Séparation de Composantes Multi-échelle pour Données Astrophysiques Multi-Canales: mGMCA

Florent SUREAU, Jérôme BOBIN, Jean-Luc STARCK

Laboratoire de Cosmologie et Statistique
CEA-DSM-IRFU-SEDI Bâtiment 709 CEA Saclay, F-91191 Gif-sur-Yvette Cedex, France
florent.sureau@cea.fr, jerome.bobin@cea.fr, jstarck@cea.fr

Résumé – L'analyse de données multi-canales motive le développement de nouvelles approches méthodologiques pour traiter de manière conjointe des données de grande taille, de résolution et de niveaux de bruit différant potentiellement. Dans ce cadre, nous nous intéressons dans ce travail à l'exploitation des données de la mission spatiale Planck, qui observe actuellement le ciel entier dans neuf canaux de résolution différant fortement (facteur 1 à 7). Les méthodes de séparation de composantes utilisées ne tiennent pas (ou mal) compte de ces différences de résolution, ce qui conduit à des résultats sous-optimaux dans l'identification de la matrice de mélange. De plus, résoudre conjointement le problème de déconvolution et de séparation de composantes est difficilement tractable pour Planck, avec des données de taille importante et des filtres de convolution à décroissance exponentielle. Nous proposons dans ce travail une approche en deux étapes : déconvolution des canaux en utilisant une approche linéaire ondelettes-vaguelettes, suivie d'une méthode de séparation de sources dans les différents niveaux de la décomposition vaguelette, précédemment développée (GMCA). Cette approche permet de conserver un modèle linéaire de mélange, et nous montrons dans des simulations simplifiées des données Planck que cette approche conduit à des niveaux de résidu inférieurs dans les cartes du fond diffus cosmologique et donc à une estimation plus précise de son spectre de puissance, comparée à l'approche GMCA classique.

Abstract – Multi-channel data analysis requires new approaches to jointly process large-scale data, which do not share the same resolution and noise levels. In this work, we deal with the problem of component separation for the full-sky Planck mission, a satellite which measures sky emissions at 9 wavelengths with highly varying resolutions (with a factor 7 in the FWHM between the first and last channel). Current component separation techniques do not tackle adequately this problem, which leads to sub-optimal identification of the mixing matrix, especially at high multipole ℓ. Besides, jointly solving the deconvolution and separation problem is numerically cumbersome, with large-scale data and convolution kernels decaying exponentially fast. We propose in this work a two-step approach: deconvolution of the channels using a linear wavelet-vaguelette decomposition, followed by the identification of the mixing matrix with a component separation approach based on the sparsity of the components in a tight frame, previously developed (GMCA). This linear approach preserves the linear mixture model and we show in simplified numerical simulations that this leads to lower foreground residuals in recovered CMB maps, and thus better estimates of the power spectrum, compared to a classical GMCA approach.

1 Introduction

Le développement récent d'imageurs multicanaux, en particulier en astrophysique, motive le développement de nouvelles approches méthodologiques pour traiter de manière conjointe des données de grande taille, de résolution et de niveaux de bruit différant potentiellement. Dans ce cadre, nous nous intéressons dans ce travail à l'exploitation des données de la mission spatiale Planck. Lancée en Mai 2009, cette mission de l'Agence Spatiale Européenne (ESA) fournit actuellement des données sur l'ensemble du ciel dans neuf bandes spectrales du domaine millimétrique et sub-millimétrique, avec une résolution et sensibilité jusqu'à présent inégalées. Planck vise notamment à mesurer les anisotropies en température et polarisation du fond diffus cosmologique (FDC) sur le ciel entier, cruciales pour mieux comprendre la structure et les premiers instants de notre univers. Pour atteindre ces objectifs, il est nécessaire de séparer les contributions des différents rayonnements observés (FDC et avant-plans d'origine galactique et extra-galactique), qui varient spatialement et en fonction de la bande spectrale observée. A cette fin, différentes approches de séparation de composantes ont déja été proposées et évaluées sur des données simulées, basées sur le $\ll$ Planck Sky Model $\gg$ [1]. Ces approches ont permis une estimation raisonnable de la carte du FDC et de son spectre de puissance. Cependant le niveau de résidu galactique dans ces cartes excédait le niveau de bruit, et aucune méthode ne pouvait être considéré comme conduisant au meilleur spectre de puissance pour tous les multipoles de cette carte. En particulier, ces approches ne tiennent pas en compte (ou de facon sous-optimale) des variations fortes de résolution dans les différents canaux d'observation, avec une largeur-à mi hauteur variant de 33 à 5 arcmin (voir Figure 1).

Ainsi parmi les huit approches proposées dans [1], six proposent de dégrader la résolution des canaux jusqu'à la résolution la plus faible avant d'effectuer la séparation de composantes, une seule effectue comme pré-traitement une déconvolution dans le domaine fréquentiel pour chaque canal (à l'aide du filtre inverse jusqu'à une fréquence fixée pour éviter que le

bruit n'explose), et la dernière résout simplement un problème de restauration dans le domaine fréquentiel (afin d'éviter des transformées coûteuses en temps de calcul). Si dégrader les données jusqu'à la résolution la plus faible permet d'estimer une matrice de mélange globale satisfaisante à partir des informations de basse fréquence présentes dans tous les canaux, cette approche conduit néanmoins à des estimations imparfaites des hautes fréquences du FDC pour Planck (qui ne peuvent être estimées qu'à partir d'un sous-ensemble de canaux, comme illustré dans la Figure 1).

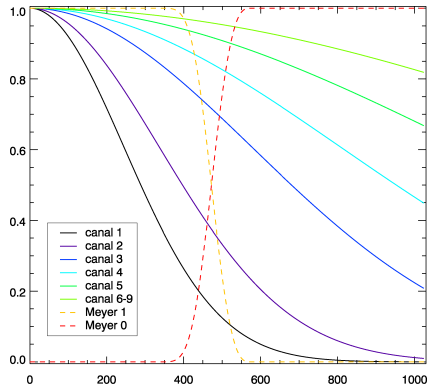

FIGURE 1 – Noyaux de convolution pour Planck et Filtres de Meyer utilisés (pointillés) pour les premiers multipoles.

L'objectif de ce travail est d'améliorer la séparation de composante en tenant mieux compte des noyaux de convolution et des informations de haute-fréquence dans les données Planck. Pour cela nous proposons une extension à une des méthodes de séparation de composantes proposée dans [1], GMCA (pour analyse en composantes morphologiques généralisées).
Cet article s'articule donc comme suit : dans la prochaine partie est présentée la méthode GMCA et son application à Planck, puis l'extension multi-échelle de GMCA (mGMCA) que nous proposons pour résoudre le problème de séparation de composantes dans le cas d'observations multicanales de résolution différente. L'intérêt de cette approche est ensuite illustré dans un cas synthétique simple, basé sur des simulations du Planck Sky Model (PSM) utilisé dans [1].

2 Séparation de sources pour Planck

Les méthodes proposées dans [1] pour Planck reposent pour la plupart sur le modèle bilinéaire suivant, qui factorise information spatiale et spectrale :

$$\left\{ o_i = b_i * \left(\sum_{c=1}^{N_c} a_{ic} s_c \right) + n_i \right\}_{i=1..N_f} \quad \text{ou } O = BAS + N \tag{1}$$

où $o_i \in \mathbb{R}^{N_e}$, $b_i \in \mathbb{R}^{N_e}$ et $n_i \in \mathbb{R}^{N_e}$ sont respectivement les données, la réponse impulsionnelle de l'instrument et un bruit gaussien pour le canal i (sur les N_f canaux observés),$N_e = N_s \times N_s \times 12$ avec N_s paramètre de résolution HEALPix (voir [3]) N_c est le nombre de composantes physiques à séparer, et la matrice de mélange globale $\{a_{ic}\}_{i=1..N_f, c=1..N_c} \in \mathbb{R}^{N_f \times N_c}$ contient le spectre des N_c différentes composantes. Enfin $s_c \in \mathbb{R}^{N_e}$ est une carte sur le ciel de la composante c. Il s'agit donc d'estimer conjointement la matrice A (ou une partie de cette matrice) et les composantes S à partir des données O, ce qui constitue un problème inverse mal posé de séparation de sources. Pour les données Planck, l'opérateur B est connu (noyau de convolution gaussien, voir Figure 1) et la convolution est réalisée dans le domaine des harmoniques sphériques. Les composantes physiques considérées sont typiquement le FDC, l'effet SZ, et les rayonnements synchrotron, de Brehmsstralung, et lié à la poussière dans la galaxie. Il est à noter que le modèle dans (1) ne permet pas de modéliser des composantes dont l'indice spectral varie spatialement, comportement attendu par exemple pour les composantes de poussière.

2.1 Approche GMCA globale

Parmi les méthodes de séparation de composantes, la méthode intitulée GMCA (pour analyse en composantes morphologiques généralisées, voir [2]) s'appuie sur la représentation parcimonieuse des composantes à extraire dans un dictionnaire $\mathcal{D}$ bien choisi (par exemple base d'ondelettes, trame d'ondelettes, dictionnaire de signaux ...) pour identifier une matrice de mélange $\hat{A}$. Cette approche consiste à résoudre le problème d'optimisation convexe suivant :

$$\begin{aligned} &\text{minimiser}_{\alpha,A} \sum_{c=1}^{N_c} \sum_{i=1}^{N_w} \lambda_{ic} |\alpha_{ic}| \\ &\text{tel que } \; ||\Sigma_N^{-1/2}(O - AS)||_2^2 \leq \tau \;\; \text{et} \;\; S = \mathcal{D}\alpha \end{aligned} \tag{2}$$

où α contient les coefficients $\{\alpha_{ic}\}_{i=1..N_w, c=1..N_c}$ de S dans $\mathcal{D}$ et Σ_N est la matrice de covariance du bruit. L'algorithme utilisé consiste à minimiser la version lagrangienne de ce problème, en estimant alternativement la matrice de mélange et les composantes. Ainsi, pour une itération n de GMCA, on résout successivement :

$$\hat{\alpha}^{(n)} = \underset{\alpha}{\text{argmin}} \, ||\Sigma_N^{-1/2}(O - \hat{A}^{(n-1)}\mathcal{D}\alpha)||_2^2 + ||\Lambda_n \alpha||_1 \tag{3}$$

$$\hat{A}^{(n)} = \underset{A}{\text{argmin}} \, ||\Sigma_N^{-1/2}(O - A\mathcal{D}\hat{\alpha}^{(n)})||_2^2 \tag{4}$$

avec Λ_n matrice diagonale contenant les pondérations λ_{ic}.

Dans le cas de la séparation de sources pour Planck, les histogrammes des coefficients dans une base d'ondelette des composantes physiques considérées présentent typiquement une decroissance exponentielle, ce qui justifie l'utilisation de l'hypothèse de parcimonie. D'autre part, l'approche GMCA suppose connues les réponses spectrales du FDC, du Brehmsstralung, de l'effet SZ, et suppose que le comportement spectral du synchrotron peut être décrit par un modèle paramétrique. La matrice A est donc partiellement connue. Les noyaux de convolution ne sont pris en compte qu'après la séparation de composante, en particulier pour corriger le spectre de puissance du FDC. Il est à noter que ceci viole le modèle linéaire global choisi en (1), puisque le modèle de mélange devient différent

pour chaque multipole considéré en harmoniques sphériques. Cette approche s'est cependant avérée compétitive pour estimer le spectre de puissance du FDC à bas multipoles où les noyaux pour chaque fréquence sont peu différents et ou l'énergie du FDC est concentrée (voir [1]).

2.2 Approche GMCA multi-échelle

Pour obtenir de meilleurs résultats à plus hauts multipoles, il est nécessaire de tenir compte des noyaux de convolution pendant la séparation de composante, c'est à dire réécrire (2) en remplaçant A par BA. Cependant résoudre un tel problème revient à résoudre un problème de déconvolution par itération de GMCA, ce qui devient rapidement intractable pour des applications avec des données de grande taille (environ 45 millions d'échantillons pour Planck), surtout en présence de noyaux de convolution décroissant exponentiellement en fréquence comme pour Planck. Une approche en deux étapes, déconvolution puis séparation de source, semble donc nécessaire dans de telles applications. Il devient cependant crucial de ne pas considérer chaque problème indépendemment, afin de ne pas détruire le modèle linéaire de mélange lors de la déconvolution.

Le problème de déconvolution est également typiquement mal posé, comme illustré dans la Figure (1) pour les premiers canaux de Planck, et doit donc être régularisé. Afin de tenir compte de la représentation parcimonieuse des composantes physiques considérées dans une base d'ondelettes, nous proposons ici une approche inspirée de la décomposition en ondelettes--vaguelettes [4], [5]. Dans les problèmes de déconvolution, les ondelettes doivent être à support compact dans l'espace fréquentiel, afin de maîtriser l'amplification du bruit lors de l'application du filtre inverse.

2.2.1 Décomposition en vaguelette

Considérons une trame ajustée basée sur des ondelettes de Meyer sur la sphère $\{\psi_{j,k}\}_{j\leq J,k\in\mathcal{K}}$ (indexé par la position k dans un ensemble dénombrable $\mathcal{K}$ et l'échelle j, en intégrant la fonction d'échelle dans cette famille pour $j = J$), et les coefficients d'un signal s_c dans cette trame à l'échelle j et à la position k :

$$\alpha_{j,k,c} = \langle s_c, \psi_{j,k}\rangle = \sum_{\ell=0}^{\ell_{max}} \sum_{m=-\ell}^{\ell} h_j(\ell) < s_c, Y_{\ell m} > Y_{\ell m}(k) \quad (5)$$

où $Y_{\ell m}$ sont les harmoniques sphériques usuels, et $h_j(\ell)$ est le filtre de Meyer à l'échelle j, comme représenté par exemple dans la Figure 1 .

La décomposition en vaguelettes$\{\varphi_{j,k}\}_{j\leq J,k\in\mathcal{K}}$ de l'observation o_i dans le cas d'un opérateur de convolution conduit à des coefficients $\beta_{j,k,i}$ qui s'écrivent :

$$\beta_{j,k,i} = \langle o_i, \varphi_{j,k}\rangle = \sum_{\ell=0}^{\ell_{max}} \sum_{m=-\ell}^{\ell} \frac{h_j(\ell)}{b_i(\ell)} < o_i, Y_{\ell m} > Y_{\ell m}(k) \quad (6)$$

où $b_i(\ell)$ indique les valeurs du noyau de convolution du canal i pour le multipôle ℓ. Dans le cas de noyaux de convolution décroissant exponentiellement en fréquence, Pensky a notamment montrée que pour des signaux de norme bornée dans un espace de Sobolev, l'estimateur linéaire construit avec des vaguelettes basées sur des ondelettes de type Meyer en ne conservant que les coefficients d'échelle pour un niveau de décomposition à déterminer est asymptotiquement optimal [5]. Ainsi, pour chaque canal dans Planck, cette approche conduit à ne conserver qu'une bande de basses fréquences, dépendant de la résolution dans chaque canal.

2.2.2 GMCA par échelle

L'étape suivante de l'approche GMCA multi-échelle consiste à estimer une matrice de mélange pour chaque échelle j de la décomposition en ondelettes de Meyer en utilisant GMCA, avec un nombre variable de canaux dans chaque bande :

$$\hat{\alpha_j}^{(n)} = \underset{\alpha}{\operatorname{argmin}} \, ||\Sigma_{N,j}^{-1/2}(\beta_j - \hat{A}_j^{(n-1)}\alpha)||_2^2 + ||\Lambda_{n,j}\alpha||_1 \quad (7)$$

$$\hat{A_j}^{(n)} = \underset{A}{\operatorname{argmin}} \, ||\Sigma_{N,j}^{-1/2}(\beta_j - A\hat{\alpha}_j^{(n)})||_2^2 \quad (8)$$

avec $\Lambda_{j,n}$ matrice diagonale contenant les pondérations pour la bande j, $\beta_j = \{\beta_{j,k,i}\}_{k\in\mathcal{K},i=1..N_f}$, $\alpha_j = \{\alpha_{j,k,i}\}_{k\in\mathcal{K},i=1..N_f}$, $\Sigma_{N,j}$ matrice de covariance inter-canale après déconvolution.

Cette approche permet donc également de conserver la validité du modèle linéaire dans (1). L'intérêt de cette approche en deux étapes est maintenant illustré dans un exemple simple dans la section suivante.

3 Méthodes et résultats

L'intérêt comparé de l'approche mGMCA et GMCA a été évalué dans des simulations dérivées du Planck Sky Model (comme utilisé dans [1]), dégradées à la résolution HEALPix $N_s = 512$ [3]. Dans ces données, cinq composantes sont simulées : rayonnements du FDC, du Brehmsstralung, du synchrotron, et deux rayonnements dûs à la poussière (effet thermique et effet de rotation). Avant analyse, les cartes originellement échantillonnées avec un paramètre $N_s = 2048$ ont été dégradées à la résolution $N_s = 512$ en multipliant les données avec un filtre passe bas de support dans les harmoniques sphériques $[0, 2N_s]$, afin d'éviter les effets de repliement de spectre et d'imprécisions liées à la transformée en harmonique sphérique de HEALPix. Les données ont été convoluées et du bruit blanc gaussien a été ajouté, selon les spécifications des instruments de Planck (voir [1]). Dans le premier cas étudié, non présenté ici (résultats similaires), un modèle de mélange purement linéaire a été estimé à partir d'une réalisation de chaque composante à chaque fréquence, afin de comparer GMCA et mGMCA dans le cas où le modèle dans (1) est valide. Dans le second cas étudié, l'émission thermique de la poussière, composante dominante dans les canaux de haute fréquence, n'a pas été linéarisée afin de se rapprocher du cas attendu pour Planck (avec un indice spectral variant spatialement). Dans ce cas, le modèle dans (1) n'est plus valide, et la matrice globale de mélange doit potentiellement identifier plus de composantes que de canaux disponibles. Dans les deux approches mGMCA et GMCA, la colonne du FDC et celle du Brehmsstralung étaient connues, et

les algorithmes ont été initialisés avec la même matrice basée sur un modèle paramétrique de la poussière. 40 itérations de GMCA ont été utilisés. Pour mGMCA, des ondelettes de Meyer représentées dans la Figure 1 ont été utilisées, et respectivement 7 (9) canaux ont été utilisés ensuite pour GMCA dans le premier (second) niveau de résolution. Les cartes obtenues pour le FDC ainsi que les spectres de puissance estimées sont présentées dans les figures 3,4,5,6. Ces figures illustrent un niveau de résidu galactique plus faible avec mGMCA dans le plan galactique, où le mélange est plus complexe, conduisant à une estimation plus fiable du spectre de puissance du FDC.

Nous projettons maintenant d'étudier le comportement de mGMCA dans des simulations plus réalistes, avec des cartes de résolution supérieure. Nous étudions également des critères pour choisir le pavage du domaine des harmoniques sphériques avec les filtres de Meyer afin d'améliorer la séparation de composante.

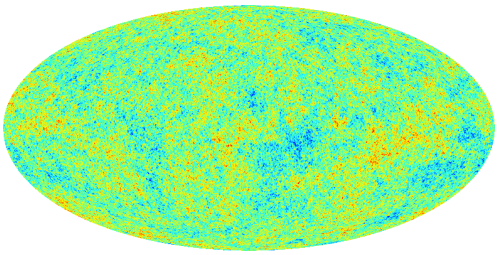

FIGURE 2 – Carte du FDC (référence des simulations).

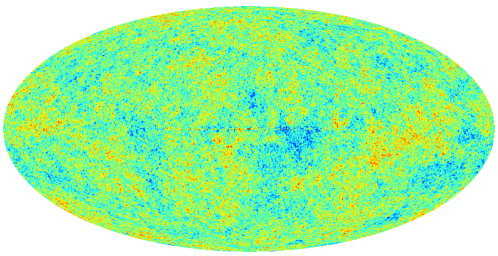

FIGURE 3 – Carte du FDC estimée avec GMCA.

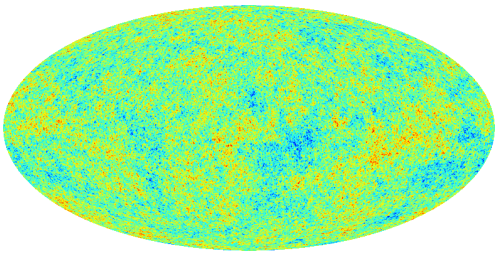

FIGURE 4 – Carte du FDC estimée avec mGMCA.

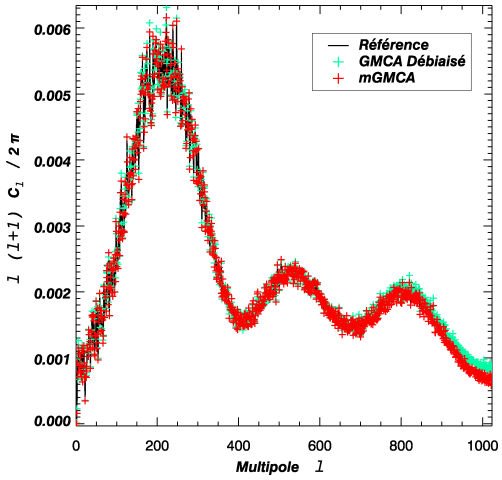

FIGURE 5 – Spectres de Puissance du FDC (référence, GMCA et mGMCA).

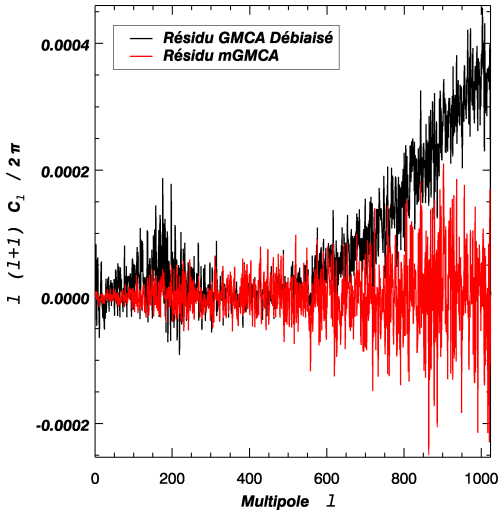

FIGURE 6 – Résidu des spectres de puissance.

Références

[1] S. Leach et al., A. & A., 491, 597-615, 2008.

[2] J. Bobin et al., IEEE Trans. Imag. Proc. vol. 16, 2662-2674

[3] Gòrski et al 2005, ApJ 622, 759-771

[4] D.L. Donoho, Appl. Comp. Harm. Anal. 1995, vol. 2, 101-126

[5] M. Pensky et B. Vidakovic 1999, Ann. Stat., vol. 27, 2033-2053