e. the local bubble). Although the present rim coincides with most of the nearby OB associations, the distribution of young stars as outlined by soft X-ray surveys [4] indicates that stellar formation is not only active along the Belt but also 100 pc or so inward, over a significant, yet poorly constrained, radial extent [5].

Although the massive-star content has been extensively studied, less is known about young low mass stars in the solar neighbourhood. Nevertheless, isolated very young stars have recently been detected. Although TW Hya displays all characteristics of T Tauri stars (TTS) – strong Hα emission and lithium absorption – it has earned notoriety mostly on the basis of the absence of dark clouds in its vicinity. The origin of TW Hya has long been a mystery and different explanations for the presence of TTS far away from molecular clouds have been proposed (i.e. run-away TTS [12]; local formation from small cloud [2]). Searches around TW Hya revealed other TTS [1, 3] in the same region of the sky sharing the same space motion. At a distance of only $\sim$ 50 pc and an age $\sim$ 12 Myr, TW Hya is now recognised as one of the closest known regions of recent star formation, i.e. the TW Hya association.

Until late 1990 the Hyades (600 Myr) and UMa (300 Myr) clusters were the only coeval, co-moving concentrations of young stars identified within 60 pc of the Sun [15]. During the last ten years, it has become increasingly evident that a considerable number of very young stars is present in the solar vicinity, which are unrelated to prominent star forming regions [9, 10]. Since the TW Hya association discovery, eight more young nearby associations have been identified so far [16], but more are expected [13, 14].

Another issue regarding young stars is known as the post-T Tauri star (PTTS) problem. Soderblom et al. [11] discussed possible explanations of the PTTS problem and, in particular, suggested that numerous PTTS may exist in isolated environments but that they are difficult to be recognized because they mostly lack the obvious optical signatures of their T Tauri progenitors.

Previous work

Selecting an appropriate sample of targets is the major difficulty in such studies. To overcome this problem, we used an X-ray biased approach. Guillout and al. [6] recently observed a sample of about 800 chromospherically active stars in the solar vicinity. These stars are the Tycho optical counterpart of ROSAT All-Sky Survey (RASS) X-ray sources in the northern hemisphere. High resolution optical spectroscopy was performed in the Hα and Li spectral regions to determine the radial and rotational velocities along with stellar parameters. The age was empirically derived from the lithium abundance. Based on their lithium and Hα properties, some stars were then classified as possible PMS stars (see Fig. 1) and their age estimated to be $\sim$ 20 – 50 Myr. Among these promising PTTS candidates we found that 4 of them display exactly the same space motion (see Fig. 2) and are located within a few degree from each others on the celestial sphere [7]. In particular, 2 of these candidates are projected just 4 arcmin apart on the sky. Even interestingly, Makarov et al. [8] have identified a common proper motion companion of a nearby young visual binary in the same area. Since all are found within 100 pc of the Sun, they have properties rather similar to the members of the TW Hydra association, although slightly older and located in the northern hemisphere.

Immediate aim

Finding such a population is the main goal of this proposal. We propose to search for co-moving young stars around our 4 post-T Tauri candidates. We have thus selected optical counterparts of XMM / RASS X-ray sources cross identified with late-type stars using sophisticated multivariate analysis methods allowing to disentangle the stellar population from the extragalatic component (galaxies and quasars) also emitting in X-ray (see Fig. 3). Our targets are spreaded over an area covering 3 000 deg^2 around the barycenter of the PTTS positions (see Fig. 4). Lithium-rich and chromospherically active stars that may be identified thanks to high resolution spectroscopy could reveal the pending members of a yet unknown young stream in the northern hemisphere. Very young nearby stars in the field are of great importance to better understand the recent local star formation history and to give new insight into the process of star formation outside standard formation regions. Moreover, PTTS offer a unique opportunity to study the evolution of circumstellar disks (in particular the transition phase of disk dispersal) as well as planet formation and migration processes.

Layout of observations

To study spectroscopic properties of these stellar soft X-ray sources, FOCES high resolution optical *echelle* spectra will be taken. The spectral coverage includes the most significant chromospheric activity indicators from Ca II H&K to Ca II IRT as well as the lithium line. In additional, radial velocity standards stars are also needed to apply the cross-correlation technique to obtain reliable radial velocities.

8b. Figures and tables

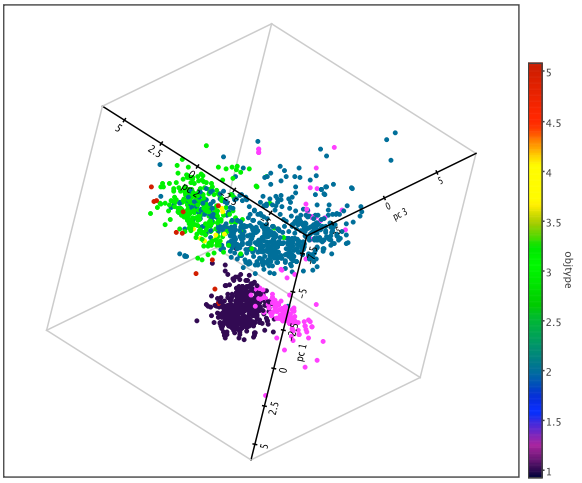


Figure 3: Result of the principal component analysis classification of XMM X-ray sources formerly identified with quasars (green symbols), galaxies (blue symbols) and stars (black symbols). Multivariate analysis methods maximize the separation of the different classes of X-ray emitters allowing to select PTTS candidates (pink symbols) [7].

Figure 1: Example of optical spectra of TTS or PTTS candidates discovered far from any star forming region. The observed spectra are displayed with black line in both panels. Filled-in profile or Hα emission (upper panel) together with strong lithium absorption line (lower panel) are the signature of possibly PMS stars. The pink dashed line (upper panel) show the Hα wavelength while a lithium-poor template of the same spectral type and gravity is over-plotted with a red line (lower panel).

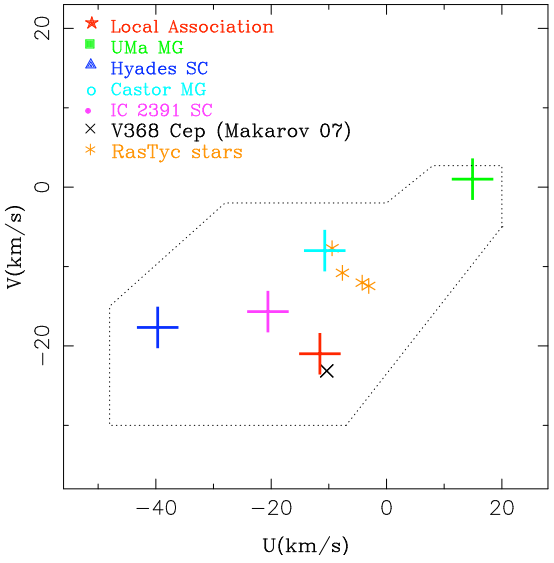


Figure 2: U-V diagram of our 4 PTTS candidates (asterisk symbols) plus that of Makarov et al. [8] (cross symbol). The locus of the young- and old-disk populations are also marked together with the mean position of the local association, Castor and UMa moving groups as well as IC2391 and Hyades superclusters [10]. Our candidates show a coherent galactic space motion [7].

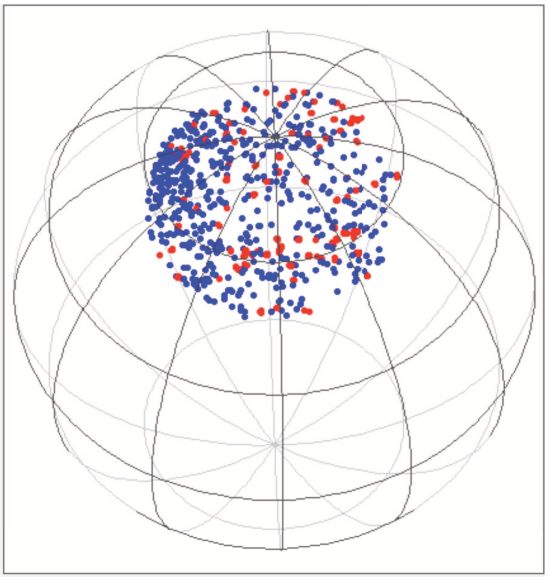


Figure 4: Sky position of XMM / RASS PTTS candidates (red and blue circles respectively) compiling our selection criteria (1: late-type star, 2: X-ray luminous: $L_X > 10^{30}$ erg s^{-1}) in a region 30^o wide encompassing couples of co-moving PTTS candidates discovered in the *RasTyc* sample.

9a. Objects to be observed

(Objects to be observed with high priority should be marked in last column.
Please limit the number of targets in this list to max. 30)

Designation	α (2000)	δ (2000)	magnitude in spectral range to be observed	priority
	h m s	° ′ ″	V SpT	
1RXS J160438.3+702212	$16^h\,04^m\,37\overset{s}{.}328$	$+70°\,22'\,14.12''$	11.21 K5	x
1RXS J163747.2+723937	$16^h\,37^m\,45\overset{s}{.}901$	$+72°\,39'\,42.15''$	11.58 G9	x
1RXS J165315.4+701554	$16^h\,53^m\,14\overset{s}{.}359$	$+70°\,15'\,59.79''$	10.75 G8	x
1RXS J194206.0+654655	$19^h\,42^m\,05\overset{s}{.}950$	$+65°\,46'\,49.68''$	10.80 G5	x
1RXS J195542.3+663207	$19^h\,55^m\,42\overset{s}{.}303$	$+66°\,32'\,05.01''$	11.58 K4	x
1RXS J195758.2+664253	$19^h\,57^m\,58\overset{s}{.}435$	$+66°\,42'\,51.30''$	10.14 G8	x
1RXS J200348.4+642542	$20^h\,03^m\,47\overset{s}{.}874$	$+64°\,25'\,47.85''$	10.97 K5	x
1RXS J210444.0+522326	$21^h\,04^m\,44\overset{s}{.}327$	$+52°\,23'\,25.61''$	12.11 K8	
1RXS J211024.8+704600	$21^h\,10^m\,24\overset{s}{.}496$	$+70°\,45'\,57.69''$	11.03 G8	x
XMM 179182	$21^h\,33^m\,02\overset{s}{.}685$	$+51°\,01'\,36.52''$	11.75 G8	
XMM 219527	$21^h\,43^m\,57\overset{s}{.}341$	$+82°\,07'\,39.08''$	11.53 M6	x
1RXS J214719.8+611618	$21^h\,47^m\,19\overset{s}{.}043$	$+61°\,16'\,17.50''$	11.59 G9	x
XMM 189583	$23^h\,36^m\,43\overset{s}{.}062$	$+62°\,04'\,33.43''$	11.52 G7	x
XMM 2282	$00^h\,19^m\,31\overset{s}{.}296$	$+81°\,22'\,11.68''$	11.63 K2	
1RXS J011319.6+585523	$01^h\,13^m\,19\overset{s}{.}589$	$+58°\,55'\,23.56''$	11.56 K5	x
XMM 19266	$02^h\,19^m\,23\overset{s}{.}865$	$+57°\,00'\,15.62''$	10.59 G2	x
1RXS J031026.1+582608	$03^h\,10^m\,26\overset{s}{.}415$	$+58°\,26'\,09.69''$	11.64 K5	x
1RXS J032802.3+642341	$03^h\,28^m\,01\overset{s}{.}981$	$+64°\,23'\,41.36''$	11.65 K2	x
1RXS J050248.7+735215	$05^h\,02^m\,49\overset{s}{.}235$	$+73°\,52'\,14.53''$	11.80 K4	
1RXS J053459.0+652146	$05^h\,34^m\,58\overset{s}{.}747$	$+65°\,21'\,43.98''$	11.54 K5	x

9b. Comments on the selection of target sample:

We propose to observe spectroscopically optical counterparts of XMM / RASS X-ray sources cross identified with late-type stars using sophisticated multivariate analysis methods (see Fig. 3). These candidates are spreaded over 3 000 deg^2 around a couple of co-moving PTTS candidates recently discovered far from any star forming regions (see Fig. 4). Lithium-rich and chromospherically active stars that may be identified thanks to high resolution spectroscopy could be the first pending of the TW Hya association (although slightly older) in the northern hemisphere.

For this observing run we have selected the brightest star of the sample ($V < 12$ mag) in order to be observed with this telescope/instrument configuration with exposure times shorter than 1 hour. For the fainter stars of the sample we are applying observing time in other large aperture telescopes and other more efficient intermediate resolution spectrographs.

10. Justification of the amount of observing time requested:

We propose to observe spectroscopically optical counterparts of XMM / RASS X-ray sources cross identified with late-type stars using sophisticated multivariate analysis methods (see Fig. 3). These candidates are spreaded over 3 000 deg^2 around a couple of co-moving PTTS candidates recently discovered far from any star forming regions (see Fig. 4). They have been selected according to their observability during the period ranging from August to October where FOCES will be available. We plan to use the Fibre Optic Cassegrain Echelle Spectrograph (FOCES) linked to the Cassegrain focus of the 2.2 m telescope, in unique fiber modus and with CCD (Site1d 2kX2k, 24 μm/pixel) and using the lower resolution mode (slit width 400μm). With this configuration, the resolution ($\lambda/\Delta\lambda$) is $\sim$ 28000 and the wavelength range covers from 3800 Å to 10000 Å in a total of 100 orders, which includes the chromospheric activity indicators Ca II H&K, He I D_3, Na I D_1, D_2, and the Balmer lines.

We would like to remark that, even if we do not detect PTTS that may pertain to the same yet undiscovered co-moving group, these spectroscopic observations will yield a very important contribution to fill the observational gap between the very young TTS and the older MS stars outside the classical star forming sites (i.e. open clusters and associations). These spectra will allow us to i) derive accurate spectral type, luminosity class and stellar parameters ($T_{\rm eff}$, $\log g$, $v \sin i$); ii) characterize their chromospheric activity (Hα line); iii) infer their age by deriving lithium abundance; iv) analyze their radial and rotational velocity distributions; v) single out spectroscopic and active binaries; vi) derive metallicity and abundances; vii) investigate kinematic properties.

The high efficiency of FOCES at the 2.2 m Calar Alto telescope will allow us to obtain a signal-to-noise (S/N) high enough with exposure times short enough in order to observe all our candidates ($V = 10 - 12$ mag) within three nights. In particular, we estimate that a S/N $\approx$ 50 is fully adequate for our purposes. To achieve this goal, FOCES at the Calar Alto telescope can provide us intermediate resolution spectra of a 11 mag late-type star in about 30 minutes (according to the exposure time calculator and our previous experience with this spectrograph). As our targets are all located in the same region of the sky, considering an average overhead time of about 5 – 10 minutes per exposure, we should be able to observe at least 8 stars in about ten hours (one night duration in August-October). So, based on our previous experience with FOCES, we could survey all our candidates in three observing nights that can be arranged in a "flexible" way, also taking into account the telescope schedule. In addition to the sample stars we need to observe radial velocity standard stars to obtain radial velocities using the cross-correlation technique.

11. Constraints for scheduling observations for this application:

Since it is not important that the observations of the program stars are collected in consecutive nights, we can accept the service mode during the period of August-October. Our only requirement is to acquire at least two spectra of radial velocity standard stars, as well as Flat lamp and Th-Ar spectra with the same instrumental setup, per each observing night. In this case we shall provide the complete source list.

12. Observational experience of observer(s) named under 2.3:
(at least one observer must have sufficient experience)

Some members of the research team have broad experience with the Calar Alto telescopes, the requested instrument, and many other telescopes and similar instruments located in other observatories (e.g. La Palma and La Silla). Our group has a wide experience in reducing *echelle* spectra (in particular with the requested instrument). In addition, we are familiar with different and powerful techniques for spectroscopic data analysis like the cross-correlation for the determination of radial and rotational velocities as well as the use of non-active standard stars and the spectral synthesis for the determination of physical parameters and for the spectral classification of our targets, both for single and multiple stars.

13. Calar Alto runs (preferably during the last 3 years)
and publications resulting from these

Telescope	instrument	date	nights	success rate	publications
2.2 m	FOCES	Jan 06	9	20%	[103, 104, 105, 106, 107] [108, 109, 110, 111, 112] [113, 114, 115]
2.2 m	FOCES	Aug 06	4	100%	[100, 101, 102]
2.2 m	FOCES	Dec 06	6	30%	[106, 107, 108] [111, 112, 113]
2.2 m	FOCES	Feb-May 07	6	70% (Service)	in analysis, [106, 107, 108] [111, 112, 113]
2.2 m	FOCES	Oct-Dec 08	5	80% (Service)	reduction phase
2.2 m	FOCES	09A	4		Not done yet

14. References for items 8 and 13:

[1] de la Reza, R., Torres, C. A. O., Quast, G., et al., 1989, ApJ, 343, 61

[2] Feigelson, E. D., 1996, ApJ, 468, 306

[3] Gregorio-Hetem, J., Lepine, J. R. D., Quast, et al., 1992, AJ, 103, 549

[4] Guillout, P., Sterzik, M. F., Schmitt, J. H. M. M. et al., 1998, A&A, 334, 540

[5] Guillout, P., Sterzik, M. F., Schmitt, J. H. M. M. et al., 1998, A&A, 337, 113

[6] Guillout, P., Klutsch, A., Frasca, A., et al., 2009, A&A, submitted

[7] Klutsch, A., 2008, PhD Thesis (Strasbourg observatory)

[8] Makarov, V. V., Zacharias, N., Hennessy, G. S., et al., 2007, ApJ, 668, L155

[9] Montes, D., López-Santiago, J., Fernández-Figueroa, et al., 2001, A&A, 379, 976

[10] Montes, D., López-Santiago, J., Gálvez, M. C., et al., 2001, MNRAS, 328, 45

[11] Soderblom, D. R., King, J. R., Siess, L., et al., 1998, ApJ, 498, 385

[12] Sterzik, M. F. and Durisen, R. H., 1995, A&A, 304, L9

[13] Torres, C. A. O., Quast, G. R., da Silva, L., et al., 2006, A&A, 460, 695

[14] Torres, C. A. O., Quast, G. R., Melo, C. F. H., Sterzik M.F., 2008, Handbook of Star Forming Regions, Vol. II, p. 757

[15] Zuckerman, B. and Song, I., 2004, ARA&A, 42, 685

[16] Zuckerman, B., Bessell, M. S., Song, I., et al., 2006, ApJ, 649, L115

[100] Biazzo, K., Frasca, A., Marilli, E., Covino, E., Alcalà, J. M., Çakirli, Ö., 2009a, "Photometric/spectroscopic analyses and magnetic activity in young late-type stars", in "The 15th Cool Stars, Stellar Systems and the Sun workshop", AIP, 1094, 636

[101] Biazzo, K., Frasca, A., Marilli, E., Covino, E., Alcalà, J. M., Çakirli, Ö., 2009b, "Magnetic activity in the young star SAO 51891" To appear in Mem. S.A.It., Proceedings of the "LII Congresso della Societa' Astronomica Italiana"

[102] Biazzo, K., Frasca, A., Marilli, E., Covino, E., Alcalà, J. M., Çakirli, Ö., Klutsch, A., Meyer, M. R.. 2009c "The young active star SAO51891 (V383 Lac)", A&A in press

[103] Maldonado J., Martínez-Arnáiz R. M., Montes D., Eiroa C., Montesinos B., Ribas I., Solano E. *Cool stars in the solar neighborhood. Preparatory activities for the Darwin mission*, 2006a, "Highlights of Spanish Astrophysics IV" Proc- VII Sc. Meeting Spanish Astro. Soc. (SEA), Springer, ISBN: 978-1-4020-5999-5, Figueras, F. et al. (Eds.)

[104] Maldonado J., Martínez-Arnáiz R. M., Montes D., Eiroa C., Montesinos B., Ribas I., Solano E. *Towards the characterization of the Darwin stars*, 2006b, in "TPF/Darwin Workshop: Star-Planet Interactions and Implications For Habitability"

[105] Maldonado J., 2007, DEA, UAM

[106] Maldonado J., Martínez-Arnáiz R. M., Montes D., Eiroa C., Montesinos B., Ribas I., Solano E., 2008a, *Exo-zodies and exo-EKB around Nearby Cool stars: The DUNES sample*, "Highlights of Spanish Astrophysics V" Proc- VIII Sc. Meeting Spanish Astro. Soc. (SEA)

[107] Maldonado J., Martínez-Arnáiz R. M., Montes D., Eiroa C., Montesinos B., Ribas I., Solano E., 2008b, *Exo-zodies and exo-EKB around Nearby Cool stars: The DUNES sample*, in "The 15th Cool Stars, Stellar Systems and the Sun workshop", AIP, 1094, 413

[108] Maldonado J., et al., 2009, A&A, in preparation

[109] Martínez-Arnáiz R. M., Maldonado J., Montes D., Eiroa C., Montesinos B., Ribas I., Solano E., 2006, *Characterizing the solar neighborhood. High Resolution spectroscopy of nearby FGK stars* , in "'The 14th Cool Stars, Stellar Systems and the Sun workshop", ASP Conf. Ser., vol. 384 (ISBN: 978-1-58381-331-7)

[110] Martínez-Arnáiz R. M., 2007, DEA, UCM

[111] Martínez-Arnáiz R. M., Maldonado J., Montes D., Eiroa C., Montesinos B., Ribas I., Solano E., 2008a, *High Resolution Spectroscopic Characterization of FGK Stars in the Solar Neighbourhood*, in "The 15th Cool Stars, Stellar Systems and the Sun workshop", AIP, 1094, 465

[112] Martínez-Arnáiz R. M., 2008b, *High resolution spectroscopic characterization of FGK stars in the Solar Neighbourhood*, "Highlights of Spanish Astrophysics V" Proc- VIII Sc. Meeting Spanish Astro. Soc. (SEA)

[113] Martínez-Arnáiz R. M., et al., 2009, A&A, in preparation

[114] Montes D, Martínez-Arnáiz R. M., Maldonado J., López-Santiago J., Crespo-Chacón I., Roa-Llamazares J. *Criteria for spectral classification of cool stars using high resolution spectra*, 2006a, in "IAU XXVIth General Assembly, Special Ses. 3: The Virtual Observatory in action: new science, new technology, and next generation facilities", SPS3, #30

[115] Montes D., Martínez-Arnáiz R. M., Maldonado J., Roa-Llamazares J., López-Santiago J., Crespo-Chacón I., Gálvez M. C., Solano E. *High resolution spectra of cool stars in the Virtual Observatory: Criteria for spectral classification*, 2006b, "Highlights of Spanish Astrophysics IV" Proc- VII Sc. Meeting Spanish Astro. Soc. (SEA), Springer, ISBN: 978-1-4020-5999-5, Figueras, F. et al. (Eds.)

15. CAHA does not cover costs of the observing run. It is the responsibility of the applicant(s) to raise the money for the travel to Calar Alto and expenses during the observing run.

    Consumables needed in larger quantities will be charged to the applicant(s).

    Furthermore the applicant(s) should consult our web page with the "Guidelines for Visiting Astronomers":

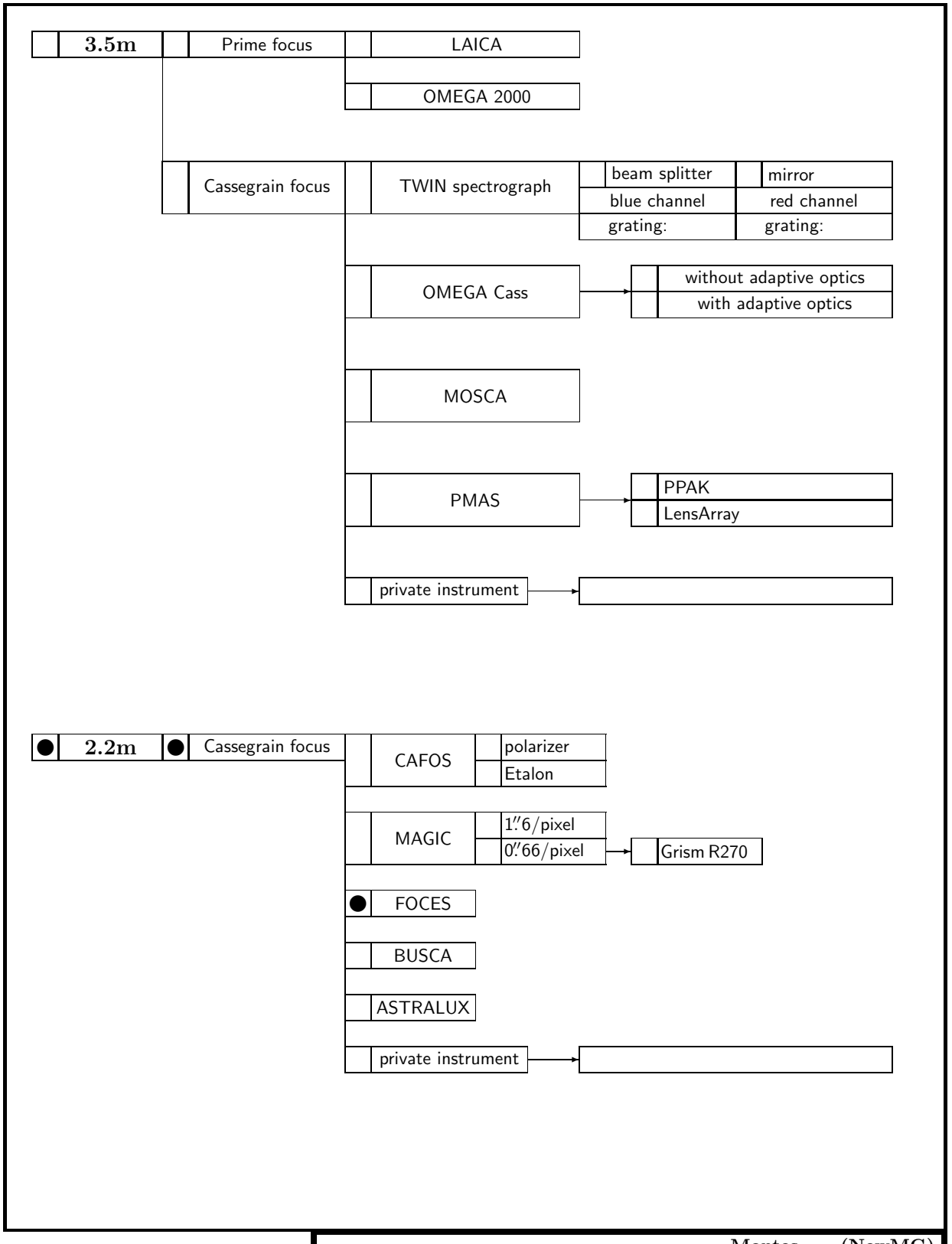
    http://www.caha.es/visast.html

16. Members of institutes of the Rat Deutscher Sternwarten (except Max Planck institutes) may apply for travel funding at the DFG with reference to this application and the letter granting observing time.

    Should observing time be granted is it planned to apply for DFG funds? yes [ ]
    no [X]

Tolerance limits for planned observations:

maximum seeing: 2.0″	**minimum transparency:** 40%	**maximum airmass:** 1.6
photometric conditions: no	**moon: max. phase / ∠ :** 1/20°	**min. / max. lag:** / nights



2 Demande de temps sur l'OHP

European Observing Proposal

Guillout

No code

Searching for co-moving young stars around post-T Tauri candidates in Cepheus

Abstract

Once mixed in the galactic plane, young stars are virtually indiscernible because neither their global photometric properties nor the presence of nearby gas can help to disentangle them from older ones. Nevertheless, very young field stars sharing the same space motion, i.e. moving groups, have been detected mainly in the southern hemisphere. Given their age and proximity, very young nearby stars in the field are of great importance to better understand the recent local star formation history and to give new insight into the process of star formation outside "standard" stellar nurseries. Moreover, they offer a unique opportunity to study debris/transition discs, planet formation and migration process. Stellar soft X-ray samples may virtually contain hundreds of such hidden young stars. We propose to find such a population around couples of co-moving post-T Tauri candidates recently discovered for the first time in the northern hemisphere, far from any star forming regions.

Science category : PNPS **Observing mode :** Service mode

Total requested nights: 2 nights

Telescope	Dark	First	Bright	Last	Instruments
OHP193	1.6		0.4		SOPHIE

Details, including preferred dates and periods to be avoided:

Since it is not important that the observations are collected in consecutive nights, we accept the service mode assuming that radial velocity standard stars, as well as calibration spectra will be acquired with the same instrumental setup. The stars selected for this program cover the whole range of right ascension but are located around the north equatorial pole. Thus, our targets can easily be inserted in other observation programs during the 2009B period.

Contact Author

Title	Dr	***Institute***	Observatoire Astronomique de Strasbourg
Name	Patrick Guillout	***Department***	
Email	guillout@astro.u-strasbg.fr	***Address***	11 rue de l'Université
Phone(first)	03 90 24 24 41	***Zipcode***	67000
Phone(second)		***City***	Strasbourg
Fax	03 90 24 24 32	***State***	
		Country	France

Pi

is the same as the contact author

Guillout

Searching for co-moving young stars aroun...

No code

Applicants

Name	Affiliation	Email	Country		Potential observer
Dr Patrick Guillout	Observatoire Astronomique de Strasbourg	guillout@astro.u-strasbg.fr	France	Pi	Yes
Dr Antonio Frasca	INAF-OA Catania	antonio.frasca@oact.inaf.it	Italy		
Prof David Montes	Universidad Complutense de Madrid	dmg@astrax.fis.ucm.es	Spain		
Dr Alexis Klutsch	Universidad Complutense de Madrid (Dept. Astrofisica, Fac. C.C. Fisicas)	klutsch@astrax.fis.ucm.es	Spain		
PhD François-Xavier Pineau	Observatoire Astronomique de Strasbourg	pineau@newb6.u-strasbg.fr	France		
Dr Ettore Marilli	INAF-OA Catania	ema@oact.inaf.it	Italy		
Dr Rubens Freire Ferrero	Observatoire Astronomique de Strasbourg	freire@astro.u-strasbg.fr	France		
Dr Nicolas Grosso	Observatoire Astronomique de Strasbourg	grosso@astro.u-strasbg.fr	France		
Dr Javier Lopez-Santiago	Universidad Complutense de Madrid	jls@astrax.fis.ucm.es	Spain		

Summary of observations

Number of targets	Telescope	Instrument	Exposure (mins.)
26 targets	OHP193	SOPHIE	1001

Is this a long term proposal: No

No PhD Students involved

Linked proposal submitted to this TAC: No

Linked proposal submitted to other TACs: Yes

A similar proposal has been submitted on the Italian time (PI A. Frasca) on the TNG telescope with the SARG spectrograph. This complementary proposal aims at observing the stars selected for this research project which were two faint to be observed within a reasonable amount of time on the OHP 1.93m telescope + SOPHIE.

Astrophysical context

It is known from statistical studies that open clusters account for only about 50% of the total galactic star formation, most of the remainder occurring in OB associations. Although newly-formed stars are usually confined to the locations of their parental clouds, many processes tend to disperse them. Finally, young stars are found scattered throughout the stellar population composed of stars born in clusters/associations which were disrupted hundreds millions or billions years ago. Once mixed in the galactic plane environment, young stars are virtually indiscernible because neither their global photometric properties nor the presence of nearby gas can help to disentangle them from older stars.

During the last decade a fascinating picture of the recent star formation history in the solar neighbourhood has emerged. About fifty Myr ago molecular clouds were forming stars near the present position of the Sun. Forty Myr ago the first generation of massive stars exploded as supernovae triggering star formation in an expanding ring-like structure (i.e. the *Gould Belt*) and modeled the gas cavity in the solar neighbourhood (i.e. the *local bubble*). Although the present rim coincides with most of the nearby OB associations, the distribution of young stars as outlined by soft X-ray surveys [1] indicates that stellar formation is not only active along the Belt but also 100 pc or so inward, over a significant, yet poorly constrained, radial extent [2]. Although the massive star content has been extensively studied, less is known about young low mass stars in the solar neighbourhood.

Until late 1990 the Hyades (600 Myr) and UMa (300 Myr) clusters were the only coeval, co-moving concentrations of young stars identified within 60 pc of the Sun [3]. However, during the last ten years, isolated very young stars have been detected and it has become increasingly evident that a considerable number of such stars, unrelated to prominent star forming regions [4][5], is present in the solar neighbourhood. Although TW Hya displays all characteristics of T Tauri stars (TTS) – strong Hα emission and lithium absorption – it has earned notoriety mostly on the basis of the absence of dark clouds in its vicinity. As an isolated TTS, the origin of TW Hya has long been a mystery for the astronomers and different scenario (i.e. *run-away TTS* [6] ; local formation from small cloud [7]) have been proposed as possible explanation for the presence of TTS far away from molecular clouds. However, searches around TW Hya revealed other TTS [8][9] in the same region of the sky sharing the same space motion. At a distance of only $\approx$ 50 pc and an age $\approx$ 12 Myr, TW Hya is now recognized as one of the closest known regions of recent star formation, i.e. the TW Hya association. Since its discovery, eight more young nearby associations have been identified so far [10], but more are expected [11][12].

Another issue regarding young stars is known as the post-T Tauri star (PTTS) problem. PTTS cover the evolutionary period between 10^6 to 10^7 years (i.e. between the T Tauri phase and zero age main sequence age). They are expected to be abundant in the vicinity of star formation regions, although studies which concentrate on regions of active star formation often fail to find the number of PTTS that are expected according to star formation and evolution theories. Soderblom et al. [13] discussed possible explanations of the PTTS problem and in particular suggested that numerous PTTS may exist in isolated environments but that they are difficult to be recognized because they mostly lack the obvious optical signatures of their T Tauri progenitors.

Given their age and proximity, very young nearby stars in the field are of great importance to better understand the recent local star formation history and to give new insight into the process of star formation outside standard formation regions (i.e. association and clusters). Moreover, PTTS offer a unique opportunity to study the evolution of circumstellar discs (in particular the transition phase of disc dispersal) as well as planet formation and migration processes. Stellar soft X-ray surveys may virtually contain hundreds of such hidden young stars. Finding such a population, for the first time in the northern hemisphere, is the main goal of this proposal.

Scientific aim and technical justification

Selecting an appropriate sample of targets is the major difficulty in such studies. To overcome this problem, we used an X-ray biased approach. Guillout and collaborators [14] recently observed a sample of about 800 chromospherically active stars in the solar vicinity. These stars are the Tycho optical counterpart of ROSAT All-Sky Survey (RASS) X-ray sources in the northern hemisphere. High resolution optical spectroscopy was performed in the Hα and lithium spectral regions to determine the radial and rotational velocities along with stellar physical parameters. The age was empirically derived from the lithium abundance (as it is commonly made for stars cooler than $B-V = 0.6$ mag). Based on their lithium and Hα properties, some stars were then classified as possible pre-main sequence (PMS) stars (Fig. 1) and their age estimated to be $\sim 20-50$ Myr. Among these promising PTTS candidates we found that 4 of them display exactly the same space motion

(Fig. 2) and are located within a few degree from each others on the celestial sphere. All may pertain to the same, yet undiscovered, co-moving group of stars [15]. Even interestingly, Makarov et al. [16] have identified a common proper motion companion of a nearby young visual binary in the same area. Since all are found within 150 pc of the Sun, they have properties rather similar to the members of the TW Hydra association, although slightly older and located in the northern hemisphere.

We propose to **search for companion co-moving young stars around the 4 post-T Tauri candidates** we have discovered in Cepheus far from any star forming regions. We have thus selected 26 optically bright counterparts of RASS / XMM X-ray sources cross identified with late-type stars using sophisticated multivariate analysis methods (Fig. 3 left panel) allowing to disentangle a-priori the stellar population from the extragalactic components (galaxies and quasars) also emitting in X-ray. Our targets are spreaded over an area covering 3 000 deg^2 around the barycenter of the PTTS positions (Fig. 3 right panel).

All our targets are brighter than $\sim 12.$ mag so that the OHP 1.93 m + SOPHIE spectrograph is the telescope of choice, allowing to reach the required resolution and wavelength coverage in the most efficient way. The High Efficiency mode allows to cover the 3800-7000 Å wavelength range, including activity (Ca II H&K, Hα) and age (Li) diagnostic lines, at a resolution of about 40 000, close to that of ELODIE and AURELIE spectrographs for which we have a large experience. In order to derive reliable equivalent widths and to enable the determination of good radial and rotational velocities, we require signal-to-noise ratio (S/N) of about 60 per resolution element in the red part of the spectrum. According to the telescope + spectro + CCD characteristics, one can achieve this S/N ratio with less than one hour exposure time per target. Such data will allow to :

(a) characterize their chromospheric activity (H_α and marginally the Ca II H&K lines)
(b) infer their age by deriving lithium abundance
(c) analyze their radial and rotational velocities
(d) single out spectroscopic and active binaries
(e) derive metalicity and abundances

We will use a synthetic approach to derive most of these parameters. Synthetic spectra will be computed at our spectral resolution and then adjusted to the observed spectra by repeatedly least square fitting in an automated manner. The code outputs several parameters among which radial velocity and apparent rotational velocity, effective temperature, gravity and metalicity. The age of the stars will be estimated by the equivalent width of the lithium absorption line as compared to galactic clusters members of the same spectral type. Once identified, the space velocity components (U,V and W) of companion PTTS candidates, will be computed from radial velocity and proper motions (derived from comparaisons of DDSS plate and current Sierra La Nave astrometry) and U-V-W kinematic diagrams will be analysed. **Lithium-rich and chromospherically active stars that may be identified thanks to high resolution spectroscopy could reveal, for the first time in the northern hemisphere, the pending members of a yet unknown young stream.**

References :

[1] Guillout, P., Sterzik, M. F., Schmitt, J. H. M. M. et al., 1998, A&A, 334, 540
[2] Guillout, P., Sterzik, M. F., Schmitt, J. H. M. M. et al., 1998, A&A, 337, 113
[3] Zuckerman, B. and Song, I., 2004, ARA&A, 42, 685
[4] Montes, D., López-Santiago, J., Fernández-Figueroa, et al., 2001, A&A, 379, 976
[5] Montes, D., López-Santiago, J., Gálvez, M. C., et al., 2001, MNRAS, 328, 45
[6] Sterzik, M. F. and Durisen, R. H., 1995, A&A, 304, L9
[7] Feigelson, E. D., 1996, ApJ, 468, 306
[8] de la Reza, R., Torres, C. A. O., Quast, G., et al., 1989, ApJ, 343, 61
[9] Gregorio-Hetem, J., Lepine, J. R. D., Quast, et al., 1992, AJ, 103, 549
[10] Zuckerman, B., Bessell, M. S., Song, I., et al., 2006, ApJ, 649, L115
[11] Torres, C. A. O., Quast, G. R., da Silva, L., et al., 2006, A&A, 460, 695
[12] Torres, C. A. O., Quast, G. R., Melo, C. F. H., 2008, Handbook of Star Forming Regions Vol.II. The Southern Sky, p757.
[13] Soderblom, D. R., King, J. R., Siess, L., et al., 1998, ApJ, 498, 385
[14] Guillout, P., Klutsch, A., Frasca, A., et al., 2009, A&A, submitted
[15] Guillout, P. et al., *Discovery of co-moving young stars in Cepheus*, 2009, in preparation
[16] Makarov, V. V., Zacharias, N., Hennessy, G. S., et al., 2007, ApJ, 668, L155

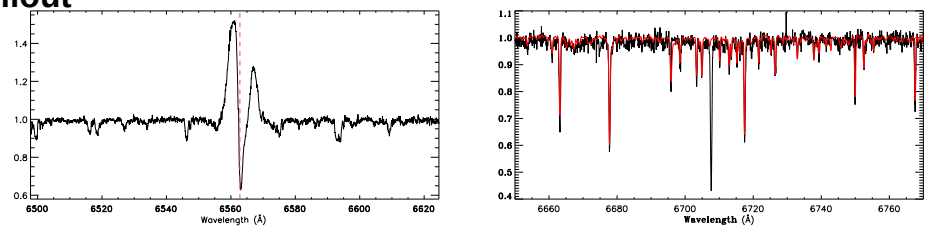


Figure 1. Example of optical spectra of TTS or PTTS candidates that we have discovered in the RASS / Tycho sample far from any star forming regions. The observed spectra are displayed with black line in both panels. Filled-in profile or Hα emission (left panel) together with strong lithium absorption line (right panel) are the signature of very young possibly PMS stars. The pink dashed line (left panel) show the Hα wavelength while a lithium-poor template of the same spectral type and gravity is over-plotted to the lithium spectra with a red line (right panel).

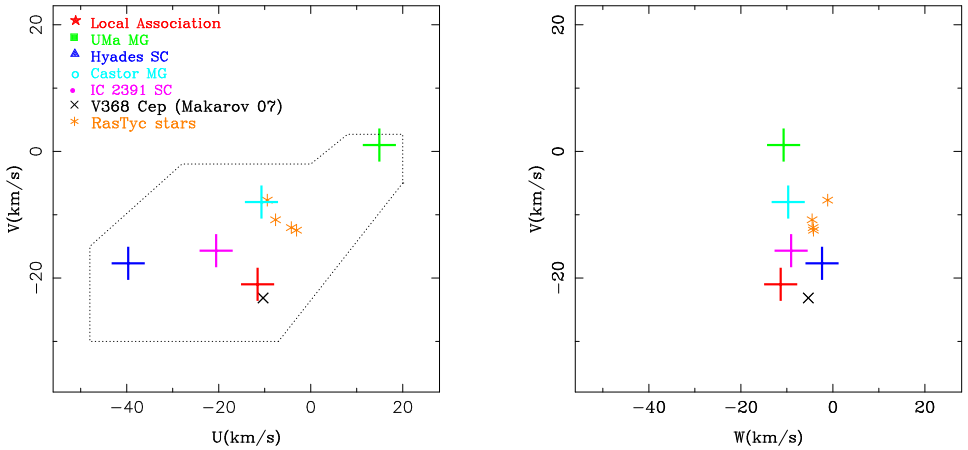


Figure 2. U-V and V-W kinematic diagrams showing the position of the 4 RasTyc PTTS candidates (brown asterisk symbols) we have discovered plus those ones of Makarov et al. [16] (black cross symbol). The locus of the young-disc and old-disc populations are also marked together with the mean position of the local association, Castor and Ursa Major (UMa) moving groups as well as IC2391 and Hyades superclusters. The 4 RasTyc PTTS candidates show a coherent space motion and may pertain to the same yet undiscovered co-moving group.

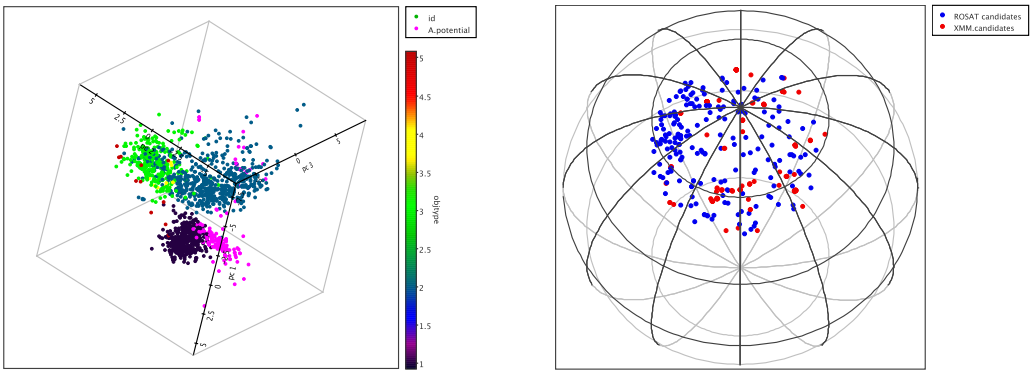


Figure 3. Left panel : Result of the principal component analysis classification of X-ray sources formerly identified with quasars (green symbols), galaxies (blue symbols) and stars (other symbols). Multivariate analysis methods maximize the separation of the different classes of X-ray emitters allowing to select a-priori, with a high probability, classes of objects of interest. We applied this method to both XMM and RASS X-ray sources to select a-priori young star candidates (pink symbols). **Right panel** : Sky position of XMM / RASS candidates (red and blue circles respectively) compiling our selection criteria (1: star - see explanation above, 2: late-type - B-V $>$ 0.6, 3: X-ray luminous - $L_X > 10^{29}$ erg s^{-1}, 4: within 170 pc of the Sun) in a region 30° wide encompassing couples of co-moving PTTS candidates that we have discovered in the RasTyc sample.

Justify the nights

The SOPHIE spectrograph in High Efficiency mode will allow us to obtain a signal-to-noise (S/N) high enough with exposure times short enough in order to observe all our candidates within a reasonable amount of time. In particular, at a resolving power R ~ 40000, a S/N~60 at 5500 angstroms is sufficient for our purposes. In fairly good sky and observing conditions, it can be achieved in 60 minutes for our faintest stars (12 mag). As our targets are all located in the same region of the sky, considering an average overhead time of about 5 minutes per exposure, we should be able to observe more than 10 stars per night (assuming ten hours as night duration). Summing up the individual exposure time (see observation details part) + 5 minutes overhead per target we could survey all our candidates in two nights.

Relevant previous Allocations: No

No additional remarks

Observation details

Telescope: OHP193

SOPHIE Instrument details

Observing mode: High-Efficiency mode (R=40000)

Read-out mode : Medium

No required scheduling constraints

No preferred scheduling constraints

Field	RA	Dec	Epoch	Exposure (min.)	Moon	Seeing	Water	S/N	Magnitude	Comments
J193003.3+663021	19:30:02.16	+66:30:24.9	J2000	4	bright	>0.9	average	60	8.9	B-V 0.52
J203903.6+640707	20:39:04.36	+64:07:07.0	J2000	7	bright	>0.9	average	60	9.5	B-V 0.44
J195758.2+664253	19:57:58.44	+66:42:51.2	J2000	12	bright	>0.9	average	60	10.1	B-V 0.75
J193141.7+641951	19:31:40.61	+64:19:51.8	J2000	13	bright	>0.9	average	60	10.2	B-V 0.64
J193026.5+580427	19:30:26.52	+58:04:33.8	J2000	17	bright	>0.9	average	60	10.5	B-V 0.84
J200348.4+642542	20:03:47.87	+64:25:47.9	J2000	26	bright	>0.9	average	60	11	B-V 1.15
J224917.6+522634	22:49:18.08	+52:26:36.2	J2000	20	bright	>0.9	average	60	10.7	B-V 0.66
J165315.4+701554	16:53:14.36	+70:15:59.7	J2000	21	bright	>0.9	average	60	10.7	B-V 0.73
J194206.0+654655	19:42:05.95	+65:46:49.7	J2000	22	bright	>0.9	average	60	10.8	B-V 0.67
J160438.3+702212	16:04:37.33	+70:22:14.1	J2000	33	bright	>0.9	average	60	11.2	B-V 1.12
J211024.8+704600	21:10:24.50	+70:45:57.7	J2000	27	bright	>0.9	average	60	11	B-V 0.73
J011319.6+585523	01:13:19.59	+58:55:23.6	J2000	45	dark	>0.9	average	60	11.6	B-V 1.16
J053459.0+652146	05:34:58.75	+65:21:44.0	J2000	45	dark	>0.9	average	60	11.6	B-V 1.14
J031026.1+582608	03:10:26.41	+58:26:09.7	J2000	49	dark	>0.9	average	60	11.7	B-V 1.17
J134050.7+792553	13:40:51.30	+79:25:55.0	J2000	44	dark	>0.9	average	60	11.5	B-V 0.95
J195542.3+663207	19:55:42.30	+66:32:05.0	J2000	49	dark	>0.9	average	60	11.7	B-V 1.04
J032802.3+642341	03:28:01.98	+64:23:41.4	J2000	49	dark	>0.9	average	60	11.7	B-V 0.94
J163747.2+723937	16:37:45.90	+72:39:42.2	J2000	46	dark	>0.9	average	60	11.6	B-V 0.79
J214719.8+611618	21:47:19.03	+61:16:17.5	J2000	46	dark	>0.9	average	60	11.6	B-V 0.77
J050248.7+735215	05:02:49.24	+73:52:14.5	J2000	59	dark	>0.9	average	60	11.9	B-V 1.04
J232346.4+620620	23:23:46.16	+62:06:19.4	J2000	54	dark	>0.9	average	60	11.8	B-V 0.91
J214909.0+591341	21:49:09.16	+59:13:34.3	J2000	64	dark	>0.9	average	60	12	B-V 1.05
J193124.1+672401	19:31:24.16	+67:23:57.4	J2000	55	dark	>0.9	average	60	11.8	B-V 0.77
J092654.8+812751	09:26:53.27	+81:27:47.1	J2000	69	dark	>0.9	average	60	12	B-V 1.02
J203549.9+594930	20:35:49.55	+59:49:38.0	J2000	61	dark	>0.9	average	60	11.9	B-V 0.82
J005300.8+682125	00:53:00.47	+68:21:22.6	J2000	64	dark	>0.9	average	60	11.9	B-V 0.85

Index

– A –
ACP ..... voir PCA
analyse en composantes principales ....voir ACP
arbre de décision 108, 110

– B –
boîte à moustaches 92
box plot 92

– C –
clustering 101

– D –
diagonalisation 88, 97, 98

– E –
errors-in-variables regression 92

– F –
feature space 96, 108, 110
fonction de poids 92
fonction noyau 97

– H –
hétéroscédasticité 95
Hopfield ..... voir réseaux de neurones
hyper-ellipse 88, 92

– I –
inliers 92

– K –
k-means 101, 106
kernel 97
  kernel PCA ..... voir KPCA
  kernel smoothing 91, 92, 111
  kernel trick 96, 97
kernel density classification 111
kernel trick 108–110
KPCA 96, 97

– L –
learning vector quantization ..... voir LVQ
LVQ 101

– M –
M-estimateurs 92, 93
machines à vecteurs de support .. voir SVM
maximum de vraisemblance 93
MCMC 111
mean shift 91, 92, 107
measurement error modeling 92
méthode de Newton 94
moindres carrés pondérés 92, 93

– O –
OC1 ..... 108, 110
orthogonal regression 92
outliers 90, 92

– P –
PCA 87
  classique 87–89
  kernel PCA ..... voir KPCA
  maximum likelihood PCA, MLPCA 92
  modifée 89, 92
  robuste 90
perceptrons multicouches .. voir réseaux de neurones
projection-pursuit 90, 91

– Q –
quartile 92

– R –
réseaux de neurones 100
  de Hopfield 101
  de Kohonen 101
    LVQ ..... voir LVQ
    SOM ..... voir SOM
    VQ ..... voir VQ
  perceptrons multicouches 100

– S –
self organizing maps ..... voir SOM
séparateurs à vaste marge ..... voir SVM
simulated annealing 91
SOM 101, 104
SVM 108, 110

– T –
total least square 92

– V –
vector quantization ..... voir VQ
VQ 101