Approche bayésienne en séparation de sources. Applications en imagerie

Hichem Snoussi

ORSAY
n° d'ordre : 7314

UNIVERSITÉ PARIS–SUD

THÈSE

présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PARIS SUD
Spécialité : Automatique et Traitement du Signal

par

Hichem Snoussi

TITRE : APPROCHE BAYÉSIENNE EN SÉPARATION DE SOURCES. APPLICATIONS EN IMAGERIE.

Soutenue **le 29 septembre 2003** devant la commission d'examen :

MM.	Pierre-Olivier Amblard	*Rapporteur*
	Jacques Delabrouille	
	Bill Fitzgerald	
	Christian Jutten	*Président*
	Ali Mohammad-Djafari	*Directeur de thèse*
	Dinh Tuan Pham	*Rapporteur*

Pour Toi, certes le plus connaisseur,

À toi, notre guide et modèle,

À ma famille,

À toi, Eya,

REMERCIEMENTS

Je commence par remercier mon directeur de thèse, Ali Mohammad-Djafari, pour m'avoir donné l'occasion d'entrer par la grande porte dans le monde fascinant des problèmes inverses. Sa confiance et son soutien m'ont beaucoup aidé à accomplir ce travail.

Je remercie également tous mes professeurs et en particulier Gérard Attal, Bernard Picinbono et Guy Demoment pour la qualité de leurs cours qui ont véritablement façonné ma vision du traitement statistique du signal.

Pierre-Olivier Amblard et Dinh-Tuan Pham ont eu la lourde tâche d'examiner cette thèse. Qu'ils trouvent ici mes considérations les plus sincères. Je remercie Christian Jutten pour avoir accepté de présider le jury. Un grand merci également à Jacques Delabrouille qui m'a donné l'occasion de travailler sur une application de cosmologie observationnelle pendant ma thèse et qui a participé à mon jury.

Je remercie tous les gens que j'ai croisés au L2S et en particulier Mehdi (mon successeur au GPI), Gio, Jérôme, Bessem, Hana, Ilhem, Sahmoudi, Safia, Myriam, Mourad, Olivier, Thomas ainsi que tous les stagiaires et thésards du L2S. Je remercie aussi Guillaume Patanchon, mon collaborateur au Collège de France.

Je ne peux pas manquer l'occasion pour adresser un remerciement particulier à toute ma famille : mon père (Essnoussi), ma mère Wassila, mon grand frère Féthi, ma soeur Sarra (avec son mari Riadh et sa petite Farah) et mon petit frère Anouar. Leur amour, leur respect et leur estime me remplissent le coeur.

Un remerciement spécial à Badi pour son amitié sincère et hors pair.

Un remerciement chaleureux à celle qui a ensoleillé ma vie, à la plus belle rencontre de ma vie, à toi Eya.

Et à Toi, en premier et en dernier lieu, à qui je dois TOUT...

Résumé

Approche bayésienne en séparation de sources. Applications en imagerie.

Dans le premier chapitre, nous présentons le problème de séparation de sources. Nous montrons que ce problème peut être abordé par deux approches duales : comme un problème de reconstruction ou comme un problème de décomposition. Cette dualité des objectifs du problème de la séparation de sources implique une dualité dans les méthodes de sa résolution :

1. la séparation de sources commme un problème de reconstruction s'inscrit naturellement dans une approche bayésienne,
2. et comme un problème de décomposition, elle s'inscrit dans une approche informationnelle.

Dans le deuxième chapitre, nous exposons l'approche bayésienne en séparation de sources. Nous distinguons l'aspect théorique de l'aspect technique de cette approche. Sur le plan théorique, cette approche présente plusieurs avantages :

1. En introduisant la loi $p(\boldsymbol{A})$, nous pouvons prendre en compte toute information *a priori* sur ses éléments. Par ailleurs, ceci nous permet de dépasser les limites imposées par l'existence de $\boldsymbol{A}^{-1}$ (nombre de sources = nombre de capteurs). On peut aussi intégrer par rapport à cette matrice pour obtenir la loi marginale des sources.
2. En introduisant une loi *a priori* pour les hyperparamètres[1], on peut aussi s'affranchir de certaines difficultés liées à la dégénérescence de la vraisemblance lorsqu'il s'agit d'estimer ces hyperparamètres.
3. En introduisant des variables cachées, on peut enrichir la modélisation des sources.
4. On tient compte explicitement du bruit dans le modèle d'observation.

Afin de profiter des avantages de l'approche bayésienne, on doit effectuer des intégrations. Ceci n'est pas toujours possible à réaliser analytiquement. Le calcul bayésien offre ainsi des méthodes numériques basées sur l'échantillonnage.

Dans les chapitres suivants, nous considérons un mélange linéaire instantané bruité. Le point commun de ces chapitres est l'exploitation de la non stationnarité que ce soit dans le domaine temporel, spatial, fréquentiel ou temps fréquence. Les algorithmes proposés intègrent implicitement la reconstruction des sources contrairement aux méthodes estimant la matrice de mélange en ajustant les statistiques. Ce point a aussi son importance car une bonne estimation de la matrice de mélange n'implique pas forcément une bonne estimation des sources.

Dans le troisième chapitre, on considère des sources monovariées (1-D) qu'on modélise par des mélanges de gaussiennes. L'estimation des variances des gaussiennes provoque une dégénérescence de la vraisemblance d'où la pénalisation par des lois inverses gamma [Snoussi et Mohammad-Djafari, 2001]. En considérant les étiquettes des gaussiennes comme des variables cachées, nous obtenons un problème doublement caché : les sources sont des variables cachées pour l'estimation de la matrice de mélange et les étiquettes sont des variables cachées pour l'estimation des paramètres des distributions des sources. Nous étudions le cas des sources modélisées par des chaînes de Markov cachées (les étiquettes forment une chaîne de Markov) en implémentant l'algorithme EM exact. Nous avons proposé et implémenté des versions sous optimales

[1] On désigne par hyperparamètres les paramètres qui ne font pas partie de l'ensemble des paramètres d'intérêt.

pour accélérer l'EM [Snoussi et Mohammad-Djafari, 2002a]. La représentation hiérarchique des sources par l'introduction des variables cachées peut être interprétée comme une exploitation de la non stationnarité des variances (voir paragraphe I.2.3) avec une partition automatique de l'intervalle $\mathcal{I} = [1\ T]$ en K sous intervalles avec K le nombre des étiquettes vectorielles des gaussiennes. Le modèle de Markov pour les étiquettes peut être considéré comme une régularisation de cette classification.

Dans le quatrième chapitre, nous exploitons la non stationnarité des variances pour séparer des images mélangées. En effet, on rencontre souvent des images homogènes par morceaux et donc qui se prêtent bien à une modélisation par mélange de gaussiennes[2]. Cependant, la classification nécessite une régularisation qui tient compte de l'homogénéité spatiale des images. Cette régularisation peut être effectuée en incorporant un modèle de champ de Markov pour les étiquettes cachées. Nous présentons une implémentation du type MCMC (Monte Carlo par Chaînes de Markov) permettant d'estimer conjointement la matrice de mélange, les sources et leurs ségmentations. Les résultats sont testés sur des images synthétiques (champs cachés de Potts) et sur des images satellitaires [Snoussi et Mohammad-Djafari, 2002c, 2003].

Dans le cinquième chapitre, nous exploitons la non stationnarité fréquentielle. En effet, avec une approximation circulante, les coefficients de la transformée de Fourier d'un processus gaussien stationnaire sont décorrélés avec une variance (spectre) qui dépend de la fréquence. Donc en utilisant le maximum de vraisemblance, le critère devient une somme pondérée des divergences de Kullback-Leibler entre des matrices spectrales. Nous étudions le cas où les spectres des sources sont connues *a priori* et le cas où on les estime en découpant le domaine de Fourier en anneaux. La minimisation de ce critère est implémentée avec l'algorithme EM et accélérée autour de la solution avec un algorithme de gradient conjugué. Nous avons appliqué cette méthode pour séparer des composantes astrophysiques en l'implémentant dans le domaine de Fourier et dans le domaine des harmoniques sphériques [Snoussi *et al.*, 2001; Cardoso *et al.*, 2002; Patanchon *et al.*, 2003]. Ce travail fait l'objet d'une collaboration avec le laboratoire *IN2P3* du Collège de France.

Le sixième chapitre traîte en détail le problème de dégénérescence du maximum de vraisemblance. Nous avons généralisé des résultats obtenus dans le cas d'une modélisation par mélange de gaussiennes de signaux monovariables au cas de signaux multivariables. La dégénérescence est produite quand les matrices de covariance approchent des matrices singulières (frontière de singularité). L'élimination de cette dégénérescence est garantie par l'utilisation d'un *a priori* inverse wishart sur les matrices de covariance sans compliquer les équations de ré-estimation de l'algorithme EM. On montre que cette dégénérescence est aussi produite en séparation de sources quand les sources sont modélisées par un mélange de gaussiennes (ou en général par un modèle de Markov caché). La pénalisation par un *a priori* inverse Wishart élimine également cette dégénérescence [Snoussi et Mohammad-Djafari, 2001].

Le septième chapitre est consacré au problème de la sélection de la loi *a priori* dans un contexte bayésien. Nous présentons une approche originale [Snoussi et Mohammad-Djafari, 2002b] basée sur la théorie de la prédiction bayésienne [Zhu et Rohwer, 1995] en utilisant les outils de la géométrie de l'information [Amari et Nagaoka, 2000]. On montre l'importance du choix de la géométrie dans l'espace des distributions de probabilité. La règle de Bayes permet de définir la masse par la loi *a posteriori*. Une fois la géométrie et la masse fixées, on construit un critère variationnel dont la minimisation donne la loi *a priori* qu'on a notée δ-*a priori*. Avec les outils de la géométrie différentielle, on introduit la notion d'*a priori* projeté pour les familles paramétriques. Ce travail est appliqué au mélange de familles δ-plates comme le mélange de familles exponentielles (0-plates) et en séparation de sources.

Le dernier chapitre ouvre quelques perspectives comme la séparation des images en utilisant une ségmentation par ensembles de niveau. Au lieu de classifier les pixels en utilisant les étiquettes discrètes modélisées par un champ de Markov, on fait évoluer au cours des itérations un contour délimitant les régions homogènes. On va aussi revenir sur la logique des questions comme un espace dual de la logique des propositions

[2] C'est d'ailleurs le but d'un traitement avancé des images où on modélise l'image par un mélange de gaussiennes afin de la ségmenter.

en essayant de l'appliquer pour séparer et ségmenter simultanément des images mélangées dans le cadre de la théorie de l'information.

Bibliographie

[Amari et Nagaoka, 2000] S. Amari et H. Nagaoka. *Methods of Information Geometry*, volume 191 of Translations of Mathematical Monographs. AMS, OXFORD, University Press, 2000.

[Cardoso *et al.*, 2002] J. Cardoso, H. Snoussi, J. Delabrouille et G. Patanchon. Blind separation of noisy gaussian stationary sources. application to cosmic microwave background imaging. In *Eusipco*, Toulouse, septembre 2002.

[Patanchon *et al.*, 2003] G. Patanchon, H. Snoussi, J. Cardoso et J. Delabrouille. Component separation for cosmic microwave background data : a blind approach based on spectral diversity. In *PSIP*, Grenoble, janvier 2003.

[Snoussi et Mohammad-Djafari, 2001] H. Snoussi et A. Mohammad-Djafari. Penalized maximum likelihood for multivariate gaussian mixture. In R. L. Fry, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 36–46. MaxEnt Workshops, Amer. Inst. Physics, août 2001.

[Snoussi et Mohammad-Djafari, 2002a] H. Snoussi et A. Mohammad-Djafari. Bayesian unsupervised learning for source separation with mixture of gaussians prior. *To appear in Int. Journal of VLSI Signal Processing Systems*, 2002.

[Snoussi et Mohammad-Djafari, 2002b] H. Snoussi et A. Mohammad-Djafari. Information Geometry and Prior Selection. In C. Williams, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 307–327. MaxEnt Workshops, Amer. Inst. Physics, août 2002.

[Snoussi et Mohammad-Djafari, 2002c] H. Snoussi et A. Mohammad-Djafari. MCMC Joint Separation and Segmentation of Hidden Markov Fields. In *Neural Networks for Signal Processing XII*, pages 485–494. IEEE workshop, septembre 2002.

[Snoussi et Mohammad-Djafari, 2003] H. Snoussi et A. Mohammad-Djafari. Fast joint separation and segmentation of mixed images. *To appear in Journal of Electronic Imaging*, 2003.

[Snoussi *et al.*, 2001] H. Snoussi, G. Patanchon, J. Macías-Pérez, A. Mohammad-Djafari et J. Delabrouille. Bayesian blind component separation for cosmic microwave background observations. In R. L. Fry, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 125–140. MaxEnt Workshops, Amer. Inst. Physics, août 2001.

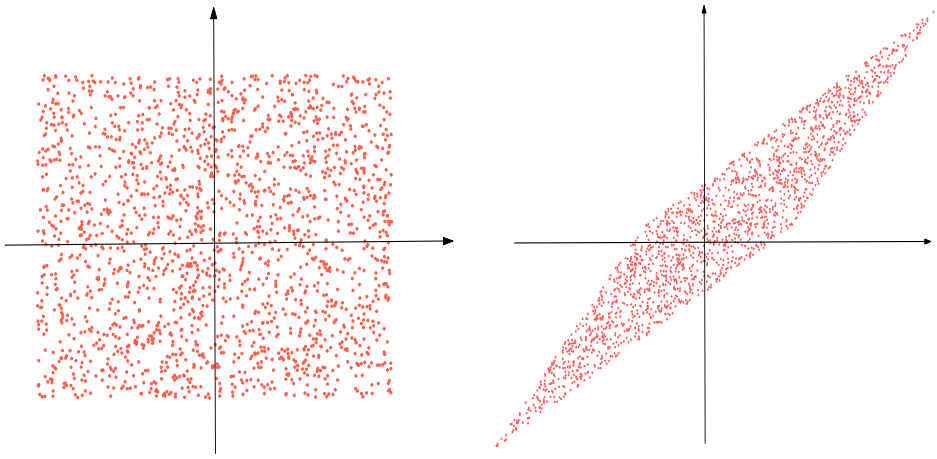
[Zhu et Rohwer, 1995] H. Zhu et R. Rohwer. Bayesian invariant measurements of generalisation. In *Neural Proc. Lett.*, volume 2 (6), pages 28–31, 1995.

Table des matières

Table des figures

Chapitre I

INTRODUCTION



Ce chapitre introductif expose le problème de séparation de sources et les motivations à la fois applicatives et théoriques qui ont poussé la communauté scientifique à se pencher sur ce problème. Le rappel des principales méthodes de séparation est présenté sous une forme comparative visant à déceler les points communs et les divergences au niveau de leurs principes. Cette introduction débouche sur les principales contributions de ce travail en indiquant le fil directeur reliant les différents chapitres de ce mémoire.

I.1 Position du problème

Avant de présenter brièvement les principales méthodes de séparation, nous allons essayer d'élucider les objectifs de ces méthodes. On va distinguer deux approches duales[1] qui se rejoignent dans un cas particulier.

[A] Séparation de sources comme un problème de reconstruction

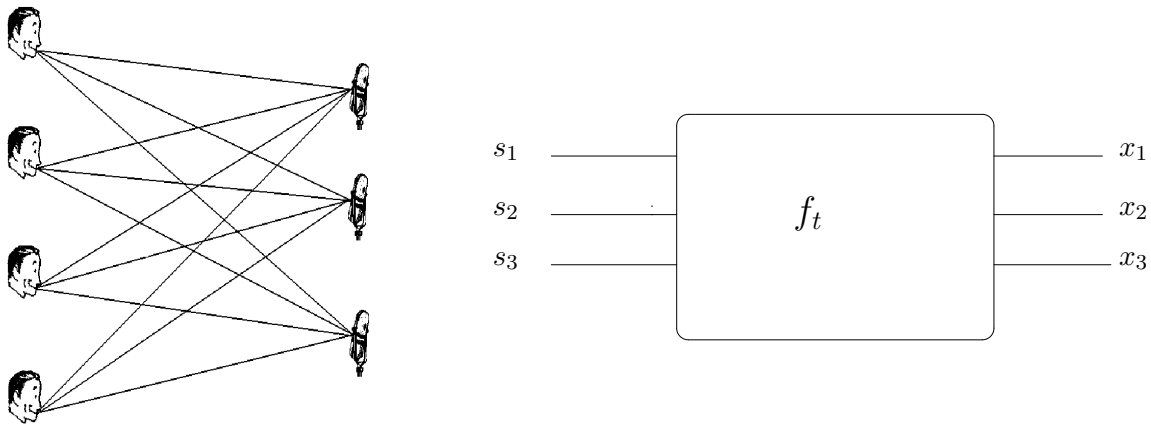
Dans un contexte physique, le problème de séparation de sources peut être considéré comme un problème d'identification. En effet, les signaux qu'on obtient sur les capteurs à la sortie d'un dispositif de mesure représentent l'image des signaux d'intérêt (les signaux **sources**) par une transformation modélisant les processus physiques de propagation et de mesure (voir figure (I.1)). Si on note $\boldsymbol{s}(t) = [s_1(t), ..., s_n(t)]^*$ le vecteur des n composantes sources à l'instant t ($t = 1, \ldots, T$), le vecteur des m observations $\boldsymbol{x}(t) = [x_1(t), ..., x_m(t)]^*$ est lié aux sources par l'équation suivante :

$$\boldsymbol{x}(t) = f_t(\boldsymbol{s}(1), ..., \boldsymbol{s}(T)),\ t = 1, \ldots, T$$

où $\{f_t\}_{t=1..T}$ est la transformation liant les sources et les observations. Quelque soit la complexité raisonable de la modélisation de cette transformation, on ne peut, dans les situations réelles, affirmer son exactitude d'où l'introduction d'un terme stochastique reflétant les erreurs de modélisation et aussi la présence d'autres sources non désirables qu'on appelle communément le **bruit**. On a alors la relation suivante entre les sources et les observations :

$$\boldsymbol{x}(t) = f_t(\boldsymbol{s}(1), ..., \boldsymbol{s}(T)) \odot \boldsymbol{\epsilon}(t),\ t = 1..T \tag{I.1}$$

où $\odot$ est l'opérateur de superposition du bruit $\boldsymbol{\epsilon}(t)$.



(a) (b)

Fig. I.1: (a) L'exemple traditionnellement repris dans la litérature est celui du "cocktail party" où plusieurs personnes parlent en même temps et les signaux sont mélangés sur les micros, (b) la modélisation du processus de mélange.

Nous avons ainsi un problème inverse : connaissant les données $\boldsymbol{x}_{1..T}$, l'objectif est de reconstruire les sources $\boldsymbol{s}_{1..T}$. Les performances de cette reconstruction sont directement liées à la forme des fonctions f_t (modélisation du problème direct) et au rapport signal sur bruit. Cette inversion est en général un problème mal posé d'où les techniques de régularisation [Tikhonov et Arsenin, 1977]. Afin d'introduire le problème de séparation de sources, on va simplifier le modèle d'observation en supposant que f_t ne dépend pas de l'instant t et qu'elle ne varie qu'en fonction de $\boldsymbol{s}(t)$ et que le bruit est additif, d'où la relation suivante :

$$\boldsymbol{x}(t) = f(\boldsymbol{s}(t)) + \boldsymbol{\epsilon}(t),\ t = 1..T. \tag{I.2}$$

[1] La notion de dualité va être reprise dans la conclusion et fera partie de l'une des perspectives théoriques de ce travail.

En séparation de sources, on introduit une difficulté supplémentaire : la fonction f n'est pas parfaitement connue. Ce n'est pas seulement la forme plus ou moins compliquée de la fonction f connue qui rend l'identification des sources difficile mais aussi la non connaissance de cette fonction. On voit clairement que le problème reste relativement difficile même si la fonction f possède une forme simple comme par exemple le cas linéaire. Dans ce cas, on introduit la matrice $\boldsymbol{A}$ de dimension $m * n$ qu'on appelle **matrice de mélange** et le modèle d'observation devient :

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t) + \boldsymbol{\epsilon}(t),\ t = 1..T \tag{I.3}$$

L'objectif est de restaurer les sources à partir des observations. La non connaissance de la matrice de mélange rend le problème mal posé (la solution n'est pas unique). Par conséquent, on doit imposer des contraintes sur les sources, sur le bruit et sur la matrice de mélange permettant d'assurer l'identifiabilité du modèle (I.3). Ces contraintes peuvent être de type statistique comme nous allons voir dans la section suivante (I.2) en rappelant les principales méthodes de séparation.

[B] Séparation de sources comme un problème de décomposition

On peut aussi considérer la séparation de sources comme la décomposition des observations multidimensionnelles $\boldsymbol{x}_{1..T}$ sur une base de signaux $\boldsymbol{y}_{1..T}$ indépendants (voir figure (I.2)). Nous avons l'habitude de manipuler des décompositions sur des bases orthogonales pertinentes [2] le long de la dimension temporelle [3] comme la transformée de Fourier ou la décomposition en ondelettes, dans l'objectif d'éliminer une certaine redondance et de capter l'information utile avec un nombre plus réduit d'échantillons. Dans le cas des signaux multi-composantes, on peut aussi envisager une décomposition le long de la dimension spatiale. Ayant plusieurs échantillons d'un vecteur de dimension suffisamment réduite (afin de distinguer la dimension spatiale de la dimension temporelle), des méthodes statistiques visant à décomposer le signal multi-composantes sur une base ayant des propriétés statistiques particulières ont prouvé leur utilité. Ainsi, l'analyse en composantes principales (ACP) est la recherche d'une base de signaux décorrélés. L'analyse en composantes indépendantes (ACI) est la recherche d'une base de signaux indépendants. A la différence des décompositions sur des bases orthogonales classiques, l'ACP ou l'ACI apprennent les bases directement à partir des données elles mêmes.

A la différence de la première approche, il n'est pas nécessaire de supposer que les données $\boldsymbol{x}_{1..T}$ proviennent physiquement de sources $\boldsymbol{s}_{1..T}$ indépendantes. C'est à partir des données $\boldsymbol{x}_{1..T}$ et une architecture fixée en avance qu'on essaie de construire une base jouissant de certaines propriétés statistiques comme la décorrélation ou l'indépendance. Dans le cas d'une architecture linéaire, on cherche à estimer une matrice $\boldsymbol{B}$ de telle façon que les signaux $\boldsymbol{y}_{1..T}$ obtenus par :

$$\boldsymbol{y}(t) = \boldsymbol{B}\boldsymbol{x}(t),\ t = 1, \ldots, T$$

soient le plus décorrélés possible (ACP) ou le plus indépendants possible (ACI) ou que leur distribution de probabilité soit le plus proche d'une distribution fixée p_0.

Dans la figure (I.3), on peut décerner la notion de dualité. En effet, dans la première approche, les données observées sont les sorties du système de mesure et on essaie de trouver les entrées qui expliquent le plus possible leur obtention. Nous voyons apparaître alors le principe du **maximum de vraisemblance**. Tandis que dans la deuxième approche, les données observées sont les entrées de notre système qui va produire les signaux $\boldsymbol{y}_{1..T}$. A titre illustratif, le premier système est à comparer à un canal de transmission où on cherche à remonter au signal émis connaissant le signal reçu. Tandis que le deuxième système est à comparer à un système de contrôle où on cherche à produire une action (les sorties $\boldsymbol{y}_{1..T}$) connaissant les informations mesurées $\boldsymbol{x}_{1..T}$.

Les deux approches peuvent se rejoindre sous certaines conditions comme nous allons voir dans le paragraphe suivant.

[2] Cette pertinence dépend bien entendu de la classe des signaux considérés.

[3] Ici, le temps indique un indice générique, ça peut être l'indice d'un pixel d'une image ou l'indice d'un cœfficient d'une transformation temps fréquence...

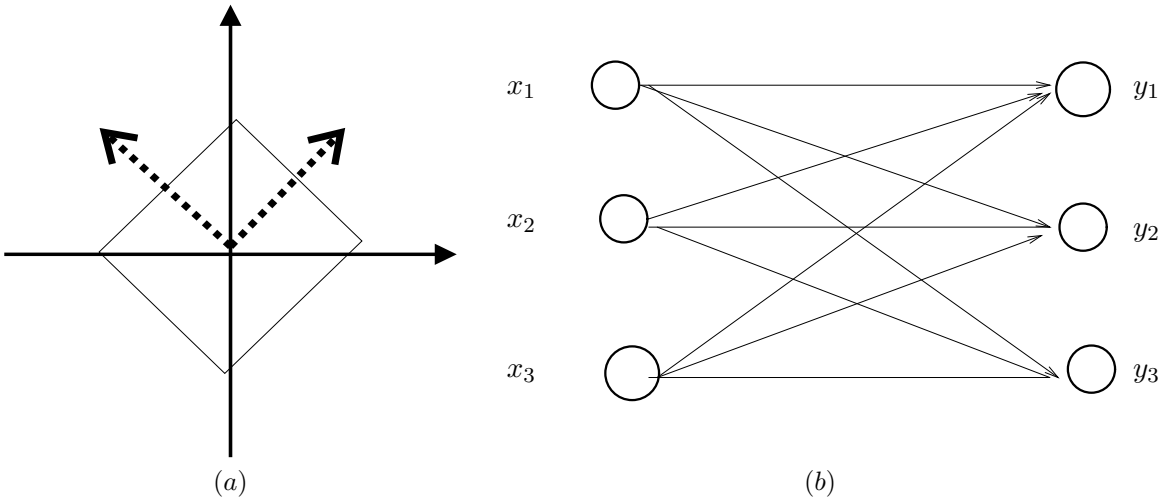


FIG. I.2: Recherche des directions indépendantes

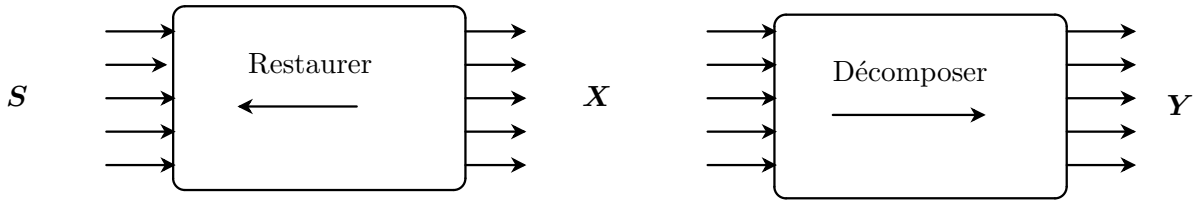


FIG. I.3: Restaurer ou décomposer ?

[C] Théorème de Darmois : Point de rencontre de ces deux approches

Dans le cas où le bruit d'observation dans l'équation (I.2) est nul et que le mélange est linéaire carré :

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t),\ t = 1..T \tag{I.4}$$

Autrement dit, les données $\boldsymbol{x}_{1..T}$ suivent le modèle de l'ACI. La recherche des composantes indépendantes $\boldsymbol{y}_{1..T}$ $(p(\boldsymbol{y}(t)) = \prod p_j(y_j(t)))$:

$$\boldsymbol{y}(t) = \boldsymbol{B}\boldsymbol{x}(t),\ t = 1..T \tag{I.5}$$

équivaut, à une permutation et à un facteur d'échelle près, à l'estimation des sources dans le modèle de mélange (I.4) à condition que ces sources soient indépendantes et au plus l'une d'entre elles soit gaussienne. Ceci est assuré par le théorème suivant de Darmois [Darmois, 1953] (regarder aussi [Comon, 1994] pour la relation de ce théorème avec l'analyse en composantes indépendantes) :

Théorème 1 *(Darmois* 1953*) Soient deux variables aléatoires X_1 et X_2 définies par :*

$$X_1 = \sum_{i=1}^{N} a_i x_i,\ \ X_2 = \sum_{i=1}^{N} b_i x_i,$$

où les x_i sont des variables aléatoires indépendantes. Si X_1 et X_2 sont indépendantes alors toutes les variables x_j tel que $a_j\, b_j \neq 0$ sont gaussiennes.

D'après les modèles (I.4) et (I.5), les signaux $\boldsymbol{y}_{1..T}$ sont liés aux signaux sources $\boldsymbol{s}_{1..T}$ par la relation linéaire suivante :

$$\boldsymbol{y}(t) = \boldsymbol{B}\boldsymbol{A}\boldsymbol{s}(t) = \boldsymbol{C}\boldsymbol{s}(t),\ t = 1..T$$

où $\boldsymbol{C}$ est le produit de la matrice séparatrice $\boldsymbol{B}$ et de la vraie matrice de mélange $\boldsymbol{A}$. En appliquant le théorème de Darmois pour des sources $\boldsymbol{s}$ indépendantes ayant au plus une composante gaussienne et en imposant l'orthogonalité de la matrice $\boldsymbol{C}$, il y a une équivalence entre les trois propositions suivantes :

(i) Les composantes de $\boldsymbol{y}$ sont deux à deux indépendantes.

(ii) Les composantes de $\boldsymbol{y}$ sont mutuellement indépendantes.

(iii) $\boldsymbol{C} = \boldsymbol{\Lambda}\boldsymbol{P}$ avec $\boldsymbol{\Lambda}$ une matrice diagonale et $\boldsymbol{P}$ une matrice de permutation

En assurant ainsi l'indépendance mutuelle (ou, de manière moins restrictive, l'indépendance deux à deux) des composantes de $\boldsymbol{y}$, on retrouve les signaux sources qui ont engendré les observations. Dans ce cas l'analyse en composantes indépendantes (ACI) est équivalente de point de vue objectif à la reconstruction des sources. On va constater cette équivalence en parcourant les principales méthodes de séparation. Cependant, on n'aboutit pas aux mêmes algorithmes de séparation car tout simplement les deux approches ne sont pas implémentées d'une manière optimale (impossible en pratique pour une raison commune qui est la non connaissance parfaite de la densité des sources) et donc au sein de chaque approche on peut avoir plusieurs variantes.

I.2 Quelques méthodes de séparation

I.2.1 Cas i.i.d

[A] Analyse en composantes indépendantes

L'analyse en composantes indépendantes peut être entrepris sans que les observations suivent le modèle de mélange :

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t) + \boldsymbol{\epsilon}(t),\ t = 1, \ldots, T.$$

En effet, supposons que les données i.i.d $[\boldsymbol{x}(1), ..., \boldsymbol{x}(T)]$ suivent la loi p_x^* et qu'on cherche une matrice $\boldsymbol{B}$ telle que les nouvelles données construites $[\boldsymbol{y}(t) = \boldsymbol{B}\boldsymbol{x}(t)]$ soient le plus indépendantes possible. Ceci peut être traduit géométriquement. On suppose tout d'abord que les données observées $\boldsymbol{x}_{1..T}$ sont décorrélées et de puissance égale à 1 (données blanchies) et comme on cherche à construire des sorties $\boldsymbol{y}_{1..T}$ indépendantes (donc décorrélées) la matrice recherchée $\boldsymbol{B}$ est une matrice unitaire ($\boldsymbol{B}\boldsymbol{B}^* = \boldsymbol{I}$). Quand la matrice $\boldsymbol{B}$ varie dans l'ensemble des matrices unitaires, la distribution du vecteur aléatoire $\boldsymbol{y}$ parcourt l'ensemble des distributions $\mathcal{Q}^*$ paramétré par la matrice $\boldsymbol{B}$:

$$\mathcal{Q}^* = \left\{p \,|\, p = p_x^*(\boldsymbol{B}^{-1}\boldsymbol{y}), \boldsymbol{B} \in \mathcal{U}_{n\times n}\right\} \subset \mathcal{P} = \{p \,|\, \int p = 1\}.$$

En choisissant une distance d entre deux distributions de probabilités[4] et en notant $\mathcal{P}_\Pi$ l'ensemble des distributions produits de leurs distributions marginales :

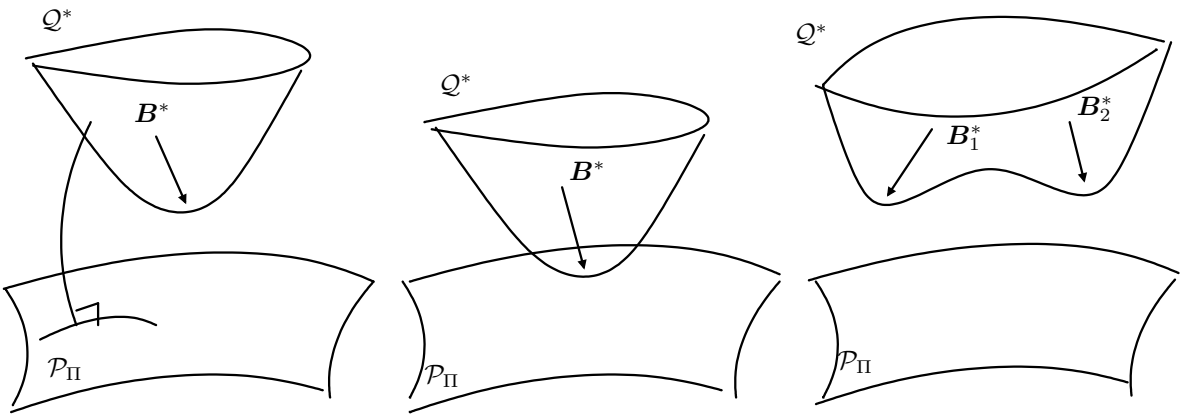
$$\mathcal{P}_\Pi = \left\{ p \,|\, p(\boldsymbol{y}) = \prod_{j=1}^{n} p_j(y_j) \right\},$$

l'analyse en composantes indépendantes repose alors sur le calcul de la distance entre les deux ensembles $\mathcal{Q}^*$ et $\mathcal{P}_\Pi$ (voir figure (I.4)) :

$$\hat{\boldsymbol{B}} = \underset{\boldsymbol{B}\in\mathcal{U}_{n\times n}}{\arg\min}\, d(p_x^*(\boldsymbol{B}^{-1}\boldsymbol{y}), \mathcal{P}_\Pi) \tag{I.6}$$

L'étude de l'existence et de l'unicité des solutions ainsi que leur calcul sont alors directement liées aux géométries des deux surfaces $\mathcal{Q}^*$ et $\mathcal{P}_\Pi$ et au choix de la distance d. Dans le cas où les observations $\boldsymbol{x}_{1..T}$ suivent un modèle de mélange linéaire non bruité et si la distance d choisie est la divergence de Kullback-Leibler, le théorème de Darmois assure l'existence et l'unicité de cette solution aux indéterminations d'échelle et de permutation près.

[4] Le choix d'une distance n'est pas arbitraire et doit être étudié dans un cadre géométrique. Le chapitre (VII) donne quelques notions nécessaires pour un raisonnement géométrique.



(a) (b) (c)

FIG. I.4: Le problème du calcul de la distance entre les ensembles $\mathcal{Q}^*$ et $\mathcal{P}_\Pi$ dépend de la géométrie de la surface $\mathcal{Q}^*$ (qui dépend de la vraie loi p^* de $\boldsymbol{x}$)

[A].1 Minimisation de l'information mutuelle

Si on choisit d la divergence de Kullback-Leibler, on obtient la définition usuelle de l'information mutuelle $\mathcal{I}(\boldsymbol{y})$:

$$\begin{aligned} \mathcal{I}(\boldsymbol{y}) &= \int p^*_{\boldsymbol{B}}(\boldsymbol{y}) \log \tfrac{p^*_{\boldsymbol{B}}(\boldsymbol{y})}{\prod p^*_j(y_j)} d\,\boldsymbol{y} \\ &= \int p^*(\boldsymbol{x}) \log \tfrac{p^*(\boldsymbol{x})}{\prod p^*_j(y_j)} d\,\boldsymbol{x} \\ &= -\textstyle\sum_{j=1}^n \mathrm{E}\big[\log p^*_j(y_j)\big] + \int p^*(\boldsymbol{x}) \log p^*(\boldsymbol{x}) d\,\boldsymbol{x} \end{aligned} \tag{I.7}$$

On note ici deux points importants qui ne sont pas mentionnés dans la littérature :

Remarque 1 *Dans toutes les équations précédentes, nous avons noté p^* pour désigner "la vraie" loi des données. Par "vraie loi", on veut dire la loi sous laquelle on intégre quand on approche le calcul de l'espérance par une moyenne empirique. Ainsi, l'espérance dans l'équation (I.7) peut être approchée par un moyennage sur les échantillons :*

$$\mathrm{E}\big[\log p^*_j(y_j)\big] \approx \frac{1}{T} \sum_{t=1}^{T} \log p^*_j(y_j(t))$$

Remarque 2 *Le passage de (I.6) à (I.7) n'est pas aussi directe. Le calcul de la distance entre les deux ensembles $\mathcal{Q}^*$ et $\mathcal{P}_\Pi$ est obtenue en prenant le minimum de la distance entre deux points q et p avec q variant dans $\mathcal{Q}^*$ (en faisant varier $\boldsymbol{B}$) et p variant dans $\mathcal{P}_\Pi$. Or dans l'équation (I.6) on ne fait varier que la matrice $\boldsymbol{B}$. Ceci est justifié par le fait que le point p dans $\mathcal{P}_\Pi$ varie aussi mais comme étant la projection de q sur l'ensemble $\mathcal{P}_\Pi$. La distribution $p^*_j(y_j)$ est alors définie par :*

$$p^*_j(y_j) = \int_{\boldsymbol{y}_{-j}} p^*_{\boldsymbol{B}}(\boldsymbol{y}) d\,\boldsymbol{y}_{-j}$$

où $\boldsymbol{y}_{-j}$ désigne le vecteur $\boldsymbol{y}$ sauf la j^{me} composante.

Autrement dit, on ne cherche pas une distribution particulière pour les $\boldsymbol{y}$ parmi toutes les distributions produits de leurs distributions marginales, seule l'indépendance suffira à retrouver cette distribution. Ce point va être repris lors de la présentation du méthode du maximum de vraisemblance où on aboutit à une remarque équivalente.

En imposant la décorrélation des composantes de $\boldsymbol{y}$, la minimisation de la somme des entropies marginales est équivalente à la minimisation de l'information mutuelle. En définissant la négentropie $J(y)$ comme la différence entre l'entropie de y et l'entropie de la variable gaussienne y_G correspondante (mesure de la distance à la gaussienne) :

$$J(y) = H(y_G) - H(y)$$

le critère (I.7) peut être ré-interprété comme la maximisation de la somme des négentropies marginales ou de la non gaussianité des composantes de $\boldsymbol{y}$. En pratique on ne connaît pas la distribution $p_j^*(y_j)$ et on est alors amené à approcher $\mathrm{E}\big[\log p_j^*(y_j)\big]$. Le calcul de l'espérance est résolu par un moyennage temporel. Cependant plusieurs approximations de $\log p_j^*(y_j)$ ont été considérées. Par exemple, une approximation polynomiale conduit à l'utilisation des cumulants d'ordre supérieurs notamment le cumulant d'ordre 4 (kurtosis) [Comon, 1994; Hyvärinen et Oja, 1997; Delfosse et Loubaton, 1995; Malouche et Macchi, 1998]. D'autres approximations non plynomiales ont été proposées dans [Hyvärinen, 1999].

[A].2 Décorrélation non linéaire

Une autre définition équivalente de l'indépendance entre deux variables y_1 et y_2 est la suivante :

$$\mathrm{E}\big[f(y_1)\, g(y_2)\big] = \mathrm{E}\big[f(y_1)\big]\, \mathrm{E}\big[g(y_2)\big] \tag{I.8}$$

pour toutes fonctions continues f et g. C'est donc une généralisation de la définition de la décorrélation obtenue en prenant les deux fonctions f et g égales à l'identité. Le travail pionnier en analyse en composantes indépendantes de Jutten, Hérault et Ans [Hérault et Ans, 1984; Hérault, 1985; Ans *et al.*, 1985; Jutten, 2000; Jutten et Herault, 1991], développé par Cichocki et Unbehauen [Cichocki et Moszczynski, 1992; Cichocki *et al.*, 1994; Cichocki et Unbehaunen, 1996] et repris d'une manière plus générale dans le cadre des "fonctions d'estimation" par Amari et Cardoso [Amari et Cardoso, 1997; Cardoso et Labeld, 1996], s'appuient sur l'équation (I.8) comme condition d'équilibre d'un algorithme de séparation. Autrement dit, on construit en général un algorithme de type gradient proportionel à la différence des deux termes de l'équation (I.8) et on étudie *a posteriori* la convergence et la stabilité. En le comparant avec la minimisation de l'information mutuelle on dégage les points suivants.

- L'information mutuelle est un critère qu'on doit minimiser tandis que (I.8) est une équation qu'on doit résoudre.
- L'information mutuelle est dérivée d'un principe géométrique plus général et peut être facilement généralisée en changeant la mesure de divergence ou l'ensemble des probabilités dans lequel on veut que la loi de $\boldsymbol{y}$ se trouve. La décorrélation non linéaire est une propriété qu'on doit vérifier et on ne distingue pas une mesure de distance à cette propriété.
- En pratique, on doit choisir les fonctions f et g et on est loin de la définition (I.8) qui exige de vérifier l'égalité pour toutes les fonctions continues f et g. Tandis qu'avec la minimisation de l'information mutuelle on doit estimer $G = \log p_j^*(y_j)$. Une étude au niveau de la forme des algorithmes montre que les deux méthodes sont équivalentes en prenant f comme dérivée de $G = \log p_j^*(y_j)$ et g la fonction identité. Le choix de G ou du couple (f, g) n'est pas en général critique. La preuve pratique est que ces méthodes réussissent, pour un choix fixe de ces non linéarités, à séparer des sources ayant des caratéristiques non étroitement liées à ce choix. Ce point va être repris dans la section suivante lors de la présentation du maximum de vraisemblance.

[A].3 Infomax

Bien que classée dans la litérature avec les méthodes du maximum de vraisemblance puisqu'elle aboutit au même algorithme de séparation, nous pensons que l'Infomax est une technique à part entière faisant partie de la classe des méthodes de décomposition et s'appuyant sur une théorie solide (logique des questions) qui commence à être développée [Knuth, 2000, 2001, 2002; Fry, 2001].

Les observations $\boldsymbol{x}$ subissent une opération linéaire $\boldsymbol{B}$ suivie d'une transformation non linéaire Φ composante par composante (voir figure (I.5)). La sortie du système est alors le vecteur $\boldsymbol{y} = \Phi(\boldsymbol{B}\boldsymbol{x})$. Le principe de l'Infomax [Bell et Sejnowski, 1995] est alors de maximiser le flux d'information entre les entrées $\boldsymbol{x}$ et

les sorties $\boldsymbol{y}$ du système. Le flux d'information est mesuré par l'information mutuelle $\mathcal{I}(\boldsymbol{x}, \boldsymbol{y})$ qui s'écrit en fonction des entropies :

$$\mathcal{I}(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{H}(\boldsymbol{y}) - \mathcal{H}(\boldsymbol{y} \mid \boldsymbol{x})$$

où $\mathcal{H}(\boldsymbol{y} \mid \boldsymbol{x})$ peut être interprétée comme une mesure du caractère aléatoire de $\boldsymbol{y}$ sachant $\boldsymbol{x}$. Comme la matrice $\boldsymbol{B}$ intervient dans la relation déterministe entre $\boldsymbol{x}$ et $\boldsymbol{y}$, $\mathcal{H}(\boldsymbol{y} \mid \boldsymbol{x})$ ne dépend pas de $\boldsymbol{B}$. Ainsi, l'estimation de $\boldsymbol{B}$ revient à maximiser l'entropie de la sortie $\boldsymbol{y}$. On montre qu'en prenant la fonction Φ comme la distribution cumulative des sources on retrouve le même algorithme de séparation qu'en utilisant le maximum de vraisemblance [Cardoso, 1997].

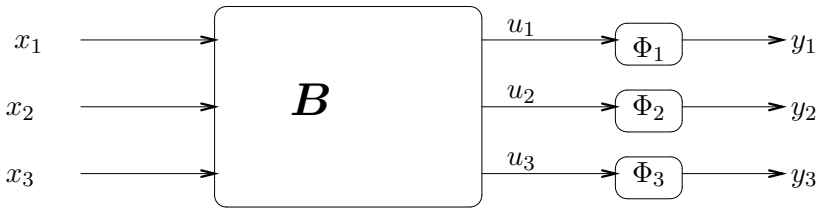


FIG. I.5: Infomax : maximiser le flux d'informations entre les entrées et les sorties du système

On peut aussi donner une interprétation géométrique de la méthode Infomax. En effet, si les échantillons $\boldsymbol{x}_{1..T}$ suivent une loi p_x^* alors les variables $\boldsymbol{u}_{1..T}$ suivent la loi paramétrique p_u^* (paramétrée par la matrice $\boldsymbol{B}$) définie par :

$$\mathrm{p}_u^*(\boldsymbol{u}) = \mathrm{p}_x^*(\boldsymbol{B}^{-1}\boldsymbol{u})|\boldsymbol{B}|^{-1}$$

et les sorties $\boldsymbol{y}_{1..T}$ suivent la loi p_y^* définie par :

$$\mathrm{p}_y^*(\boldsymbol{y}) = \mathrm{p}_u^*(\boldsymbol{u}) \Big/ \prod_j |\Phi_j'(u_j)|$$

où $\prod_j |\Phi_j'(u_j)| = \mathcal{J}_\Phi(\boldsymbol{u})$ est le Jacobien de la transformation Φ.

En développant l'expression de l'entropie du vecteur $\boldsymbol{y}$ (qui est une fonction de B puisque la distribution de $\boldsymbol{y}$ est une fonction de $\boldsymbol{B}$), on trouve :

$$\begin{aligned} \mathcal{H}(\boldsymbol{y}) &= -\int \mathrm{p}_y^*(\boldsymbol{y}) \log \mathrm{p}_y^*(\boldsymbol{y}) d\,\boldsymbol{y} \\ &= -\int \mathrm{p}_u^*(\boldsymbol{u}) \log \frac{\mathrm{p}_u^*(\boldsymbol{u})}{\prod_j |\Phi_j'(u_j)|} d\,\boldsymbol{u} \end{aligned}$$

L'opposé de l'entropie $\mathcal{H}(\boldsymbol{y})$ est donc une divergence de Kullback-Leibler entre p_u^* et $\prod_j |\Phi_j'(u_j)|$. On retrouve ainsi l'interprétation géométrique donnée dans le paragraphe ([A].1) en considérant le terme $\prod_j |\Phi_j'(u_j)|$ comme la projection de la distribution p_u^* sur l'espace $\mathcal{P}_\Pi$ définie plus haut[5] (ce qui justifie l'application de Φ terme à terme).

L'équivalence de l'Infomax avec la méthode du maximum de vraisemblance se produit dans le cas où :

$$\prod_j |\Phi_j'(u_j)| = \prod_j \mathrm{p}_s(u_j) \implies \forall\, j\ \Phi_j(u_j) = \int_{-\infty}^{u_j} \mathrm{p}_s(s_j) d\, s_j,$$

autrement dit lorsque la non linéarité Φ est la distribution cumulative des sources.

[B] Maximum de vraisemblance

Contrairement à l'analyse en composantes indépendantes, le maximum de vraisemblance est conceptuellement lié au problème de reconstruction. En supposant dans un premier temps que $\boldsymbol{x}_{1..T}$ suivent un modèle de mélange linéaire, carré, inversible et non bruité :

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t),\ t = 1..T$$

[5] Par conséquent un traitement rigoureux consiste à adapter la non linéarité Φ

il s'agit de trouver la matrice $\boldsymbol{A}$ qui explique le plus possible les données $\boldsymbol{x}_{1..T}$. Autrement dit, on doit maximiser la probabilité que la proposition "a = La matrice de mélange est $\boldsymbol{A}$ ¡' implique la proposition "x = Les données observées sont $\boldsymbol{x}_{1..T}$!" : C'est le principe du maximum de vraisemblance,

$$\hat{\boldsymbol{A}} = \arg\max_{\boldsymbol{A}} p(\boldsymbol{x}_{1..T} \,|\, \boldsymbol{A}).$$

En notant p_s la distribution des sources, l'opposé du logarithme normalisé de la vraisemblance s'écrit :

$$\begin{array}{rcl} -\frac{1}{T}\log p(\boldsymbol{x}_{1..T} \,|\, \boldsymbol{A}) & = & -\frac{1}{T}\sum_{t=1}^{T}\log p_s(\boldsymbol{A}^{-1}\boldsymbol{x}(t)) \\ & \approx & \mathrm{E}\big[\log p_s(\boldsymbol{A}^{-1}\boldsymbol{x})\big] \end{array} \tag{I.9}$$

En notant $\boldsymbol{B} = \boldsymbol{A}^{-1}$, on retrouve le critère de la minimisation de l'information mutuelle (I.7) à une constante près avec deux petites différences :

1. Le moyennage temporel dans (I.7) est une approximation tandis que dans (I.9) est une simple normalisation d'une somme qui existe déjà.
2. L'un des reproches concernant le maximum de vraisemblance qu'on retrouve dans la litérature de la séparation de sources est que la connaissance de p_s est nécessaire pour appliquer le MV tandis qu'avec l'information mutuelle on n'a besoin que de la connaissance de la forme de $\log p_s(\boldsymbol{A}^{-1}\boldsymbol{x})$ (du moment non quadratique). Cet inconvénient n'existe plus si on fixe la forme de p_s et qu'on estime ses paramètres. En effet, on n'a pas besoin de considérer toutes les statistiques de $\boldsymbol{x}$ (toute la fonction p_x) pour retrouver la matrice $\boldsymbol{A}$. Nous pensons que les approximations de $\log p_s(\boldsymbol{A}^{-1}\boldsymbol{x})$ par une fonction non quadratique ou d'une manière équivalente l'approximation de p_s par une distribution manipulable visent à capter des statistiques suffisantes pour retrouver la matrice $\boldsymbol{A}$ et n'ont pas pour rôle de représenter au mieux les sources.

La technique du MV a été appliquée avec succés en séparation de sources [Gaeta et Lacoume, 1990; Pham *et al.*, 1992; Pham, 1996]. L'introduction du gradient naturel par Amari [Amari *et al.*, 1996] ou le gradient relatif par Cardoso [Cardoso et Labeld, 1996] ont amélioré l'aspect algorithmique en exploitant la particularité du problème.

Le principe du maximum de vraisemblance présente des avantages qui le distinguent de la minimisation de l'information mutuelle :

1. On peut tenir compte du bruit dans la modélisation du mélange [Mohammad-Djafari, 1999; Bermond, 2000; Belouchrani et Cardoso, 1995; Moulines *et al.*, 1997], en maximisant :

   $$p(\boldsymbol{x}_{1..T} \mid \boldsymbol{A}) = \int p_\epsilon(\boldsymbol{x}_{1..T} \mid \boldsymbol{s}_{1..T}, \boldsymbol{A}) p(\boldsymbol{s}_{1..T}) d\,\boldsymbol{s}_{1..T}$$

   où p_ϵ est la loi du bruit.
2. On n'a plus besoin de l'hypothèse de l'indépendance des sources et toute information sur leur dépendance peut être prise en compte dans p_s.
3. Le modèle de mélange peut être enrichi (mélange non linéaire, introduction d'autres variables...) sans avoir de conséquences sur la méthodologie du maximum de vraisemblance.

[B].1 Approche bayésienne

En applicant le principe du maximum de vraisemblance dans le paragraphe précédent, on est déjà dans une logique bayésienne. En effet, les sources sont considérées comme des variables aléatoires et ont été marginalisées pour obtenir la vraisemblance $p(\boldsymbol{x}_{1..T} \,|\, \boldsymbol{A})$. On peut aussi modéliser les informations *a priori* sur les éléments de la matrice $\boldsymbol{A}$ en leur attribuant une loi de probabilité et former ainsi la loi jointe :

$$p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{A} \,|\, \boldsymbol{I}) = p(\boldsymbol{x}_{1..T} \mid \boldsymbol{s}_{1..T}, \boldsymbol{A}, \boldsymbol{I}) p(\boldsymbol{s}_{1..T} \mid \boldsymbol{A}, \boldsymbol{I}) p(\boldsymbol{A} \mid \boldsymbol{I})$$

où $\boldsymbol{I}$ représente notre connaissance *a priori* sur le problème comme par exemple la forme des distributions des sources, la forme de la loi *a priori* de $\boldsymbol{A}$...

Plusieurs directions peuvent alors être envisagées comme l'estimation jointe de $\boldsymbol{s}_{1..T}$ et de $\boldsymbol{A}$, marginalisation de $\boldsymbol{s}_{1..T}$ et estimation de $\boldsymbol{A}$ ou le contraire. Ceci est devenu possible avec le développement des techniques du calcul bayésien comme les méthodes de Monte Carlo par chaînes de Markov (MCMC).

Nous allons revenir en détail sur cette approche et sur l'échantillonnnage bayésien dans le chapitre (II). Les premiers travaux sur l'application de l'approche bayésienne en séparation de sources sont ceux de [Knuth, 1999],[Mohammad-Djafari, 1999], [Rowe, 1998]. Les techniques bayésiennes ont été utilisées dans [Senecal, 2000, 2002; Snoussi et Mohammad-Djafari, 2002c].

[B].2 Méthodes algébriques : JADE / FOBI

Bien que différentes des méthodes de maximum de vraisemblance, nous préférons les classer parmi celles-ci. En effet, avec le maximum de vraisemblance on cherche à trouver une matrice $\boldsymbol{A}$ qui explique le plus possible la loi des observations $p(\boldsymbol{x}_{1..T} \mid \boldsymbol{A})$ et donc nécessairement à rapprocher toutes les statistiques $\mathrm{E}\big[f(\boldsymbol{x})\big]$ aux statistiques du modèle du mélange $\mathrm{E}\big[f(\boldsymbol{A}\boldsymbol{s})\big]$. En exploitant l'indépendance des sources et la linéarité du mélange il est souvent possible de se contenter aux cumulants d'ordre supérieur à deux comme les cumulants d'ordre 4 pour identifier la matrice $\boldsymbol{A}$: c'est le principe des méthodes algébriques (ou tensorielles).

Les cumulants d'ordre 4 de $\boldsymbol{x}$ $[cum(x_i, x_j, x_k, x_l)]_{i,j,k,l=1..n}$ forment un tenseur (matrice d'ordre 4). L'opérateur linéaire C induit par ce tenseur agissant sur l'espace des matrices est par définition :

$$[\mathrm{C}(\boldsymbol{M})]_{ij} = \sum_{kl} m_{kl}\; cum(x_i, x_j, x_k, x_l).$$

Si on note $\boldsymbol{B} = \boldsymbol{A}^{-1}$, on montre que la matrice $\boldsymbol{M} = \boldsymbol{b}_j \boldsymbol{b}_j^*$ où $\boldsymbol{b}_j$ est la j^{me} ligne de $\boldsymbol{B}$ est une matrice propre de l'opérateur C avec la valeur propre $kurt(s_j)$:

$$\mathrm{C}(\boldsymbol{b}_j \boldsymbol{b}_j^*) = kurt(s_j)\, \boldsymbol{b}_j \boldsymbol{b}_j^*$$

Si ces valeurs propres sont distinctes[6] alors la matrice $\boldsymbol{B}$ est identifiée en calculant les matrices propres $\boldsymbol{M}_j$ de C. Si elles ne sont pas distinctes, une décomposition en valeurs singulières de ces matrices est alors nécessaire en espérant que cette nouvelle décomposition ne redonne pas des valeurs singulières identiques.

En pratique, cette méthode est implémentée par la technique JADE (Joint Approximate Diagonalization of Eigenmatrices) [Cardoso et Souloumiac, 1993]. Elle se base sur le fait que la matrice $\boldsymbol{B}$ diagonalise toutes les matrices $\mathrm{C}(\boldsymbol{M})$ $\forall \boldsymbol{M}$. Un choix judicieux est de prendre les $\boldsymbol{M}_j$, matrices propres de C, et de diagonaliser conjointement les matrices $\mathrm{C}(\boldsymbol{M}_j)$. Le critère de diagonalité est la minimisation de la somme des carrés des termes non diagonaux ou d'une manière équivalente la maximisation des termes diagonaux[7] :

$$\mathcal{J}_{JADE}(\boldsymbol{B}) = \sum_j \|diag(\boldsymbol{B}\mathrm{C}(\boldsymbol{M}_j)\boldsymbol{B}^*)\|^2 \tag{I.10}$$

En comparant cette méthode à la technique du maximum de vraisemblance, on dégage les deux points suivants.

1. Dans la méthode du maximum de vraisemblance, l'équivalent des méthodes algébriques serait de prendre des distributions des sources p_s dont le logarithme est une fonction polynomiale. Cependant, le maximum de vraisemblance impose une distance particulière entre les tenseurs statistiques. A titre illustratif, si on travaille avec les matrices de covariance (tenseurs de second ordre), le critère du maximum de vraisemblance est la distance de Kullback-Leibler entre la covariance emprique des observations et la covariance théorique. Cette distance pourrait rendre le problème d'optimisation difficile mais bénéficie des propriétés asymptotiques comme la consistance et l'efficacité.
2. Dans JADE, on remarque que la diagonalisation des matrices $\boldsymbol{B}\,\mathrm{C}(\boldsymbol{M}_j)\,\boldsymbol{B}^*$ suffit à retrouver la matrice $\boldsymbol{B}$. Les valeurs des éléments diagonaux qui représentent les statistiques des sources vont être ainsi estimées. On trouve la même constatation en utilisant le maximum de vraisemblance. Quand on fixe la forme de la distribution p_s, il ne faut pas aussi fixer les valeurs des statistiques des sources mais il faut les estimer dans l'algorithme de séparation.

[6] Dans ce cas la méthode FOBI [Cardoso, 1989] est plus simple à mettre en oeuvre.

[7] Cette équivalence est due au fait que la somme des carrés de tous les termes est constante.

I.2.2 Exploitation de la corrélation

Dans les méthodes précédentes, on a supposé que les sources sont temporellement indépendantes et identiquement distribuées[8]. Ce qui nous a empêché de modéliser les sources par des gaussiennes (ou ce qui revient au même à utiliser les statistiques d'ordre 2). Cependant, si on abandonne l'hypothèse de l'indépendance temporelle, la matrice de mélange peut être identifiable en exploitant les matrices d'intercovariance. Autrement dit, on peut supposer que les sources sont gaussiennes (en se limitant aux statistiques d'ordre deux) pourvu que les fonctions d'autocorrélation des sources soient différentes et non réduites à la distribution de Dirac δ.

Comme dans le cas des sources i.i.d. où nous avons le choix entre la maximisation de la vraisemblance (I.9) ou l'ajustement des statistiques d'ordre supérieure (I.10), on a deux types de méthodes :

1. `Maximum de vraisemblance` :
   Chaque source $[s_j(1), ..., s_j(T)]$ (en considérant tous les échantillons) est un processus gaussien avec une matrice de covariance $\boldsymbol{C}_j$:
   $$[s_j(1), ..., s_j(T)] \sim \mathcal{N}(0, \boldsymbol{C}_j)$$
   On suppose de plus que les sources sont stationnaires et donc que les matrices $\boldsymbol{C}_j$ sont des matrices de Toeplitz. Sous l'approximation circulante [Hunt, 1971], ces matrices se diagonalisent dans la base de Fourier :
   $$\boldsymbol{W}\boldsymbol{C}_j\boldsymbol{W}^* \approx \boldsymbol{\Lambda}_j$$
   où $\boldsymbol{W}$ est la matrice de Fourier et $\boldsymbol{\Lambda}_j$ la matrice diagonale contenant le spectre de la j^{me} source. En passant dans le domaine de Fourier[9], le critère du maximum de vraisemblance se met sous la forme d'une somme pondérée de divergences de Kullback-Leibler entre les matrices de covariance spectrales [Snoussi *et al.*, 2001; Cardoso *et al.*, 2002] :
   $$\mathcal{J}_{MV} = \sum_{\nu} \delta_\nu \mathcal{D}_{KL}(\boldsymbol{R}_{xx}(\nu) \;||\; \boldsymbol{A}\boldsymbol{R}_{ss}(\nu)\boldsymbol{A}^* + \boldsymbol{R}_\epsilon) \tag{I.11}$$
   où les spectres des sources $\boldsymbol{R}_{ss}$ et la covariance du bruit $\boldsymbol{R}_\epsilon$ sont à estimer lors de la séparation. Nous allons revenir en détail sur cette méthode dans le chapitre (V).

2. `Ajustement des matrices d'autocovariance` :
   Comme le maximum de vraisemblance cherche à ajuster toutes les matrices de covariance (I.11) avec une mesure qui découle de son expression (dans le cas gaussien, c'est la divergence de Kullback-Leibler), on peut essayer de se contenter de calculer les matrices d'autocovariance $\boldsymbol{R}_{xx}(\tau) = \mathrm{E}\big[\boldsymbol{x}(t)\boldsymbol{x}(t+\tau)^*\big]$ pour quelques retards temporels $\tau = 1, \ldots, K$ et les ajuster aux matrices théoriques $\boldsymbol{A}\boldsymbol{R}_{ss}(\tau)\boldsymbol{A}^*$ en choisissant une autre mesure de rapprochement entre deux matrices comme par exemple la distance quadratique. Si le mélange est non bruité ou si on connaît la covariance du bruit, l'ajustement des covariances se transforme en un problème de diagonalisation conjointe avec le critère suivant à minimiser :
   $$\mathcal{J}(\boldsymbol{B}) = \sum_{\tau} off(\boldsymbol{B}\boldsymbol{R}_{xx}(\tau)\boldsymbol{B}^*)$$
   avec $\boldsymbol{B} = \boldsymbol{A}^{-1}$ la matrice unitaire recherchée. C'est le principe de l'algorithme SOBI [Belouchrani *et al.*, 1997]. Cette diagonalisation conjointe peut être aussi effectuée sur les matrices d'autocovariance spectrales [Rahbar et Reilly, 2001] ou les matrices d'autocovariance temps fréquence [Belouchrani et Amin, 1997].

I.2.3 Exploitation de la non stationnarité

On peut aussi se baser sur la non stationnarité des sources. Par exemple, en supposant que les sources sont temporellement indépendantes mais que leurs variances varient en fonction du temps, on peut se limiter aux

[8] Ici l'hypothèse i.i.d. est considérée pour la dimension temporelle.
[9] En passant dans le domaine de Fourier, on va plutôt exploiter la non stationnarité spectrale.

statistiques du second ordre pour séparer les sources. En effet, à chaque instant t, la matrice de covariance théorique du vecteur $\boldsymbol{x}(t)$ se met sous la forme :

$$\begin{aligned} \boldsymbol{R}_{xx}(t) &= \boldsymbol{A}\mathrm{E}\big[\boldsymbol{s}(t)\boldsymbol{s}(t)^*\big]\boldsymbol{A}^* + \mathrm{E}\big[\boldsymbol{\epsilon}(t)\boldsymbol{\epsilon}(t)^*\big] \\ &= \boldsymbol{A}\boldsymbol{R}_{ss}(t)\boldsymbol{A}^* + \boldsymbol{R}_\epsilon \end{aligned} \tag{I.12}$$

On peut ainsi exploiter la variation de la covariance des sources $\boldsymbol{R}_{ss}(t)$ (ce qui implique la variation de $\boldsymbol{R}_{xx}(t)$) au cours du temps pour identifier la matrice de mélange $\boldsymbol{A}$ en ajustant les matrices de covariance empiriques de $\boldsymbol{x}$ aux matrices de covariance théoriques (I.12). Cette identification s'accompagne nécessairement de l'estimation des covariances des sources et du bruit.

De même que dans les cas précédents, on a le choix entre le maximum de vraisemblance (ajustement des matrices avec la divergence de Kullback-Leibler) et la méthode algébrique (utilisation d'une autre métrique pour l'ajustement) :

1. `Maximum de vraisemblance :`
   Se limitant aux techniques de second ordre se traduit par la modélisation des sources par des gaussiennes. Comme on ne peut pas estimer toutes les variances des sources (autant de variances que d'échantillons), on divise l'intervalle temporel[10] en L sous intervalles ($\mathcal{I} = \bigcup_{l=1}^{L} \mathcal{I}_l$) [Pham et Cardoso, 2001; Cardoso *et al.*, 2002] où les variances sont supposées constantes. Le critère du maximum de vraisembance est alors une somme pondérée de distances de Kullback-Leibler entre les covariances empiriques $\hat{\boldsymbol{R}}_{xx}(l)$ et les covariances théoriques $\boldsymbol{A}\boldsymbol{R}_{ss}(l)\boldsymbol{A}^* + \boldsymbol{R}_\epsilon$:

   $$\mathcal{J}_{MV} = \sum_{l=1}^{L} \alpha_l D_{KL}(\hat{\boldsymbol{R}}_{xx}(l) \parallel \boldsymbol{A}\boldsymbol{R}_{ss}(l)\boldsymbol{A}^* + \boldsymbol{R}_\epsilon)$$

   Remarque 3 *Cette répartition en sous intervalles est fixée en avance selon une connaissance a priori du profil des variances. La modélisation des sources par des mélanges de gaussiennes est très similaire à cette approche avec une répartition automatique des échantillons en groupes. Chaque groupe est représenté par une gaussienne. Ceci débouche sur une remarque intéressante brièvement exposé dans [Pham et Cardoso, 2001] et qu'on va développer dans le chapitre (III) sur la connection entre la non stationnarité et la non gaussianité des sources.*

2. `Ajustement des matrices de covariance :`
   Dans le cas non bruité, l'ajustement des matrices de covariance est un problème de diagonalisation conjointe[11]. On cherche à diagonaliser les matrices $\boldsymbol{B}\boldsymbol{R}_{xx}(t)\boldsymbol{B}$ en minimisant leur distances à leurs matrices diagonales. En choisissant la distance de Kullback-Leibler, on retrouve le critère de l'information mutuelle (ou d'une manière équivalente le maximum de vraisemblance). Les matrices de covariance sont calculées en divisant l'intervalle [1 T] en sous intervalles [Choi et Cichocki, 2000; Souloumiac, 1995] comme dans le cas du maximum de vraisemblance ou calculées localement en utilisant un noyau h [Matsuoka *et al.*, 1995] :

   $$\boldsymbol{R}_{xx}(t) \approx \sum_{\tau} h(\tau)\boldsymbol{x}(t-\tau)\boldsymbol{x}(t-\tau)^*$$

I.3 Contributions et organisation du document

Dans le chapitre (II), nous exposons l'approche bayésienne en séparation de sources. Nous distinguons l'aspect théorique de l'aspect technique de cette approche. Sur le plan théorique, cette approche présente plusieurs avantages :

1. En introduisant la loi $p(\boldsymbol{A})$, nous pouvons prendre en compte toute information *a priori* sur ses éléments. Par ailleurs, ceci nous permet de dépasser les limites imposées par l'existence de $\boldsymbol{A}^{-1}$ (nombre de sources = nombre de capteurs). On peut aussi intégrer par rapport à cette matrice pour obtenir la loi marginale des sources.

[10] On rappelle que le temps t est un indice générique qui peut aussi désigner l'indice d'un pixel d'une image, la fréquence, l'indice temps fréquence...

[11] Arès une étape de blanchiement.

2. En introduisant une loi *a priori* pour les hyperparamètres[12], on peut aussi s'affranchir de certaines difficultés liées à la dégénérescence de la vraisemblance lorsqu'il s'agit d'estimer ces hyperparamètres.
3. En introduisant des variables cachées, on peut enrichir la modélisation des sources.
4. On tient compte explicitement du bruit dans le modèle d'observation.

Afin de profiter des avantages de l'approche bayésienne, on doit effectuer des intégrations. Ceci n'est pas toujours possible à réaliser analytiquement. Le calcul bayésien offre ainsi des méthodes numériques basées sur l'échantillonnage.

Dans les chapitres suivants, nous considérons un mélange linéaire instantané bruité. Le point commun de ces chapitres est l'exploitation de la non stationnarité que ce soit dans le domaine temporel, spatial, fréquentiel ou temps fréquence. Les algorithmes proposés intègrent implicitement la reconstruction des sources contrairement aux méthodes estimant la matrice de mélange en ajustant les statistiques. Ce point a aussi son importance car une bonne estimation de la matrice de mélange n'implique pas forcément une bonne estimation des sources.

Dans le chapitre (III), on considère des sources monovariées (1-D) qu'on modélise par des mélanges de gaussiennes. L'estimation des variances des gaussiennes provoque une dégénérescence de la vraisemblance d'où la pénalisation par des lois inverses gamma [Snoussi et Mohammad-Djafari, 2001]. En considérant les étiquettes des gaussiennes comme des variables cachées, nous obtenons un problème doublement caché : les sources sont des variables cachées pour l'estimation de la matrice de mélange et les étiquettes sont des variables cachées pour l'estimation des paramètres des distributions des sources. Nous étudions le cas des sources modélisées par des chaînes de Markov cachées (les étiquettes forment une chaîne de Markov) en implémentant l'algorithme EM exact. Nous avons proposé et implémenté des versions sous optimales pour accélérer l'EM [Snoussi et Mohammad-Djafari, 2002a]. La représentation hiérarchique des sources par l'introduction des variables cachées peut être interprétée comme une exploitation de la non stationnarité des variances (voir paragraphe I.2.3) avec une partition automatique de l'intervalle $\mathcal{I} = [1\ T]$ en K sous intervalles avec K le nombre des étiquettes vectorielles des gaussiennes. Le modèle de Markov pour les étiquettes peut être considéré comme une régularisation de cette classification.

Dans le chapitre (IV), nous exploitons la non stationnarité des variances pour séparer des images mélangées. En effet, on rencontre souvent des images homogènes par morceaux et donc qui se prêtent bien à une modélisation par mélange de gaussiennes[13]. Cependant, la classification nécessite une régularisation qui tient compte de l'homogénéité spatiale des images. Cette régularisation peut être effectuée en incorporant un modèle de champ de Markov pour les étiquettes cachées. Nous présentons une implémentation du type MCMC (Monte Carlo par Chaînes de Markov) permettant d'estimer conjointement la matrice de mélange, les sources et leurs ségmentations. Les résultats sont testés sur des images synthétiques (champs cachés de Potts) et sur des images satellitaires [Snoussi et Mohammad-Djafari, 2002c; **?**].

Dans le chapitre (V), nous exploitons la non stationnarité fréquentielle. En effet, avec une approximation circulante, les coefficients de la transformée de Fourier d'un processus gaussien stationnaire sont décorrélés avec une variance (spectre) qui dépend de la fréquence. Donc en utilisant le maximum de vraisemblance, le critère devient une somme pondérée des divergences de Kullback-Leibler entre des matrices spectrales. Nous étudions le cas où les spectres des sources sont connues *a priori* et le cas où on les estime en découpant le domaine de Fourier en anneaux (l'équivalent d'intervalles dans le paragraphe I.2.3). La minimisation de ce critère est implémentée avec l'algorithme EM et accélérée autour de la solution avec un algorithme de gradient conjugué. Nous avons appliqué cette méthode pour séparer des composantes astrophysiques en l'implémentant dans le domaine de Fourier et dans le domaine des harmoniques sphériques [Snoussi *et al.*, 2001; Cardoso *et al.*, 2002; Patanchon *et al.*, 2003]. Ce travail fait l'objet d'une collaboration avec le laboratoire *IN2P3* du Collège de France.

[12] On désigne par hyperparamètres les paramètres qui ne font pas partie de l'ensemble des paramètres d'intérêt.

[13] C'est d'ailleurs le but d'un traitement avancé des images où on modélise l'image par un mélange de gaussiennes afin de la ségmenter.

Le chapitre (VI) traîte en détail le problème de dégénérescence du maximum de vraisemblance. Nous avons généralisé des résultats obtenus dans le cas d'une modélisation par mélange de gaussiennes de signaux monovariables au cas de signaux multivariables. La dégénérescence est produite quand les matrices de covariance approchent des matrices singulières (frontière de singularité). L'élimination de cette dégénérescence est garantie par l'utilisation d'un *a priori* inverse wishart sur les matrices de covariance sans compliquer les équations de ré-estimation de l'algorithme EM. On montre que cette dégénérescence est aussi produite en séparation de sources quand les sources sont modélisées par un mélange de gaussiennes (ou en général par un modèle de Markov caché). La pénalisation par un *a priori* inverse Wishart élimine également cette dégénérescence [Snoussi et Mohammad-Djafari, 2001].

Le chapitre (VII) est consacré au problème de la sélection de la loi *a priori* dans un contexte bayésien. Nous présentons une approche originale [Snoussi et Mohammad-Djafari, 2002b] basée sur la théorie de la prédiction bayésienne [Zhu et Rohwer, 1995] en utilisant les outils de la géométrie de l'information [Amari et Nagaoka, 2000]. On montre l'importance du choix de la géométrie dans l'espace des distributions de probabilité. La règle de Bayes permet de définir la masse par la loi *a posteriori*. Une fois la géométrie et la masse fixées, on construit un critère variationnel dont la minimisation donne la loi *a priori* qu'on a notée δ-*a priori*. Avec les outils de la géométrie différentielle, on introduit la notion d'*a priori* projeté pour les familles paramétriques. Ce travail est appliqué au mélange de familles δ-plates comme le mélange de familles exponentielles (0-plates) et en séparation de sources.

Le dernier chapitre (VIII) ouvre quelques perspectives comme la séparation des images en utilisant une ségmentation par ensembles de niveau. Au lieu de classifier les pixels en utilisant les étiquettes discrètes modélisées par un champ de Markov, on fait évoluer au cours des itérations un contour délimitant les régions homogènes. On va aussi revenir sur la logique des questions comme un espace dual de la logique des propositions en essayant de l'appliquer pour séparer et ségmenter simultanément des images mélangées dans le cadre de la théorie de l'information.

Bibliographie

[Amari et Nagaoka, 2000] S. Amari et H. Nagaoka. *Methods of Information Geometry*, volume 191 of Translations of Mathematical Monographs. AMS, OXFORD, University Press, 2000.

[Amari et Cardoso, 1997] S.-I. Amari et J.-F. Cardoso. Blind source separation — semiparametric statistical approach. *IEEE Trans. Signal Processing*, 45 (11) : 2692–2700, novembre 1997.

[Amari *et al.*, 1996] S.-I. Amari, A. Cichocki et H. Yang. A new learning algorithm for blind source separation. In *Advances in Neural Information Processing Systems 8*, pages 757–763. MIT Press, 1996.

[Ans *et al.*, 1985] B. Ans, J. Hérault et C. Jutten. Adaptive neural architectures : detection of primitives. In *Proc. of COGNITIVA'85*, pages 593–597, Paris, France, 1985.

[Bell et Sejnowski, 1995] A. J. Bell et T. J. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7 (6) : 1129–1159, 1995.

[Belouchrani *et al.*, 1997] A. Belouchrani, K. Abed Meraim, J.-F. Cardoso et Éric Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. Signal Processing*, 45 (2) : 434–44, février 1997.

[Belouchrani et Amin, 1997] A. Belouchrani et M. Amin. Blind source separation using time-frequency distributions : algorithm and asymptotic performance. In *Proc. ICASSP*, pages 3469 – 3472, Munchen, 1997.

[Belouchrani et Cardoso, 1995] A. Belouchrani et J.-F. Cardoso. Maximum likelihood source separation by the expectation-maximization technique : deterministic and stochastic implementation. In *Proc. NOLTA*, 1995.

[Bermond, 2000] O. Bermond. *Méthodes statistiques pour la séparation de sources.* thèse de doctorat, Ecole Nationale Supérieure des Télécommunications, 2000.

[Cardoso et Labeld, 1996] J. Cardoso et B. Labeld. Equivariant adaptative source separation. *Signal Processing*, 44 : 3017–3030, 1996.

[Cardoso *et al.*, 2002] J. Cardoso, H. Snoussi, J. Delabrouille et G. Patanchon. Blind separation of noisy gaussian stationary sources. application to cosmic microwave background imaging. In *Eusipco*, Toulouse, septembre 2002.

[Cardoso, 1989] J.-F. Cardoso. Source separation using higher order moments. In *Proc. ICASSP*, pages 2109–2112, 1989.

[Cardoso, 1997] J. F. Cardoso. Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing*, 4 : 112–114, avril 1997.

[Cardoso et Souloumiac, 1993] J.-F. Cardoso et A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140 (6) : 362–370, décembre 1993.

[Choi et Cichocki, 2000] S. Choi et A. Cichocki. Blind separation of nonstationary sources in noisy mixtures. *Electronics Letters*, 36(9) : 848–849, apr 2000.

[Cichocki et Moszczynski, 1992] A. Cichocki et L. Moszczynski. A new learning algorithm for blind separation of sources. *Electronics Letters*, 28(21) : 1986–1987, 1992.

[Cichocki *et al.*, 1994] A. Cichocki, R. Unbehauen et E. Rummert. Robust learning algorithm for blind separation of signals. *Electronics Letters*, 30(17) : 1386–1387, 1994.

[Cichocki et Unbehaunen, 1996] A. Cichocki et R. Unbehaunen. Robust neural networks with on-line learning for blind identification and blind separation of sources. *IEEE Trans. on Circuits and Systems*, 43 (11) : 894–906, 1996.

[Comon, 1994] P. Comon. Independent Component Analysis, a new concept ? *Signal processing, Special issue on Higher-Order Statistics, Elsevier*, 36 (3) : 287–314, avril 1994.

[Darmois, 1953] G. Darmois. Analyse Générale des Liaisons Stochastiques. *Rev. Inst. Internat. Stat.*, 21 : 2–8, 1953.

[Delfosse et Loubaton, 1995] N. Delfosse et P. Loubaton. Adaptive blind separation of independent sources : a deflation approach. *Signal Processing*, 45 : 59–83, 1995.

[Fry, 2001] R. Fry. The engineering of cybernetic systems. In R. L. Fry, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 497–528. MaxEnt Workshops, Amer. Inst. Physics, août 2001.

[Gaeta et Lacoume, 1990] M. Gaeta et J.-L. Lacoume. Source separation without prior knowledge : the maximum likelihood solution. In *Proc. EUSIPCO'90*, pages 621–624, 1990.

[Hérault, 1985] J. Hérault. Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé. In *Actes 10e coll. GRETSI*, pages 1017–1022, Nice, France, 1985.

[Hérault et Ans, 1984] J. Hérault et B. Ans. Circuits neuronaux à synapses modifiables : décodage de messages composites par apprentissage non supervisé. *C. R. de l'Académie des Sciences*, 299 (III-13) : 525–528, 1984.

[Hunt, 1971] B. R. Hunt. A matrix theory proof of the discrete convolution theorem. *IEEE Trans. Automat. Contr.*, AC-19 : 285–288, 1971.

[Hyvärinen, 1999] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3) : 626–634, 1999.

[Hyvärinen et Oja, 1997] A. Hyvärinen et E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7) : 1483–1492, 1997.

[Jutten, 2000] C. Jutten. Source separation : from dusk till dawn. In *Proc. of 2nd Int. Workshop on Independent Component Analysis and Blind Source Separation (ICA'2000)*, pages 15–26, Helsinki, Finland, 2000.

[Jutten et Herault, 1991] C. Jutten et J. Herault. Blind separation of sources .1. an adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24 (1) : 1–10, 1991.

[Knuth, 1999] K. Knuth. A Bayesian approach to source separation. In *Proceedings of Independent Component Analysis Workshop*, pages 283–288, 1999.

[Knuth, 2000] K. Knuth. Source separation as an exercise in logical induction. In A. Mohammad-Djafari, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 340–349, Gif-sur-Yvette, juillet 2000. Proc. of MaxEnt, Amer. Inst. Physics.

[Knuth, 2001] K. Knuth. Inductive logic : From experimental design to data analysis. In R. L. Fry, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 392–404. MaxEnt Workshops, Amer. Inst. Physics, août 2001.

[Knuth, 2002] K. Knuth. What is a question ? In C. J. Williams, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 227–242, Moscow, Idaho, août 2002. MaxEnt Workshops, Amer. Inst. Physics.

[Malouche et Macchi, 1998] Z. Malouche et O. Macchi. Adaptive unsupervised extraction of one component of a linear mixture with a single neuron. *IEEE Trans. on Neural Networks*, 9(1) : 123–138, 1998.

[Matsuoka *et al.*, 1995] K. Matsuoka, M. Ohya et M. Kawamoto. A neural net for blind separation of nonstationary sources. *Neural Networks*, 8(3) : 411–419, 1995.

[Mohammad-Djafari, 1999] A. Mohammad-Djafari. A Bayesian approach to source separation. In J. R. G. Erikson et C. Smith, éditeurs, *Bayesian Inference and Maximum Entropy Methods*, Boise, IH, USA, juillet 1999. MaxEnt Workshops, Amer. Inst. Physics.

[Moulines *et al.*, 1997] E. Moulines, J. Cardoso et E. Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *ICASSP-97*, Munich, Allemagne, avril 1997.

[Patanchon *et al.*, 2003] G. Patanchon, H. Snoussi, J. Cardoso et J. Delabrouille. Component separation for cosmic microwave background data : a blind approach based on spectral diversity. In *PSIP*, Grenoble, janvier 2003.

[Pham, 1996] D.-T. Pham. Blind separation of instantaneous mixture sources via independent component analysis. *IEEE Trans. Signal Processing*, 44, 1996.

[Pham et Cardoso, 2001] D.-T. Pham et J. Cardoso. Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. Signal Processing*, 49, 9 (11) : 1837–1848, 2001.

[Pham *et al.*, 1992] D.-T. Pham, P. Garrat et C. Jutten. Separation of a mixture of independent sources through a maximum likelihood approach. In *Proc. EUSIPCO'92*, pages 771–774, 1992.

[Rahbar et Reilly, 2001] K. Rahbar et J. Reilly. Blind source separation of convolved sources by joint approximate diagonalization of cross-spectral density matrices. In *Proc. ICASSP*, 2001.

[Rowe, 1998] D. Rowe. *Correlated Bayesian Factor analysis.* thèse de doctorat, Department of Statistics, University of California, Riverside, 1998.

[Senecal, 2000] P. Senecal, S.Amblard. MCMC methods for discrete source separation. In *Bayesian Inference and Maximum Entropy Methods*, pages 350–360, Gif-sur-Yvette, juillet 2000. Proc. of MaxEnt, Amer. Inst. Physics.

[Senecal, 2002] S. Senecal. *Méthodes de simulation Monte-Carlo par chaînes de Markov pour l'estimation de modèles. Applications en séparation de sources et en égalisation.* thèse de doctorat, INPG (Grenoble), 2002.

[Snoussi et Mohammad-Djafari, 2001] H. Snoussi et A. Mohammad-Djafari. Penalized maximum likelihood for multivariate gaussian mixture. In R. L. Fry, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 36–46. MaxEnt Workshops, Amer. Inst. Physics, août 2001.

[Snoussi et Mohammad-Djafari, 2002a] H. Snoussi et A. Mohammad-Djafari. Bayesian unsupervised learning for source separation with mixture of gaussians prior. *To appear in Int. Journal of VLSI Signal Processing Systems*, 2002.

[Snoussi et Mohammad-Djafari, 2002b] H. Snoussi et A. Mohammad-Djafari. Information Geometry and Prior Selection. In C. Williams, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 307–327. MaxEnt Workshops, Amer. Inst. Physics, août 2002.

[Snoussi et Mohammad-Djafari, 2002c] H. Snoussi et A. Mohammad-Djafari. MCMC Joint Separation and Segmentation of Hidden Markov Fields. In *Neural Networks for Signal Processing XII*, pages 485–494. IEEE workshop, septembre 2002.

[Snoussi *et al.*, 2001] H. Snoussi, G. Patanchon, J. Macías-Pérez, A. Mohammad-Djafari et J. Delabrouille. Bayesian blind component separation for cosmic microwave background observations. In R. L. Fry, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 125–140. MaxEnt Workshops, Amer. Inst. Physics, août 2001.

[Souloumiac, 1995] A. Souloumiac. Blind source detection and separation using second order nonstationarity. In *Proc. ICASSP*, pages 1912–1915, 1995.

[Tikhonov et Arsenin, 1977] A. Tikhonov et V. Arsenin. *Solutions of Ill-Posed Problems*. Winston, Washington, DC, USA, 1977.

[Zhu et Rohwer, 1995] H. Zhu et R. Rohwer. Bayesian invariant measurements of generalisation. In *Neural Proc. Lett.*, volume 2 (6), pages 28–31, 1995.

Chapitre II

APPROCHE BAYÉSIENNE EN SÉPARATION DE SOURCES

Thomas Bayes 1702 – 1761

Dans ce chapitre, on introduit la méthode bayésienne comme une alternative à l'approche classique fréquentiste en statistiques. Les probabilités sont présentées comme une extension de la logique des propositions prouvant ainsi leur auto-consistance. La règle de Bayes n'est qu'une simple conséquence des règles que vérifient le calcul des probabilités.
On évoque le problème du choix des lois de probabilités qui a donné lieu à deux points de vue : les probabilités subjectives et les probabilités logiques.
On traîte aussi les techniques de calcul qui permettent la mise en oeuvre de la méthodologie bayésienne. Parmi ces techniques, on décrit l'algorithme EM (*Expectation-Maximization*) et les méthodes MCMC (Monte Carlo par Chaînes de Markov).
On illustre l'application de la théorie bayésienne sur le plan méthodologique et technique en considérant le problème de séparation de sources.

II.1 Inférence logique

Supposons qu'à l'issue d'une expérience $\mathcal{E}$, on récupère T données $\boldsymbol{x}_{1..T}$. La démarche scientifique consiste à affirmer l'existence d'un processus (physique) générant ces données. Si on répète l'expérience $\mathcal{E}$, dans strictement les mêmes conditions, on obtient les mêmes données $\boldsymbol{x}_{1..T}$. De plus, nous supposons, dans ce travail, que ce processus est une composition de transformations de signaux d'origine $\boldsymbol{s}_{1..T}$ et qu'il peut être modélisé mathématiquement par une relation (fonction $\mathcal{F}$) de cause (signaux d'origine) à effet (observations). Autrement dit, les données observées $\boldsymbol{x}_{1..T}$ et les signaux d'origine $\boldsymbol{s}_{1..T}$ sont respectivement les sorties et les entrées d'un système modélisé par la relation $\mathcal{F}$:

$$\boldsymbol{x}_{1..T} = \mathcal{F}(\boldsymbol{s}_{1..T}).$$

La distinction entre les entrées et les sorties d'un système est parfois ambiguë. Cette ambiguïté est en général soulevée en considérant le sens entrées-sorties le sens dans lequel on perd de l'information. Le système n'est alors autre qu'un opérateur de projection. En outre, cette classification des variables en entrées et sorties n'est pas pertinente. Nous préférons la répartition des variables du problème en trois classes :

1. les observations $\boldsymbol{x}_{1..T}$,
2. les grandeurs recherchées $\boldsymbol{\theta}$,
3. et toutes les informations $\boldsymbol{I}$ qu'on a *a priori* sur le problème.

L'information recherchée $\boldsymbol{\theta}$ peut représenter les sources d'origine $\boldsymbol{s}_{1..T}$ qu'on veut reconstruire et/ou un paramètre d'une famille de transformations $f_{\boldsymbol{\theta}}$ liant les entrées et les sorties du système $\mathcal{F}$. Nous ne faisons pas de distinction entre des paramètres d'intérêt et des paramètres de nuisance. Tous les paramètres considérés sont des paramètres d'intérêt. L'introduction des paramètres de nuisance est un artifice pour faciliter des traitements d'inférence ou de prédiction. Un traitement optimal profite de la présence de tels paramètres en assurant que le résultat final n'en dépend pas.

$\boldsymbol{I}$ représente toutes les informations *a priori* qu'on possède sur le problème. On distingue $\boldsymbol{I}$ des données $\boldsymbol{x}_{1..T}$. En effet, $\boldsymbol{I}$ inclut des données de nature différente de celle des données observées $\boldsymbol{x}_{1..T}$. Elle représente par exemple des données qualitatives et quantitatives sur le système $\mathcal{F}$ comme la nature de la famille paramétrique $f_{\boldsymbol{\theta}}$ (modélisant $\mathcal{F}$), la présence d'un bruit additif...

On peut définir l'inférence comme la recherche de l'information $\boldsymbol{\theta}$ à partir de la connaissance de $\boldsymbol{x}_{1..T}$ et de $\boldsymbol{I}$. On peut distinguer trois sortes d'inférence.

1. Si $\boldsymbol{\theta}$ représente les signaux d'entrée alors l'inférence est une **reconstruction**.
2. Si $\boldsymbol{\theta}$ représente les paramètres du système $\mathcal{F}$ alors l'inférence est une **identification**.
3. Si on cherche à prédire le vecteur $\boldsymbol{x}_{T+1}$ à partir des données observées $\boldsymbol{x}_{1..T}$, l'inférence est une **prédiction**. La prédiction se distingue de l'identification dans le sens où son objectif n'est pas l'estimation de $\boldsymbol{\theta}$ mais plutôt l'estimation de la densité de probabilité $p(\boldsymbol{x}_{T+1} \mid \boldsymbol{x}_{1..T})$. Cette densité peut ne pas appartenir à la famille $\{f_{\boldsymbol{\theta}}\}$ et la prédiction ne fournit pas ainsi un point $\hat{\boldsymbol{\theta}}$. Cependant, la famille $\{f_{\boldsymbol{\theta}}\}$ détermine complètement la géométrie du problème.

L'inférence peut prendre un caractère logique en considèrant les variables du problème étudié comme des propositions :
- $\boldsymbol{x}_{1..T}$ est la proposition : << les données observées sont $\boldsymbol{x}_{1..T}$! >>
- $\boldsymbol{\theta}$ est la proposition : << les paramètres recherchés sont $\boldsymbol{\theta}$! >>
- $\boldsymbol{I}$ est la proposition : << l'information *a priori* est $\boldsymbol{I}$! >>

Dans le cas parfait, la valeur de la proposition $\mathcal{I} = (\boldsymbol{x}_{1..T} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{\theta})$ peut être construite avec les règles de la logique[1] et vaudra 0 ou 1. Si $\boldsymbol{x}_{1..T}$, $\boldsymbol{I}$ et $\boldsymbol{\theta}$ varient dans leurs espaces respectifs $\mathcal{X}$, $\mathcal{H}$ et Θ, il y aura autant de propositions que d'éléments dans $\mathcal{X} \cup \mathcal{H} \cup \Theta$. En fixant les valeurs de $\boldsymbol{x}_{1..T}$ et $\boldsymbol{I}$, la solution du problème

[1] L'information et les lois de la physique sont contenues dans la proposition $\boldsymbol{I}$.

d'inférence est la valeur $\hat{\boldsymbol{\theta}}$ telle que la proposition $\mathcal{I} = (\boldsymbol{x} \wedge \boldsymbol{I} \longrightarrow \hat{\boldsymbol{\theta}})$ soit vraie (vaut 1). Malheureusement, la proposition $\mathcal{I} = (\boldsymbol{x}_{1..T} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{\theta})$ ne peut pas être évaluée avec exactitude (on ne peut pas affirmer si elle est vraie ou fausse). Ceci est dû à, au moins, trois raisons.

1. L'information *a priori* $\boldsymbol{I}$ ne renseigne pas suffisamment sur la physique du problème.
2. Les données $\boldsymbol{x}_{1..T}$ ne contiennent pas suffisamment d'informations sur le paramètre $\boldsymbol{\theta}$. Le problème est sous-déterminé.
3. La physique du problème est très compliquée. L'évaluation des propositions $\mathcal{I} = (\boldsymbol{x} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{\theta})$ est très complexe.

Cependant, on veut parfois exprimer une certaine incertitude sur la proposition $\mathcal{I}$. Les probabilités représentent une mesure de cette incertitude consistante avec les règles de la logique (voir les travaux de Cox pour la dérivation des relations que doivent vérifier les probabilités [Cox, 1946, 1961, 1979]). Le problème d'inférence (ou de prédiction) est alors complètement décrit par la fonction $\mathrm{P}r(\boldsymbol{x} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{\theta})$. Autrement dit, la quantité $\mathrm{P}r(\boldsymbol{x} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{\theta})$ contient toute l'information disponible pour une inférence sur $\boldsymbol{\theta}$. Nous notons que la manipulation des probabilités sur les propositions ne fait pas la distinction entre tout ce qui est connu de tout ce qui n'est pas connu. $\boldsymbol{x}_{1..T}$, $\boldsymbol{I}$ et $\boldsymbol{\theta}$ représentent trois propositions et on veut mesurer le degré d'implication entre les différentes combinaisons de ces propositions en respectant les règles de calcul des probabilités.

Remarque 4 *La définition des probabilités comme une mesure d'implication entre deux propositions montre que la notion de la probabilité d'une proposition* $\mathrm{P}r(A)$ *n'existe pas. Cette notion existe au sens de la théorie fréquentiste où* A *n'est pas une proposition mais plutôt un événement et* $\mathrm{P}r(A)$ *est la fréquence de cet événement dans une infinité de réalisations. Cependant, parfois dans la littérature de l'inférence logique, on trouve la notation* $\mathrm{P}r(A)$ *mais rigoureusement ceci représente* $\mathrm{P}r(\boldsymbol{I} \longrightarrow A)$ *où* $\boldsymbol{I}$ *est toute l'information a priori qu'on possède.*

II.2 Règle de Bayes

La règle de Bayes est une conséquence de la consistance des probabilités avec l'algèbre booléenne. En effet, en appliquant la règle de produit (conséquence de l'associativité) :

$$\mathrm{P}r(\boldsymbol{I} \longrightarrow \boldsymbol{\theta} \wedge \boldsymbol{x}_{1..T}) = \mathrm{P}r(\boldsymbol{x}_{1..T} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{\theta})\,\mathrm{P}r(\boldsymbol{I} \longrightarrow \boldsymbol{x}_{1..T}),$$

avec la règle de la commutativité entre $\boldsymbol{\theta}$ et $\boldsymbol{x}_{1..T}$, on obtient le théorème de Bayes :

$$\mathrm{P}r(\boldsymbol{x}_{1..T} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{\theta}) = \frac{\mathrm{P}r(\boldsymbol{\theta} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{x}_{1..T})\,\mathrm{P}r(\boldsymbol{I} \longrightarrow \boldsymbol{\theta})}{\mathrm{P}r(\boldsymbol{I} \longrightarrow \boldsymbol{x}_{1..T})}. \tag{II.1}$$

L'incertitude sur la proposition de l'inférence $\mathcal{I} =: (\boldsymbol{x}_{1..T} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{\theta})$ est ainsi exprimée d'une manière simple en fonction des incertitudes d'autres propositions qui sont mieux abordables par le physicien.

Dans le cas où les grandeurs manipulées sont continues, la proposition $(B \longrightarrow A)$ est transformée en $B \longrightarrow A \in \mathcal{V}(A)$ où $\mathcal{V}(A)$ est un voisinage de la variable continue A et l'incertitude est mesurée par $\mathrm{d\,P}(B \longrightarrow A) = \mathrm{P}r(B \longrightarrow A \in \mathcal{V}(A))$. En changeant les notations $\mathrm{d\,P}(B \longrightarrow A)$ par $\mathrm{d\,P}(A \mid B)$ et la conjonction $(A \wedge B)$ par $(A\,,\,B)$, le théorème de Bayes (II.1) s'écrit :

$$\mathrm{d\,P}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{I}) = \frac{\mathrm{d\,P}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}, \boldsymbol{I})\ \mathrm{d\,P}(\boldsymbol{\theta} \mid \boldsymbol{I})}{\mathrm{d\,P}(\boldsymbol{x}_{1..T} \mid \boldsymbol{I})}. \tag{II.2}$$

Si chaque distribution de probabilité possède une densité p par rapport à une mesure $\boldsymbol{\mu}$ de l'espace correspondant, on peut ré-écrire (II.2) :

$$\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{I}) = \frac{\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}, \boldsymbol{I})\,\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{I})}{\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{I})}. \tag{II.3}$$

$\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{I})$ est la densité *a priori* de $\boldsymbol{\theta}$ et $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}, \boldsymbol{I})$ est la vraisemblance de $\boldsymbol{\theta}$. La règle de Bayes peut être interprétée comme la combinaison logique de ces deux sources d'informations pour donner l'information

a posteriori (la densité *a posteriori*). C'est l'une des raisons principales de l'intérêt qu'a suscité l'approche bayésienne pour résoudre les problèmes d'inférence. On arrive à injecter de l'information *a priori* dans un cadre probabiliste consistant avec le raisonnement logique. Le terme $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{I})$ est l'évidence des données. Il peut être interprété comme un coefficient de normalisation en imposant que $\int \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{I}) d\,\boldsymbol{\theta} = 1$ [2].

La méthode bayésienne se distingue de l'approche classique fréquentiste d'un point de vue fondamental et méthodologique :

1. Au niveau fondamental :
   Dans l'approche fréquentiste, $\boldsymbol{x}_{1..T}$ sont des variables aléatoires et les observations acquises représentent une réalisation particulière de ce phénomène aléatoire. Les probabilités prennent alors le sens d'une fréquence d'un événement dans une infinité de réalisations. Quant à $\boldsymbol{\theta}$, il est considéré comme un paramètre fixe mais inconnu et on ne peut pas parler de probabilité de $\boldsymbol{\theta}$. Par contre, dans l'approche bayésienne, les données $\boldsymbol{x}_{1..T}$ et le paramètre $\boldsymbol{\theta}$ sont manipulés de la même façon en tant que propositions. Le fait de parler d'événement parmi beaucoup de réalisations souvent n'a pas de sens. Les probabilités représentent alors le degré d'incertitude des implications entre les différentes propositions. Le vrai mérite de l'approche bayésienne est de s'opposer à l'approche fréquentiste en évitant des arguments métaphysiques. En effet, quand les fréquentistes manipulent $\boldsymbol{\theta}$ comme un paramètre fixe, les bayésiens ne s'y opposent pas en prétendant que $\boldsymbol{\theta}$ est de nature aléatoire [3] mais ils approchent le problème d'une façon constructive basée sur le raisonnement logique. L'introduction des probabilités reflète notre ignorance et la limitation de nos capacités à comprendre tout ce qui passe dans ce monde. Par abus de langage, on qualifie $\boldsymbol{\theta}$ de variable aléatoire mais ceci ne présente aucun jugement sur sa nature aléatoire ou non !

2. Au niveau méthodologique :[4]
   La différence au niveau des fondements des deux approches classique et bayésienne a une conséquence directe sur la méthodologie de l'estimation de $\boldsymbol{\theta}$ à partir des observations $\boldsymbol{x}_{1..T}$. Dans l'approche classique, on cherche à construire des estimateurs $\hat{\boldsymbol{\theta}}(\boldsymbol{x}_{1..T})$ et à comparer leurs performances (biais et variance) en les considérant comme des variables aléatoires (puisque $\boldsymbol{x}_{1..T}$ le sont). Par contre, dans l'approche bayésienne, on considère que toute l'information est contenue dans la distribution *a posteriori* $\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{I})$ et que toute inférence doit être basée sur cette distribution. Une fois observées, les données $\boldsymbol{x}_{1..T}$ cessent d'être aléatoires et la seule variable est le paramètre $\boldsymbol{\theta}$. Selon le contexte du problème qu'on traîte, on choisit un coût $\mathrm{C}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ et l'estimateur $\hat{\boldsymbol{\theta}}$ est le minimiseur de l'espérance *a posteriori* de ce coût :
   $$\hat{\boldsymbol{\theta}} = \arg\min \int \mathrm{C}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \mathrm{p}(\boldsymbol{\theta}^* \mid \boldsymbol{x}_{1..T}, \boldsymbol{I}) d\,\boldsymbol{\theta}^*,$$
   ce qui revient à considérer une caractéristique particulière de la distribution *a posteriori* comme la moyenne, la médiane, le mode,...

II.3 Choix de la loi *a priori* ou choix des probabilités ?

L'une des reproches à la théorie bayésienne est le choix de la densité *a priori* $\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{I})$. Autrement dit, ayant l'information *a priori* $\boldsymbol{I}$, quelle est le degré de plausibilité de la proposition $\boldsymbol{I} \longrightarrow \boldsymbol{\theta}$?. Nous notons que cette question fait partie d'un problème plus général lors de l'introduction des probabilités dans la section précédente et qu'elle n'est pas spécifique à la densité *a priori* ou à l'application de la règle de Bayes (II.3). On doit aussi se poser cette même question à propos de la vraisemblance $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}, \boldsymbol{I})$.

En remontant à l'introduction de la notion de probabilité, on l'a définie comme une mesure de l'incertitude d'une implication entre deux propositions. Ainsi $\mathrm{P}r(A \mid B)$ est l'incertitude de la proposition $B \longrightarrow A$ et la question qui se pose est quelle est la valeur de cette incertitude ?

[2] Le fait que la somme totale $\int p$ soit égale à 1 est une conséquence de la consistance logique des probabilités

[3] ce qui va aboutir à une discussion sur la nature de ce monde.

[4] L'objectif de ce paragraphe n'est pas de montrer l'avantage de la méthode bayésienne sur l'approche classique mais d'expliquer leur divergence au niveau méthodologique.

Cette question a partagé les scientifiques entre deux théories : les probabilités subjectives et les probabilités logiques :

1. **Probabilités subjectives** :
   Le choix de la valeur de $\mathrm{P}r(A \mid B)$ est propre à l'utilisateur. A travers ce choix, ce dernier exprime sa propre incertitude qui reflète sa connaissance du problème. Les probabilités prennent ainsi un caractère subjectif ou personnel. Deux personnes différentes ayant les mêmes informations B peuvent attribuer des probabilités différentes. Cependant, cette théorie n'est pas déconnectée de la démarche scientifique. L'utilisateur doit respecter la consistance du calcul des probabilités. Par exemple, si $\mathrm{P}r(A \mid B)$ est fixée alors $\mathrm{P}r(\overline{A} \mid B)$ ne peut pas être fixée librement et doit être égale à $1 - \mathrm{P}r(A \mid B)$.
2. **Probabilités logiques** :
   D'après cette théorie, les probabilités représentent une extension de la logique. Deux personnes ayant les mêmes connaissances doivent attribuer les mêmes probabilités aux quantités manipulées dans le problème étudié. Ce qui suppose qu'il existe des règles universelles de choix de lois de probabilités.

L'approche logique du choix des probabilités est théoriquement bien fondée quoique difficile à mettre en oeuvre et reste ainsi à un stade idéal. L'approche subjective est par contre souvent adoptée dans les situations pratiques.

En revenant à notre problème d'inférence, il s'agit de choisir les probabilités $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}, \boldsymbol{I})$ et $\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{I})$. On note que dans les deux cas, on traîte le même type de problème. La vraisemblance modélise le processus qui a généré les observations $\boldsymbol{x}_{1..T}$ et l'*a priori* modélise le processus (éventuellement virtuel) qui a généré le paramètre $\boldsymbol{\theta}$. En pratique, Le choix est fait par des considérations subjectives (le modèle gaussien pour le bruit en est un exemple). Cependant, si on ne possède pas d'informations *a priori* sur le paramètre $\boldsymbol{\theta}$, l'approche logique des probabilités peut être abordée. En effet, on se trouve dans un cas commun à beaucoup de problèmes qu'on qualifie d'**ignorance**. On peut alors essayer de trouver des règles de calcul de probabilités pour exprimer cette ignorance [Kass et Wasserman, 1994; Rodríguez, 1991; Snoussi et Mohammad-Djafari, 2002].

II.4 Structure hiérarchique

Dans certains cas, notre objectif n'est pas l'inférence de tout le vecteur $\boldsymbol{\theta}$ mais seulement d'un sous-vecteur $\boldsymbol{\theta}_I$. La distribution *a posteriori* de $\boldsymbol{\theta}_I$ est obtenue avec la règle de Bayes :

$$\mathrm{p}(\boldsymbol{\theta}_I \mid \boldsymbol{x}_{1..T}, \boldsymbol{I}) = \frac{\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}_I, \boldsymbol{I})\, \mathrm{p}(\boldsymbol{\theta}_I \mid \boldsymbol{I})}{\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{I})} \tag{II.4}$$

Le problème $(\boldsymbol{\theta}_I \wedge \boldsymbol{I} \longrightarrow \boldsymbol{x}_{1..T})$ est en général plus difficile (et aussi différent) que $(\boldsymbol{\theta} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{x}_{1..T})$ (puisqu'on possède moins d'informations). Redéfinir la vraisemblance posera sans doute des problèmes de cohérence. Cependant, on peut obtenir l'expression de $\mathrm{p}(\boldsymbol{\theta}_I \mid \boldsymbol{x}_{1..T}, \boldsymbol{I})$ par marginalisation de la distribution *a posteriori* de $\boldsymbol{\theta}$ qui est plus simple à obtenir :

$$\begin{cases} \mathrm{p}(\boldsymbol{\theta}_I \mid \boldsymbol{x}_{1..T}, \boldsymbol{I}) = \int_{\boldsymbol{\theta}_{-I}} \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{I}) d\,\boldsymbol{\theta}_{-I} \\ \\ \boldsymbol{\theta} = (\boldsymbol{\theta}_I, \boldsymbol{\theta}_{-I}) \end{cases}$$

L'opération inverse de la marginalisation est l'augmentation de paramètres. On introduit d'autres variables $\boldsymbol{c}$ d'une manière naturelle ou artificielle de telle façon que l'inférence $(\boldsymbol{x}_{1..T} \longrightarrow \boldsymbol{\theta} \wedge \boldsymbol{c})$ soit plus simple à modéliser. D'un point de vue logique, ceci revient à intercaler une proposition $\boldsymbol{c}$ entre l'information *a priori* $\boldsymbol{I}$ et le paramètre $\boldsymbol{\theta}$ pour faciliter le raisonnement logique ou à compléter l'information dans $\boldsymbol{\theta}$ pour expliquer les données $\boldsymbol{x}_{1..T}$:

$$\begin{cases} (\boldsymbol{I} \longrightarrow \boldsymbol{\theta}) \rightsquigarrow (\boldsymbol{I} \longrightarrow \boldsymbol{c} \longrightarrow \boldsymbol{\theta}) \\ \text{et / ou} \\ (\boldsymbol{\theta} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{x}_{1..T}) \rightsquigarrow (\boldsymbol{\theta} \wedge \boldsymbol{I} \wedge \boldsymbol{c} \longrightarrow \boldsymbol{x}_{1..T}) \end{cases}$$

D'un point de vue probabiliste, cette procédure consiste à introduire dans un premier temps des variables cachées $\boldsymbol{c}$ d'une manière à limiter les choix subjectifs des probabilités intervenant dans le problème d'inférence. Dans un deuxième temps, on marginalise par rapport à ces variables pour ne garder que le paramètre d'intérêt $\boldsymbol{\theta}$:

$$\begin{aligned} \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{I}) &\propto \int_{\boldsymbol{c}} \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{c} \mid \boldsymbol{x}_{1..T}, \boldsymbol{I}) d\,\boldsymbol{c} \\ &\propto \frac{\int_{\boldsymbol{c}} \mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{I})\, \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{c}, \boldsymbol{I})\, \mathrm{p}(\boldsymbol{c} \mid \boldsymbol{I})\, d\,\boldsymbol{c}}{\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{I})} \end{aligned} \tag{II.5}$$

On note ici la consistance de la théorie des probabilités avec la logique. L'introduction des variables cachées possède un sens logique et elle est naturellement interprétée en probabilités. Cette procédure peut être infiniment ré-itérée en introduisant d'autres couches de variables cachées :

$$(\boldsymbol{I} \longrightarrow \boldsymbol{\theta}) \rightsquigarrow (\boldsymbol{I} \longrightarrow \boldsymbol{c}_0 \longrightarrow \boldsymbol{\theta}) \rightsquigarrow (\boldsymbol{I} \longrightarrow ... \longrightarrow \boldsymbol{c}_1 \longrightarrow \boldsymbol{c}_0 \longrightarrow \boldsymbol{\theta})$$

Cette structure hiérarchique facilite le choix des probabilités intervenant dans le calcul de la distribution *a posteriori*. Ainsi, au lieu de choisir la vraisemblance $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}, \boldsymbol{I})$, on construit plutôt la fonction $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{I})$ en profitant de l'augmentation de l'information. Le choix de l'*a priori* est aussi facilité par la décomposition logique de $(\boldsymbol{I} \longrightarrow \boldsymbol{\theta})$ en $(\boldsymbol{I} \longrightarrow \boldsymbol{c} \longrightarrow \boldsymbol{\theta})$. On obtient ainsi l'expression de la distribution *a posteriori* (II.5) sous forme intégrale. La question qui se pose est comment mener l'inférence sur le paramètre $\boldsymbol{\theta}$ à l'aide de cette expression. Même si on réussit parfois à obtenir une forme analytique de la distribution *a posteriori*, le calcul de ses caractéristiques (comme le mode, la moyenne, la médiane...) est en général très difficile. Nous allons voir dans la section suivante des méthodes adaptées pour la structure à variables cachées. La première méthode est l'algorithme EM qui permet de calculer le maximum *a posteriori* et la deuxième est l'échantillonnage bayésien qui est une technique plus générale (non réservée aux problèmes à variables cachées) permettant d'approcher numériquement toutes les caractéristiques de la distribution *a posteriori*.

II.5 Quelques techniques de calcul

II.5.1 Algorithme EM

En prenant le coût d'estimation $C(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ suivant :

$$C(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = 1 - \delta_{(\boldsymbol{\theta} - \boldsymbol{\theta}^*)}$$

où δ est la distribution de dirac et $\boldsymbol{\theta}^*$ est la vraie valeur du paramètre recherché $\boldsymbol{\theta}$. L'estimé de $\boldsymbol{\theta}$ est alors le minimiseur de la moyenne *a posteriori* du coût $C(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$:

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \int_{\boldsymbol{\theta}^*} C(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \mathrm{p}(\boldsymbol{\theta}^* \mid \boldsymbol{x}_{1..T}, \boldsymbol{I}) d\,\boldsymbol{\theta}^* \\ &= \arg\max_{\boldsymbol{\theta}} \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{I}) \end{aligned}$$

L'estimation de $\boldsymbol{\theta}$ revient alors à résumer la distribution *a posteriori* par son maximum (MAP). Comme le logarithme est une fonction strictement croissante, la solution MAP est aussi le maximiseur du logarithme de la distribution *a posteriori* [5] :

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \log \int \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{c} \mid \boldsymbol{x}_{1..T}, \boldsymbol{I}) d\,\boldsymbol{c} \\ &= \arg\max_{\boldsymbol{\theta}} \log \int \mathrm{p}(\boldsymbol{x}_{1..T}, \boldsymbol{c} \mid \boldsymbol{\theta}, \boldsymbol{I}) d\,\boldsymbol{c} + \log p(\boldsymbol{\theta} \mid \boldsymbol{I}) \end{aligned} \tag{II.6}$$

où $\mathrm{p}(\boldsymbol{x}_{1..T}, \boldsymbol{c} \mid \boldsymbol{\theta}, \boldsymbol{I})$ est appelée dans la littérature la vraisemblance complétée.

[5] l'introduction du logarithme est justifiée par le fait que souvent on manipule des familles exponentielles

Le calcul et l'optimisation de la vraisemblance $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}, \boldsymbol{I})$ sont en général compliqués du fait de la présence de l'intégration. L'algorithme EM [Dempster *et al.*, 1977] est un algorithme itératif qui permet d'approcher numériquement la solution de (II.6). Le principe de l'EM est la construction d'une suite déterministe $(\boldsymbol{\theta}^{(k)})_{k\in\mathbb{N}}$ qui converge vers le MAP. En partant d'une valeur initiale $\boldsymbol{\theta}^{(0)}$, cette suite est définie par une transformation $\mathcal{M}$:

$$\boldsymbol{\theta}^{(k+1)} = \mathcal{M}(\boldsymbol{\theta}^{(k)})$$

La transformation $\mathcal{M}$ consiste en deux étapes :

1. Etape **E** (Expectation) :
   Dans cette étape, on calcule une fonction auxiliaire $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ espérance *a posteriori* du logarithme de la distribution *a posteriori* jointe de $(\boldsymbol{\theta}, \boldsymbol{c})$:

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) &= \mathrm{E}\big[\log \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{c} \mid \boldsymbol{x}_{1..T}, \boldsymbol{I}) \mid \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^{(k)}\big] \\ &= \mathrm{E}\big[\log \mathrm{p}(\boldsymbol{x}_{1..T}, \boldsymbol{c} \mid \boldsymbol{\theta}, \boldsymbol{I}) \mid \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^{(k)}\big] + \log p(\boldsymbol{\theta} \mid \boldsymbol{I}) + K \end{aligned} \tag{II.7}$$

   où K est une constante indépendante de $\boldsymbol{\theta}$ et l'espérance est calculée par rapport à la variable cachée $\boldsymbol{c}$ conditionnellement aux données $\boldsymbol{x}_{1..T}$ et au paramètre de l'itération précédente $\boldsymbol{\theta}^{(k)}$:

$$\mathrm{E}\big[f(\boldsymbol{c}) \mid \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^{(k)}\big] = \int f(\boldsymbol{c}) \mathrm{p}(\boldsymbol{c} \mid \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^{(k)}) d\,\boldsymbol{c}.$$

2. Etape **M** (Maximization) :
   Dans cette étape, on maximise la fonctionnelle $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ calculée dans l'étape **E** par rapport à $\boldsymbol{\theta}$ pour calculer le terme suivant de la suite numérique :

$$\boldsymbol{\theta}^{(k+1)} = \mathcal{M}(\boldsymbol{\theta}^{(k)}) = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}).$$

La propriété clé de l'algorithme EM est que la distribution *a posteriori* marginalisée $\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{I})$ croît à chaque itération (monotonie) :

$$\mathrm{p}(\boldsymbol{\theta}^{(k+1)} \mid \boldsymbol{x}_{1..T}, \boldsymbol{I}) \;\geq\; \mathrm{p}(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{x}_{1..T}, \boldsymbol{I})$$

Cette propriété est due essentiellement à l'inégalité de Jensen pour les fonctions concaves. En effet,

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)}) - \mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}) &= \mathrm{E}\big[\log \mathrm{p}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{c} \mid \boldsymbol{x}_{1..T})\big] - \mathrm{E}\big[\log \mathrm{p}(\boldsymbol{\theta}^{(k)}, \boldsymbol{c} \mid \boldsymbol{x}_{1..T})\big] \\ &= \int \log \tfrac{\mathrm{p}(\boldsymbol{c} \mid \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^{(k+1)})}{\mathrm{p}(\boldsymbol{c} \mid \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^{(k)})} \mathrm{p}(\boldsymbol{c} \mid \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^{(k)}) d\,\boldsymbol{c} + \log \tfrac{\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}^{(k+1)}) \mathrm{p}(\boldsymbol{\theta}^{(k+1)})}{\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}^{(k)}) \mathrm{p}(\boldsymbol{\theta}^{(k)})} \\ &\leq \log \int \tfrac{\mathrm{p}(\boldsymbol{c} \mid \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^{(k+1)})}{\mathrm{p}(\boldsymbol{c} \mid \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^{(k)})} \mathrm{p}(\boldsymbol{c} \mid \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^{(k)}) d\,\boldsymbol{c} + \log \tfrac{\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}^{(k+1)}) \mathrm{p}(\boldsymbol{\theta}^{(k+1)})}{\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}^{(k)}) \mathrm{p}(\boldsymbol{\theta}^{(k)})} \\ &\leq \log(1) + \log \tfrac{\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}^{(k+1)}) \mathrm{p}(\boldsymbol{\theta}^{(k+1)})}{\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}^{(k)}) \mathrm{p}(\boldsymbol{\theta}^{(k)})} \end{aligned} \tag{II.8}$$

où l'inégalité est due à la concavité de la fonction logarithme. D'après l'inégalité (II.8), la croissance de la distribution *a posteriori* est supérieure ou égale à la croissance de la fonctionnelle $\mathcal{Q}$ [6].

La convergence de l'algorithme EM est liée à la contraction de la transformation $\mathcal{M}$. Dans ce cas, la suite $(\boldsymbol{\theta}^{(k)})_{k\in\mathbb{N}}$ converge vers $\boldsymbol{\theta}$ tel que :

$$\boldsymbol{\theta} = \mathcal{M}(\boldsymbol{\theta}) \tag{II.9}$$

[6] On note qu'on peut se contenter à chaque itération d'une valeur de $\boldsymbol{\theta}^{(k+1)}$ telle que $\mathcal{Q}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)}) \geq \mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)})$ et on obtient ainsi l'algorithme GEM (Generalized EM).

Dans le cas du maximum de vraisemblance ($\mathrm{p}(\boldsymbol{\theta}) \propto cte$), des conditions de continuité [Dempster *et al.*, 1977] de la fonctionnelle $\mathcal{Q}$ garantissent la contraction de la transformation $\mathcal{M}$ dans le voisinage du maximum de vraisemblance $\hat{\boldsymbol{\theta}}$ et que $\hat{\boldsymbol{\theta}}$ est solution de l'équation (II.9). On peut consulter [Wu, 1983; Boyles, 1983; McLachlan et Krishnan, 1997] pour plus de détails sur la convergence de l'EM dans le cas du maximum de vraisemblance.

Ces résultats de convergence ont été étendus pour le cas du maximum *a posteriori* (avec la fonctionnelle $\mathcal{Q}$ définie dans (II.7)) [Hero et Fessler, 1993; Green, 1990].

II.5.2 Techniques du calcul bayésien

La méthode bayésienne ne se limite pas à résumer la densité *a posteriori* $\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{h})$ par son maximum. D'autres caractéristiques de cette densité sont utiles selon le contexte du problème étudié. En général, on sera amené à calculer l'espérance *a posteriori* d'une fonction $h(\boldsymbol{\theta})$:

$$\mathrm{E}\big[h(\boldsymbol{\theta})\big] = \int h(\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{h}) d\,\boldsymbol{\theta} \tag{II.10}$$

En pratique, étant données la vraisemblance $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}, \boldsymbol{h})$ et la densité *a priori* $\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{h})$, on obtient la densité *a posteriori* à une constante près en appliquant la règle de Bayes :

$$\begin{aligned} f(\boldsymbol{\theta}) &= \mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}, \boldsymbol{h})\, \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{h}) \\ &\propto \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{h}) \end{aligned}$$

Le calcul de l'espérance de $h(\boldsymbol{\theta})$ nécessite alors deux intégrations :

$$\mathrm{E}\big[h(\boldsymbol{\theta})\big] = \frac{\int h(\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\,\boldsymbol{\theta}}{\int f(\boldsymbol{\theta}) d\,\boldsymbol{\theta}} \tag{II.11}$$

En général, ce calcul est difficile à mener pour les raisons suivantes.

- La forme de la fonction h est compliquée.
- La forme de la fonction f est compliquée.
- On ne possède pas une forme analytique de f comme c'est le cas dans les problèmes à variables cachées où f est donnée sous forme intégrale (II.5).

On peut alors essayer d'approcher la densité *a posteriori* par des fonctions simples. L'approximation normale est la plus naturelle du fait de sa validité asymptotique [Lindley, 1980; Tierney et Kadane, 1986]. Pour tenir compte de l'asymétrie de f, on peut pousser le développement limité de $\log f(\boldsymbol{\theta})$ à l'ordre 3. On peut aussi envisager des approximations directement sur les quantités à intégrer ($h(\boldsymbol{\theta}) f(\boldsymbol{\theta})$) [Lindley, 1980; Kass *et al.*, 1988; Tierney *et al.*, 1986]. Cependant, ces approximations ne peuvent pas s'adapter à toutes les formes possibles de la fonctions f. Avec le développement des moyens de calcul, les méthodes de Monte Carlo présentent une alternative efficace.

Etant donnés M échantillons $(\tilde{\boldsymbol{\theta}}^{(1)}, ..., \tilde{\boldsymbol{\theta}}^{(M)})$ générés selon la distribution *a posteriori* de $\boldsymbol{\theta}$:

$$\tilde{\boldsymbol{\theta}}^{(m)} \sim \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{h}), \quad m = 1, .., M$$

la moyenne empirique de $h(\boldsymbol{\theta})$:

$$\mathrm{E}^*_M\big[h(\boldsymbol{\theta})\big] = \frac{1}{M} \sum_{m=1}^{M} h(\tilde{\boldsymbol{\theta}}^{(m)})$$

est une variable aléatoire qui converge presque sûrement, selon la loi forte des grands nombres, vers la moyenne théorique (II.10) quand M tend vers l'infini. La variance de la moyenne empirique peut être calculée

à condition que la fonction $h^2(\boldsymbol{\theta})\, f(\boldsymbol{\theta})$ soit intégrable :

$$var(\mathrm{E}_M^*\big[h(\boldsymbol{\theta})\big]) = \frac{1}{M}\left[\int h^2(\boldsymbol{\theta})\, f(\boldsymbol{\theta}) d\,\boldsymbol{\theta} - \big(\mathrm{E}_f\big[h(\boldsymbol{\theta})\big]\big)^2\right].$$

L'échantillonnage exacte de la distribution *a posteriori* n'est pas toujours possible[7]. L'expression (II.10) peut être approximée par une moyenne empirique basée sur un jeu d'échantillons $(\tilde{\boldsymbol{\theta}}^{(1)}, ...\tilde{\boldsymbol{\theta}}^{(M)})$ générés selon une autre distribution g :

$$\sum_1^M h(\tilde{\boldsymbol{\theta}}^{(m)}) \frac{f(\tilde{\boldsymbol{\theta}}^{(m)})}{g(\tilde{\boldsymbol{\theta}}^{(m)})} \Big/ \sum_1^M \frac{f(\tilde{\boldsymbol{\theta}}^{(m)})}{g(\tilde{\boldsymbol{\theta}}^{(m)})} \tag{II.12}$$

C'est ce qu'on appelle l'échantillonnage pondéré. Quand M tend vers l'infini, l'expression (II.12) converge vers la moyenne théorique :

$$\sum_1^M h(\tilde{\boldsymbol{\theta}}^{(m)}) \frac{f(\tilde{\boldsymbol{\theta}}^{(m)})}{g(\tilde{\boldsymbol{\theta}}^{(m)})} \Big/ \sum_1^M \frac{f(\tilde{\boldsymbol{\theta}}^{(m)})}{g(\tilde{\boldsymbol{\theta}}^{(m)})} \xrightarrow{M\to\infty} \int h(\boldsymbol{\theta}) \frac{f(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} g(\boldsymbol{\theta}) d\,\boldsymbol{\theta} \Big/ \int \frac{f(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} g(\boldsymbol{\theta}) d\,\boldsymbol{\theta}$$

$$\xrightarrow{M\to\infty} \mathrm{E}_f\big[h(\boldsymbol{\theta})\big]$$

Comme dans le cas de l'échantillonnage direct, l'estimateur (II.12) possède une variance finie si la quantité :

$$\mathrm{E}_g\big[h^2(\boldsymbol{\theta}) \left[\frac{f(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}\right]^2\big] = \mathrm{E}_f\big[h^2(\boldsymbol{\theta}) \frac{f(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}\big]$$

est finie. Cette condition limite le choix de la distribution instrumentale g. D'une manière qualitative, on doit choisir une distribution g à queues plus épaisses que celles de f. On trouve dans [Geweke, 1989] des conditions suffisantes sur g pour garantir une variance finie de l'estimateur de l'échantillonnage pondéré.

Les performances de l'échantillonnage pondéré sont directement liées à la similarité entre la distribution instrumentale g et la distribution *a posteriori* $\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{h})$. Autrement dit, l'approximation de la moyenne théorique par une moyenne empirique est optimale quand on échantillonne selon la loi *a posteriori* [8]. Les techniques MCMC (Monte Carlo par Chaînes de Markov) permettent de générer des échantillons suivant la distribution *a posteriori* (simulation par chaînes de Markov) et de garantir l'application des méthodes de Monte Carlo sur les échantillons obtenus. On évite ainsi le recours à l'échantillonnage pondéré. Afin d'introduire les méthodes MCMC, on rappelle d'abord la théorie générale des chaînes de Markov [Feller, 1968].

[A] Chaînes de Markov

Soit $(\boldsymbol{Y}_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires discrètes ou continues à valeur dans un espace $\boldsymbol{E}$. Dans la suite, on suppose que $\mathcal{E}$ est un borélien sur $\boldsymbol{E}$. Si l'information induite par la connaissance de toutes les variables antérieures à l'instant n est restreinte à celle contenue dans les k instants précédents n :

$$\mathrm{P}(\boldsymbol{Y}_n \mid \boldsymbol{Y}_{n-1}, .., \boldsymbol{Y}_1) = \mathrm{P}(\boldsymbol{Y}_n \mid \boldsymbol{Y}_{n-1}, .., \boldsymbol{Y}_{n-k}),$$

alors on dit que la suite $(\boldsymbol{Y}_n)_{n\in\mathbb{N}}$ est une chaîne de Markov d'ordre k. Dans la suite, on ne considère que les chaînes de Markov d'ordre 1 :

$$\mathrm{P}(\boldsymbol{Y}_n \mid \boldsymbol{Y}_{n-1}, .., \boldsymbol{Y}_1) = \mathrm{P}(\boldsymbol{Y}_n \mid \boldsymbol{Y}_{n-1})$$

[7] Si on sait parfaitement échantillonner la densité $\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{h})$ le recours aux techniques de Monte-Carlo pour des fonctions h simples n'est pas justifié.

[8] Pour une fonction h précise, l'optimalité est obtenue pour $g(\boldsymbol{\theta}) = \frac{|h(\boldsymbol{\theta})| f(\boldsymbol{\theta})}{\int |h(\boldsymbol{\theta})| f(\boldsymbol{\theta}) d\,\boldsymbol{\theta}}$ [Rubinstein, 1981].

$\mathrm{P}(\boldsymbol{Y}_n \mid \boldsymbol{Y}_{n-1})$ est appelée le noyau de transition $\mathcal{K}_n(. \mid .)$ de la chaîne. La chaîne de markov est dite *h*omogène si le noyau de transition ne varie pas avec l'indice n, on le note $\mathcal{K}(. \mid .)$. Le noyau $\mathcal{K}(. \mid .)$ peut être considéré comme un opérateur qui transforme la densité de probabilité π_n de $\boldsymbol{Y}_n$ en π_{n+1} la d.d.p de $\boldsymbol{Y}_{n+1}$:

$$\pi_{n+1}(\boldsymbol{y}) = \int \mathcal{K}(\boldsymbol{y} \mid \boldsymbol{y}')\pi_n(\boldsymbol{y}')d\,\boldsymbol{y}'$$

On note aussi $\mathcal{K}^n(. \mid .)$ la probabilité de transition entre l'état initial et l'état à l'instant n :

$$\mathcal{K}^n(A \mid \boldsymbol{y}) = \mathrm{P}(\boldsymbol{Y}_n \in A \mid \boldsymbol{Y}_0 = \boldsymbol{y})$$

La chaîne de Markov $(\boldsymbol{Y}_n)_{n\in\mathbb{N}}$ est ainsi définie par une distribution initiale π_0 et par un noyau de transition $\mathcal{K}(. \mid .)$.

Une réalisation de la chaîne est obtenue par l'algorithme itératif suivant :

Réalisation d'une chaîne de Markov

$1 - \boldsymbol{y}_0 \sim \pi_0$

$2 - \boldsymbol{y}_{n+1} \sim K(\boldsymbol{y}_{n+1} \mid \boldsymbol{y}_n)$

(II.13)

L'étude de la convergence des chaînes de Markov est bien établie dans la littérature [Meyn et Tweedie, 1993]. Le chapitre 3 de l'ouvrage [Robert, 1996] et [Tierney, 1994] représentent une bonne synthèse de ce domaine. L'annexe B de [Senecal, 2002] résume les notions et théorèmes nécessaires pour l'étude des algorithmes MCMC. On rappelle brièvement quelques définitions nécessaires pour établir deux théorèmes importants concernant la convergence d'une chaîne de Markov.

Invariance : Une distribution ϕ est invariante par le noyau de transition $\mathcal{K}(. \mid .)$ si $\phi = \mathcal{K}\phi$. Une transition donnée $\mathcal{K}$ peut avoir une seule, plusieurs ou ne pas avoir de distributions invariantes.

Irréductibilité : Un noyau de transition $\mathcal{K}(. \mid .)$ est dit ϕ-irréductible si pour tout $\boldsymbol{y}_0 \in \boldsymbol{E}$ et tout ensemble $A \in \mathcal{E}$ de mesure non nulle par rapport à ϕ ($\phi(A) > 0$), il existe un entier $n \geq 1$ tel que $\mathcal{K}^n(A \mid \boldsymbol{y}_0) > 0$. Cette propriété signifie que la chaîne de Markov visite tous les ensembles de mesure ϕ non nulle et donc qu'elle ne reste pas bloquée dans une région de l'espace $\boldsymbol{E}$. La notion de ϕ-irréductibilité est liée à la connectivité de l'espace $\boldsymbol{E}$ sous la transition $\mathcal{K}$.

L'irréductibilité garantit la visite de tout l'espace mais ne renseigne pas sur la fréquence de cette visite. On introduit alors la propriété de **récurrence** d'une chaîne ϕ-irréductible qui signifie que tout ensemble A de mesure $\phi(A) > 0$ est visité une infinité de fois avec une probabilité non nulle à partir de tout point $\boldsymbol{y}_0 \in \boldsymbol{E}$ et avec une probabilité 1 pour ϕ-presque tout point $\boldsymbol{y}_0$:

Définition 1 *Un noyau de transition $\mathcal{K}(. \mid .)$ ϕ-irréductible est récurrent si et seulement si*

$$\begin{cases} \forall \boldsymbol{y}_0 \in \boldsymbol{E} \text{ et } A \in \mathcal{E} \text{ tel que } \phi(A) > 0 \\ \mathrm{P}(\boldsymbol{Y}_n \in A \text{ infiniment souvent } \mid \boldsymbol{y}_0) > 0 \\ \text{Pour } \phi\text{-presque tout point } \boldsymbol{y}_0 \in \boldsymbol{E} \text{ et } A \in \mathcal{E} \text{ tel que } \phi(A) > 0 \\ \mathrm{P}(\boldsymbol{Y}_n \in A \text{ infiniment souvent } \mid \boldsymbol{y}_0) = 1 \end{cases}$$

Une chaîne est dite **récurrente au sens de Harris** si A est visité une infinité de fois avec une probabilité 1 pour tout $\boldsymbol{y}_0 \in \boldsymbol{E}$ (en autorisant les ensembles de $\boldsymbol{y}_0$ de mesure nulle).

Lorsque $\mathcal{K}$ admet π comme densité invariante, il est naturel de considérer la π-irréductibilité. On a le résultat suivant [Tierney, 1994] qui assure, pour une chaîne irréductible, l'unicité de la distribution invariante :

Proposition 1 *Si une chaîne de Markov est irréductible et admet une distribution invariante π, alors la chaîne est π-irréductible, π est l'unique distribution invariante de la chaîne et la chaîne est récurrente positive*[9].

Apériodicité : Une chaîne de Markov est dite apériodique si elle n'a pas un comportement périodique lors de ses transitions. La période m d'un noyau de transition peut être définie comme le cardinal minimum d'une partition $(C_1, ..., C_m)$ de $\boldsymbol{E}$ vérifiant :

$$\forall \boldsymbol{y} \in C_i,\ \mathcal{K}(C_{i+1[m]} \mid \boldsymbol{y}) = 1$$

Une chaîne est apériodique si la période m est égale à 1.

Ergodicité : Une chaîne est ergodique si elle est irréductible, apériodique, admet une distribution invariante et récurrente au sens de Harris.

Avec ces définitions, on peut énoncer quelques résultats sur le comportement asymptotique des échantillons générés par l'algorithme (II.13) et sur la convergence des sommes empiriques $\frac{1}{M}\sum_{m=1}^{M} h(\boldsymbol{y}_m)$.

Théorème 2 *Soit une chaîne de Markov de noyau de transition $\mathcal{K}(. \mid .)$ irréductible et π-invariante. En construisant la mesure $\bar{\mathcal{K}}^M(. \mid \boldsymbol{y}_0)$ sur $\mathcal{E}$ par :*

$$\forall A \in \mathcal{E}, \bar{\mathcal{K}}^M(A \mid \boldsymbol{y}_0) = \frac{1}{M}\sum_{m=1}^{M} \mathcal{K}^m(A \mid \boldsymbol{y}_0)$$

Alors, pour π-presque tout point de départ $\boldsymbol{y}_0$,

$$\lim_{M\to\infty} \|\bar{\mathcal{K}}^M(. \mid \boldsymbol{y}_0) - \mu_\pi(.)\|_{VT} = 0$$

et pour toute fonction h π-intégrable :

$$\frac{1}{M}\sum_{m=1}^{M} h(\boldsymbol{y}_m) \xrightarrow{p.s.} \mathrm{E}_\pi\big[h\big]$$

où $\|\mu_1 - \mu_2\|_{VT}$ est la norme de la variation totale entre deux mesures définie par :

$$\|\mu_1 - \mu_2\|_{VT} = \sup_A |\mu_1(A) - \mu_2(A)|$$

Le théorème (2) assure la convergence de la moyenne des distributions $\mathcal{K}^m$ vers la distribution invariante π et celle des sommes empiriques vers les espérances théoriques, pour π-presque tout point de départ $\boldsymbol{y}_0$. En ajoutant la propriété de l'apériodicité de la chaîne, on assure en plus la convergence en loi vers π de la distribution $\mathcal{K}^m(. \mid \boldsymbol{y}_0)$:

Théorème 3 *Soit une chaîne de Markov de noyau de transition $\mathcal{K}(. \mid .)$ irréductible, π-invariante et apériodique. Alors, pour π-presque tout point de départ $\boldsymbol{y}_0$,*

$$\lim_{M\to\infty} \|\mathcal{K}^M(. \mid \boldsymbol{y}_0) - \mu_\pi(.)\|_{VT} = 0 \tag{II.14}$$

Si la chaîne de Markov est en plus ergodique (en ajoutant la propriété de la récurrence au sens de Harris), les convergences dans les deux théorèmes précédents sont assurées pour tout point de départ $\boldsymbol{y}_0 \in \boldsymbol{E}$ en autorisant ainsi les ensembles de mesure nulle par rapport à π.

[9] Une chaîne est récurrente positive si elle est irréductible, récurrente et admet une distribution invariante.

[B] Algorithmes MCMC

Etant donnée une densité π^* difficile à échantillonner, le but des algorithmes MCMC est double :

1. Construire une chaîne de Markov $(\boldsymbol{Y}_n)_{n\in\mathbb{N}}$ selon (II.13) qui converge en loi vers la densité d'intérêt π^* (problème inverse de l'étude de convergence d'un noyau $\mathcal{K}$).
2. Approcher asymptotiquement les espérances $\mathrm{E}_\pi\big[h\big]$ par des moyennes empiriques $\frac{1}{M}\sum_{m=k+1}^{k+M} h(\boldsymbol{y}_m)$, k est le temps supposé nécessaire pour la convergence[10] de la chaîne ("temps de chauffe").

Si on exige que la convergence en loi soit au sens de (II.14) pour tout point de départ $\boldsymbol{y}_0$, la chaîne de Markov doit être alors ergodigue π^*-invariante. Parmi les algorithmes MCMC, l'échantillonnage de Gibbs et l'algorithme de Metropolis-Hastings sont les plus connus, étudiés et utilisés en pratique.

Echantillonnage de Gibbs : Soit $\pi^*(\boldsymbol{y})$ la distribution d'intérêt à échantillonner. On note $\pi_j^*(y_j \mid y_{-j}) = \pi_j^*(y_j \mid y_1, .., y_{j-1}, y_{j+1}, ..., y_p)$ la distribution conditionnelle de la composante j du vecteur $\boldsymbol{y}$ connaissant toutes les autres composantes y_{-j}. L'échantillonnage de Gibbs consiste à simuler le vecteur des composantes de $\boldsymbol{y}$ d'une manière cyclique. En partant d'un point initial $\boldsymbol{y}^0 = (y_1^0, ..., y_p^0)$, on génère la suite $\boldsymbol{y}^1, \boldsymbol{y}^2, \boldsymbol{y}^3$,..., avec $\boldsymbol{y}^{m+1}$ obtenue à partir de $\boldsymbol{y}^m$ de la façon suivante :

Algorithme de Gibbs

$$\left\{\begin{array}{lll} y_1^{m+1} & \text{est échantillonné selon} & \pi_1^*(y_1 \mid y_2^m, y_3^m, ..., y_p^m) \\ y_2^{m+1} & \text{est échantillonné selon} & \pi_2^*(y_2 \mid y_1^{m+1}, y_3^m, y_4^m, ..., y_p^m) \\ \ldots & \ldots & \ldots \\ y_p^{m+1} & \text{est échantillonné selon} & \pi_p^*(y_p \mid y_1^{m+1}, y_2^{m+1}, ..., y_{p-1}^{m+1}) \end{array}\right. \tag{II.15}$$

La suite $(\boldsymbol{y}^m)_{m\in\mathbb{N}}$ est bien une chaîne de Markov d'ordre 1. En effet, l'échantillonnage de $\boldsymbol{y}^{m+1}$ ne dépend que de $\boldsymbol{y}^m$. Le noyau de transition $\mathcal{K}_G(. \mid .)$ est défini par :

$$\begin{aligned}\mathcal{K}_G(\boldsymbol{y}^{m+1} \mid \boldsymbol{y}^m) \quad &= \pi_1^*(y_1^{m+1} \mid y_2^m, y_3^m, ..., y_p^m)\, \pi_2^*(y_2^{m+1} \mid y_1^{m+1}, y_3^m, y_4^m, ..., y_p^m) \\ &\quad \ldots \pi_p^*(y_p^{m+1} \mid y_1^{m+1}, y_2^{m+1}, ..., y_{p-1}^{m+1})\end{aligned}$$

On montre facilement que la distribution π^* est invariante par le noyau $\mathcal{K}_G$. L'irréductibilité (et donc l'unicité de la distribution invariante d'après la proposition (1)) et l'apériodicité ne sont pas automatiquement vérifiées par le noyau $\mathcal{K}_G$. Elles dépendent de la distribution π^*.

On note que l'échantillonnage de Gibbs ne suppose pas la connaissance de la distribution π^*. Seule la connaissance des lois conditionnelles est nécessaire pour la construction de la chaîne. C'est souvent le cas dans les problèmes à variables cachées traités dans la section (II.4). En effet, la distribution *a posteriori* $\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{h})$ d'un paramètre d'intérêt $\boldsymbol{\theta}$ se met sous la forme d'une intégrale, en général, difficile à calculer analytiquement ou à échantillonner :

$$\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{h}) = \int \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{c} \mid \boldsymbol{x}_{1..T}) d\,\boldsymbol{c}.$$

[10] cette convergence n'est qu'asymptotique mais en pratique on ne dispose que d'un nombre fini d'échantillons et on essaie d'éliminer les premiers échantillons qui peuvent altérer le calcul empirique des espérances.

Cependant, les lois conditionnelles $\mathrm{p}_\theta(\boldsymbol{\theta} \mid \boldsymbol{c}, \boldsymbol{x}_{1..T})$ et $\mathrm{p}_c(\boldsymbol{c} \mid \boldsymbol{\theta}, \boldsymbol{x}_{1..T})$ sont simulables. L'implémentation de l'échantillonnage de Gibbs est très utile dans ce cas. On part des points initiaux $\boldsymbol{\theta}^0$ et $\boldsymbol{c}^0$ quelconques et on itère le cycle suivant :

$$\left\{ \begin{array}{lll} \boldsymbol{\theta}^{m+1} & \text{est échantillonné selon} & \pi^*_\theta(\boldsymbol{\theta} \mid \boldsymbol{c}^m, \boldsymbol{x}_{1..T}) \\ \boldsymbol{c}^{m+1} & \text{est échantillonné selon} & \pi^*_c(\boldsymbol{c} \mid \boldsymbol{\theta}^{m+1}, \boldsymbol{x}_{1..T}) \end{array} \right. \tag{II.16}$$

La somme empirique $\frac{1}{M} \sum_{m=1}^{M} h(\boldsymbol{\theta}^m)$ converge presque sûrement vers l'espérance *a posteriori* $\mathrm{E}_{apost}\big[h\big] = \int h(\boldsymbol{\theta}) \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{h}) d\,\boldsymbol{\theta}$.

L'échantillonnage à partir des lois conditionnelles est parfois impossible. Si on peut évaluer la distribution π^* à une constante près, alors l'algorithme de Metropolis-Hastings [Metropolis *et al.*, 1953; Hastings, 1970] est une bonne alternative.

Algorithme de Metropolis-Hastings : La transition entre deux valeurs successives $\boldsymbol{y}_m$ et $\boldsymbol{y}_{m+1}$ se passe de la manière suivante : à partir de $\boldsymbol{y}_m$, on échantillonne un candidat $\boldsymbol{z}$ selon une distribution choisie (quelconque) $g(\boldsymbol{z} \mid \boldsymbol{y}_m)$ qu'on appelle la distribution instrumentale. On accepte $\boldsymbol{z}$ (ie, $\boldsymbol{y}_{m+1} = \boldsymbol{z}$) ou on le rejette en gardant la valeur $\boldsymbol{y}_m$ (ie, $\boldsymbol{y}_{m+1} = \boldsymbol{y}_m$) avec une probabilité $\rho(\boldsymbol{z}, \boldsymbol{y}_m) = min(\frac{\pi^*(\boldsymbol{z}) g(\boldsymbol{y}_m \mid \boldsymbol{z})}{\pi^*(\boldsymbol{y}_m) g(\boldsymbol{z} \mid \boldsymbol{y}_m)}, 1)$.

Algorithme de Metropolis-Hastings

1. initialiser $\boldsymbol{y}_0 \sim \pi_0(\boldsymbol{y})$,

2. à l'itération $m+1$:

   - proposer un candidat $\boldsymbol{z} \sim g(\boldsymbol{z} \mid \boldsymbol{y}_m)$

   - accepter $\boldsymbol{z}$ avec la probabilité $\rho(\boldsymbol{z}, \boldsymbol{y}_m)$:

     simuler $u \sim \mathcal{U}_{[0,1]}$

     - Si $u < \rho(\boldsymbol{z}, \boldsymbol{y}_m)$ alors $\boldsymbol{y}_{m+1} = \boldsymbol{z}$ sinon $\boldsymbol{y}_{m+1} = \boldsymbol{y}_m$

3. $m \leftarrow m+1$ et retourner à (2)

La probabilité d'acceptation étant :

$\rho(\boldsymbol{z}, \boldsymbol{y}_m) = min(\frac{\pi^*(\boldsymbol{z}) g(\boldsymbol{y}_m \mid \boldsymbol{z})}{\pi^*(\boldsymbol{y}_m) g(\boldsymbol{z} \mid \boldsymbol{y}_m)}, 1)$

Le noyau de transition $\mathcal{K}_{MH}(. \mid .)$ qui est par définition la distribution (résultante) de $\boldsymbol{y}_{m+1}$ connaissant $\boldsymbol{y}_m$ s'écrit :

$$\mathcal{K}_{MH}(\boldsymbol{y}_{m+1} \mid \boldsymbol{y}_m) = g(\boldsymbol{y}_{m+1} \mid \boldsymbol{y}_m)\, \rho(\boldsymbol{y}_{m+1}, \boldsymbol{y}_m),$$

si $\boldsymbol{y}_{m+1} \neq \boldsymbol{y}_m$ avec,

$$\mathrm{P}(\boldsymbol{y}_{m+1} = \boldsymbol{y}_m \mid \boldsymbol{y}_m) = 1 - \int g(\boldsymbol{y} \mid \boldsymbol{y}_m)\, \rho(\boldsymbol{y}, \boldsymbol{y}_m) d\,\boldsymbol{y}.$$

$\mathcal{K}_{MH}(. \mid .)$ peut se mettre sous la forme compacte suivante :

$$\mathcal{K}_{MH}(\boldsymbol{y}_{m+1} \mid \boldsymbol{y}_m) = g(\boldsymbol{y}_{m+1} \mid \boldsymbol{y}_m)\, \rho(\boldsymbol{y}_{m+1}, \boldsymbol{y}_m) + \left(1 - \int g(\boldsymbol{y} \mid \boldsymbol{y}_m)\, \rho(\boldsymbol{y}, \boldsymbol{y}_m) d\, \boldsymbol{y}\right) \delta_{\boldsymbol{y}_m}(\boldsymbol{y}_{m+1}).$$

On montre que π^* est une distribution invariante par la transition $\mathcal{K}_{MH}$ [Robert, 1996]. L'irréductibilité et l'apériodicité sont à étudier selon le contexte (la distribution π^* et la distribution instrumentale $g(. \mid .)$) [Robert, 1996].

Version hybride : L'échantillonnage de Gibbs présente des avantages et des inconvénients par rapport à l'algorithme de Metropolis-Hastings. Son principal avantage est que le noyau de transition $\mathcal{K}_G$ est construit uniquement à partir des lois conditionnelles de π^* et ne fait pas appel à une distribution arbitraire $g(.|.)$. L'algorithme de Gibbs exploite ainsi la structure de la distribution d'intérêt π^*. Cependant, il n'est pas toujours possible d'échantillonner selon les lois conditionnelles tandis qu'avec l'échantillonnage de Metropolis-Hastings on ne rencontre pas ce type de problème. L'autre inconvénient est le risque de blocage de l'échantillonneur de Gibbs causé par une forte corrélation entre les échantillons successifs de la chaîne des $\boldsymbol{y}_m$. Cette corrélation entre les échantillons $\boldsymbol{y}_m$ est due essentiellement à la corrélation[11] entre les différentes composantes y^j du vecteur $\boldsymbol{y}$.

Des versions hybrides de l'échantillonneur de Gibbs combinant l'algorithme de Gibbs et l'algorithme de Métropolis-Hastings peuvent être proposées. Un schéma proposé dans [Müller, 1991, 1992] consiste à reprendre l'algorithme de Gibbs mais en appliquant une itération de Métropolis-Hastings à chaque loi conditionnelle. On remplace la simulation de la loi conditionnelle $\pi_j^*(y_j \mid y_{-j})$ dans (IV.3) par une simulation selon une loi instrumentale $g_j(y_j \mid y_{-j})$.

Version hybride de Gibbs

- à l'itération m et à la composante j :
1. Simuler $z_j \sim g_j(z_j \mid y_1^{m+1}, ..., y_{j-1}^{m+1}, y_{j+1}^m ..., y_p^m)$
2. Prendre

$$y_j^{m+1} = \begin{cases} z_j & \text{avec probabilité} \quad \rho \\ y_j^m & \text{avec probabilité} \quad 1-\rho \end{cases}$$

avec

$$\rho = min\left(1, \frac{\pi_j^*(z_j \mid y_1^{m+1},...,y_{j-1}^{m+1},y_{j+1}^m ...,y_p^m)}{g_j(z_j \mid y_1^{m+1},...,y_{j-1}^{m+1},y_{j+1}^m ...,y_p^m)} \right.$$

$$\left. \Big/ \frac{\pi_j^*(y_j^m \mid y_1^{m+1},...,y_{j-1}^{m+1},y_{j+1}^m ...,y_p^m)}{g_j(y_j^m \mid y_1^{m+1},...,y_{j-1}^{m+1},y_{j+1}^m ...,y_p^m)} \right)$$

L'introduction de la loi instrumentale est motivée par deux raisons.

1. Les lois conditionnelles $\pi_j^*(y_j \mid y_{-j})$ ne sont pas simulables. On greffe alors une procédure de *Métropolis-Hastings* avec une seule itération.
2. Les composantes du vecteur $\boldsymbol{y}$ sont très corrélées. Simuler selon une autre distribution arbitraire peut débloquer l'échantillonneur de Gibbs.

[11] On peut parfois corriger cet inconvénient par une reparamétrisation ou un repartitionnement du vecteur $\boldsymbol{y}$.

II.6 Application en séparation de sources

Dans l'approche classique des statistiques, on peut apporter plusieurs solutions à un même problème donné. On commence par construire des estimateurs qu'on évalue *a posteriori* par leur biais et leur variance. Cet aspect est qualifié dans [De Finetti, 1974] par la "adhockery". L'avantage de l'approche bayésienne est qu'elle applique la même méthodologie et ne s'appuie pas sur des estimateurs *ad hoc* ou des schémas prédéfinis. L'information sur le problème direct est codée dans la vraisemblance et l'information *a priori* est codée dans la distribution *a priori*. Avec la règle de Bayes, on combine ces deux sources d'informations pour obtenir la distribution *a posteriori*. En minimisant une fonction coût choisie par le décideur, on obtient un estimateur qui reflète l'une des caractéristiques de la distribution *a posteriori*. Le problème de la séparation de sources constitue un bon exemple pour illustrer la méthodologie bayésienne.

On suppose dans la suite de ce paragraphe que les données $\boldsymbol{x}_{1..T}$ sont un mélange linéaire instantané bruité des sources :

$$\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{s}_t + \boldsymbol{\epsilon}_t, \quad t = 1..T \tag{II.17}$$

où $\boldsymbol{x}_t$ est le vecteur $(m \times 1)$ des observations à l'instant t, $\boldsymbol{s}_t$ est le vecteur $(n \times 1)$ des sources et $\boldsymbol{\epsilon}_t$ est le vecteur $(m \times 1)$ du bruit. $\boldsymbol{A}$ est la matrice de mélange $(m \times n)$.

Remarque 5 *t est un indice générique, il désigne le temps dans le chapitre (III), le pixel d'une image dans le chapitre (IV), la fréquence dans le chapitre (V) ou l'indice temps-fréquence.*

Seules les observations $\boldsymbol{x}_{1..T}$ sont connues. Le problème de séparation de sources admet plusieurs sous-problèmes selon la spécification du paramètre d'intérêt $\boldsymbol{\theta}$. Dans tous ces sous-problèmes, on va suivre la même méthodologie : (i) définition du problème d'inférence, (ii) construction de la distribution *a posteriori* avec la règle de Bayes, (iii) choix des probabilités (vraisemblance et *a priori*), (iv) choix du critère (v) et finalement choix de l'algorithme d'optimisation.

Estimation de $\boldsymbol{A}$: Le problème d'inférence considéré est $\mathcal{I} := (\boldsymbol{I} \wedge \boldsymbol{x}_{1..T} \longrightarrow \boldsymbol{A})$ où $\boldsymbol{I}$ représente toute l'information *a priori* qu'on possède sur le problème comme la nature du mélange, la loi du bruit $\boldsymbol{\epsilon}$, le nombre des sources et des observations...

En appliquant la règle de Bayes, la distribution *a posteriori* s'écrit :

$$\mathrm{p}(\boldsymbol{A} \mid \boldsymbol{x}_{1..T}, \boldsymbol{I}) \propto \mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{A}, \boldsymbol{I})\, \mathrm{p}(\boldsymbol{A} \mid \boldsymbol{I}) \tag{II.18}$$

Pour déterminer l'expression de la vraisemblance, on utilise la structure à **variables cachées** naturelle au problème inférentiel de séparation de sources :

$$\begin{aligned} \mathrm{p}(\boldsymbol{A} \mid \boldsymbol{x}_{1..T}, \boldsymbol{I}) &\propto \left[\int \mathrm{p}(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T} \mid \boldsymbol{A}, \boldsymbol{I}) d\,\boldsymbol{s}_{1..T}\right] \mathrm{p}(\boldsymbol{A} \mid \boldsymbol{I}) \\ &\propto \left[\int \mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{s}_{1..T}, \boldsymbol{A}, \boldsymbol{I}) \mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{I}) d\,\boldsymbol{s}_{1..T}\right] \mathrm{p}(\boldsymbol{A} \mid \boldsymbol{I}) \end{aligned}$$

Dans la suite, afin d'alléger les notations, on élimine la proposition $\boldsymbol{I}$ des différentes expressions et la probabilité p est indexée par la variable à laquelle elle se rapporte. Par exemple, p_s désigne $\mathrm{p}(\boldsymbol{I} \longrightarrow \boldsymbol{s}_{1..T})$.

Connaissant la matrice de mélange et les sources, l'incertitude sur les observations est due entièrement au bruit. Par changement de variables entre $\boldsymbol{x}_{1..T}$ et $\boldsymbol{\epsilon}_{1..T}$ (le jacobien valant 1 puisque le bruit est additif) :

$$\mathrm{p}(\boldsymbol{x}_{1..T} \mid , \boldsymbol{s}_{1..T}, \boldsymbol{A}) = \mathrm{p}_\epsilon(\boldsymbol{\epsilon}_{1..T}) = \mathrm{p}_\epsilon(\boldsymbol{x}_{1..T} - \boldsymbol{A}\,\boldsymbol{s}_{1..T}).$$

La distribution *a posteriori* de $\boldsymbol{A}$ s'écrit alors :

$$\mathrm{p}(\boldsymbol{A} \mid \boldsymbol{x}_{1..T}) \propto \left[\int \mathrm{p}_\epsilon(\boldsymbol{x}_{1..T} - \boldsymbol{A}\,\boldsymbol{s}_{1..T})\, \mathrm{p}_s(\boldsymbol{s}_{1..T}) d\,\boldsymbol{s}_{1..T}\right] \mathrm{p}_A(\boldsymbol{A}) \tag{II.19}$$

Pour exploiter l'expression (II.19), on doit choisir les probabilités p_ϵ, p_s et p_A. Bien qu'en pratique, on adopte le point de vue subjectif (paragraphe II.3) pour effectuer le choix des probabilités, ce choix n'est pas arbitraire et doit être basé sur notre connaissance physique du problème de mélange. Dans certains cas, un changement de base (passage dans le domaine de Fourrier, des ondelettes, temps-fréquence) peut faciliter considérablement le choix des probabilités. On est aussi guidé par des considérations pratiques d'implémentation et par l'interprétation finale du critère d'estimation (le choix des lois gaussiennes signifie qu'on veut se limiter aux statistiques d'ordre deux).

Une fois l'estimateur fixé (une caractéristique particulière de la distribution *a posteriori* de $\boldsymbol{A}$) par le choix d'une fonction coût $\mathrm{C}(\boldsymbol{A}, \boldsymbol{A}^*)$, on se trouve face à un problème technique d'optimisation. Souvent le calcul explicite des caractéristiques de la distribution *a posteriori* (II.19) (comme le Maximum *a posteriori* MAP, espérance *a posteriori* EAP, MAP marginal, EAP marginal...) n'est pas possible. On profite alors de la structure à variables cachées pour implémenter l'algorithme EM ou l'échantillonnage de Gibbs.

Remarque 6
Relation avec l'ACI *: L'estimateur particulier* MAP *peut être interprété comme une régularisation de l'estimateur du maximum de vraisemblance en analyse en composantes indépendantes*[12]*. En effet, avec le modèle non bruité* $\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{s}_t$*, la distribution a posteriori (II.19) s'écrit :*

$$\mathrm{p}(\boldsymbol{A} \mid \boldsymbol{x}_{1..T}) \propto |\boldsymbol{A}|^{-1} \mathrm{p}_s(\boldsymbol{A}^{-1}\boldsymbol{x}_{1..T}) \mathrm{p}_A(\boldsymbol{A}).$$

En faisant le changement de variables entre la matrice $\boldsymbol{A}$ *et son inverse* $\boldsymbol{B} = \boldsymbol{A}^{-1}$ *et en supposant que les sources* $\boldsymbol{s}_{1..T}$ *sont i.i.d. :*

$$\mathrm{p}(\boldsymbol{B} \mid \boldsymbol{x}_{1..T}) \propto \prod_{t=1}^{T} |\boldsymbol{B}| \mathrm{p}_s(\boldsymbol{B}\boldsymbol{x}_t) \mathrm{p}_B(\boldsymbol{B}) \tag{II.20}$$

avec $\mathrm{p}_B(\boldsymbol{B})$ *la loi a priori de la matrice* $\boldsymbol{B}$ *obtenue à partir de celle de* $\boldsymbol{A}$ [13] *:*

$$\mathrm{p}_B(\boldsymbol{B}) = \mathrm{p}_A(\boldsymbol{B}^{-1}) |\frac{\partial \boldsymbol{B}^{-1}}{\partial \boldsymbol{B}}|$$

L'incrément $\Delta\boldsymbol{B}$ *d'un algorithme de gradient maximisant le logarithme de la distribution a posteriori de* $\boldsymbol{B}$ *s'écrit :*

$$\Delta\boldsymbol{B} = \underbrace{\sum_{t=1}^{T} (\phi_s(\boldsymbol{y}_t)\boldsymbol{y}_t^T + \boldsymbol{I})\boldsymbol{B}^{-T}}_{\text{Incrément de l'ACI}} + \underbrace{\frac{\partial}{\partial \boldsymbol{B}} \log \mathrm{p}_B(\boldsymbol{B})}_{\text{Terme de régularisation}}$$

où $\phi_s = \frac{\mathrm{p}_s'}{\mathrm{p}_s}$ *est la fonction score et* $\boldsymbol{y}_t = \boldsymbol{B}\boldsymbol{x}_t$*. Le terme* $\log \mathrm{p}_B(\boldsymbol{B})$ *peut être ainsi considéré comme un terme de régularisation* [14]*.*

Estimation de $\boldsymbol{s}_{1..T}$: Dans ce cas le problème d'inférence est défini par $\mathcal{I} := (\boldsymbol{x}_{1..T}, \boldsymbol{I} \longrightarrow \boldsymbol{s}_{1..T})$. Connaissant les données $\boldsymbol{x}_{1..T}$ et l'information *a priori* $\boldsymbol{I}$ (incluant le fait que les données $\boldsymbol{x}_{1..T}$ suivent le modèle de mélange linéaire instantané bruité (II.17)), l'objectif est de reconstruire les sources $\boldsymbol{s}_{1..T}$.

En appliquant la règle de Bayes, la distribution *a posteriori* des sources est :

$$\mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{x}_{1..T}, \boldsymbol{I}) \propto \mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{s}_{1..T}, \boldsymbol{I}) \mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{I})$$

Le modèle direct ($\boldsymbol{s}_{1..T} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{x}_{1..T}$) n'est pas un modèle de mélange linéaire. En supposant, dans une approche bayésienne, que la matrice $\boldsymbol{A}$ est une variable aléatoire et en l'intégrant hors du problème

[12] Dans l'approche classique, on ne trouve pas l'équivalent d'autres estimateurs comme l'EAP ou l'EAPM.
[13] En général, on possède une information physique sur la matrice de mélange $\boldsymbol{A}$ et non sur son inverse $\boldsymbol{B}$.
[14] La propriété de l'équivariance [Cardoso et Labeld, 1996] peut être conservée pour une certaine classe de lois *a priori* sur la matrice $\boldsymbol{B}$.

la relation entre les données $\boldsymbol{x}_{1..T}$ et les sources $\boldsymbol{s}_{1..T}$ n'est plus linéaire. La vraisemblance, modélisant le problème direct, s'écrit sous forme intégrale :

$$\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{s}_{1..T}, \boldsymbol{I}) = \int \mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{s}_{1..T}, \boldsymbol{A}, \boldsymbol{I})\, \mathrm{p}(\boldsymbol{A} \mid \boldsymbol{I})\, d\,\boldsymbol{A}.$$

En faisant un changement de variables entre le vecteur des observations $\boldsymbol{x}_{1..T}$ et le bruit $\boldsymbol{\epsilon}_{1..T}$ et en introduisant les probabilités indexées p_ϵ, p_s et p_A du bruit, des sources et de la matrice de mélange, la distribution *a posteriori* des sources s'écrit :

$$\mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{x}_{1..T}) \propto \left[\int \mathrm{p}_\epsilon(\boldsymbol{x}_{1..T} - \boldsymbol{A}\,\boldsymbol{s}_{1..T})\, \mathrm{p}_A(\boldsymbol{A})\, d\,\boldsymbol{A}\right]\, \mathrm{p}_s(\boldsymbol{s}_{1..T}). \tag{II.21}$$

On obtient ainsi une expression symétrique de la distribution *a posteriori* de la matrice $\boldsymbol{A}$ (II.19). La discussion sur le choix des probabilités p_ϵ, p_s et p_A et sur l'aspect algorithmique est la même que celle dans le paragraphe précédent (estimation de $\boldsymbol{A}$). En général, on n'a pas une forme explicite de la distribution *a posteriori* (II.21) ou le calcul des caractéristiques de cette distribution comme l'estimateur MAP ou l'estimateur EAP est difficile à mener. On profite alors de la structure à variables cachées ($\boldsymbol{A}$ étant la variable cachée) en implémentant l'algorithme EM (pour le calcul du MAP) ou les techniques bayésiennes MCMC (pour le calcul des estimateurs du type $\mathrm{E}\big[h(\boldsymbol{s}_{1..T})\big]$).

Remarque 7 *Le fait d'estimer dans un premier temps la matrice de mélange* $\hat{\boldsymbol{A}} = \arg\max \mathrm{p}(\boldsymbol{A} \mid \boldsymbol{x}_{1..T}, \boldsymbol{I})$ *et de reconstruire dans un deuxième temps les sources* $\hat{\boldsymbol{s}}_{1..T} = \arg\max \mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{x}_{1..T}, \hat{\boldsymbol{A}}, \boldsymbol{I})$ *ne s'inscrit pas dans une méthodologie bayésienne rigoureuse et constitue plutôt une méthode approximée.*

Estimation conjointe : Le problème d'inférence est défini par $\mathcal{I} := (\boldsymbol{x}_{1..T} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{s}_{1..T} \wedge \boldsymbol{A})$. Connaissant les données $\boldsymbol{x}_{1..T}$, on veut estimer conjointement la matrice de mélange et les sources. La distribution *a posteriori* s'écrit :

$$\mathrm{p}(\boldsymbol{s}_{1..T}, \boldsymbol{A} \mid \boldsymbol{x}_{1..T}) \propto \mathrm{p}_\epsilon(\boldsymbol{x}_{1..T} - \boldsymbol{A}\,\boldsymbol{s}_{1..T})\mathrm{p}_A(\boldsymbol{A})\, \mathrm{p}_s(\boldsymbol{s}_{1..T}). \tag{II.22}$$

La forme analytique de la distribution *a posteriori* est explicite en fonction de la matrice de mélange et des sources. En effet, cette expression ne fait pas intervenir des intégrales comme c'est le cas avec les expressions (II.19) et (II.21). Cependant, le calcul exact des caractéristiques (MAP, EAP,...) de cette distribution n'est pas en général abordable et on fait appel aux techniques numériques itératives. Pour le calcul du MAP, on pourrait être tenté par utiliser la technique de relaxation en profitant de la décomposition naturelle du paramètre d'intérêt $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{s}_{1..T})$ en sous vecteurs $\boldsymbol{\theta}_1 = \boldsymbol{A}$ et $\boldsymbol{\theta}_2 = \boldsymbol{s}_{1..T}$, mais cette technique n'évite pas les maxima locaux. Avec cette même décomposition du vecteur $\boldsymbol{\theta}$, l'échantillonneur de Gibbs ou sa version hybride sont mieux adaptés. Le schéma de l'échantillonneur de Gibbs est le suivant :

Echantillonneur de Gibbs

1. simuler $(\boldsymbol{A}^{(0)}, \boldsymbol{s}_{1..T}^{(0)}) \sim \pi_0(\boldsymbol{A}, \boldsymbol{s}_{1..T})$

2. à l'itération m :

$$\begin{cases} \boldsymbol{s}_{1..T}^{(m)} & \sim & \pi_s(\boldsymbol{s}_{1..T} \mid \boldsymbol{x}_{1..T}, \boldsymbol{A}^{(m-1)}) \\ \boldsymbol{A}^{(m)} & \sim & \pi_A(\boldsymbol{A} \mid \boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}^{(m)}) \end{cases}$$

3. $m \leftarrow m + 1$ et retour à (2)

Les conditions de convergence de la chaîne de Markov sont liées aux probabilités p_ϵ, p_s et p_A à travers les lois conditionnelles $\pi_s(\boldsymbol{s}_{1..T} \mid \boldsymbol{x}_{1..T}, \boldsymbol{A})$ et $\pi_A(\boldsymbol{A} \mid \boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T})$. On approche alors les espérances du type $\mathrm{E}\big[h(\boldsymbol{A}, \boldsymbol{s}_{1..T})\big]$ par des sommes empiriques en se basant sur les échantillons $(\boldsymbol{A}^{(m)}, \boldsymbol{s}_{1..T}^{(m)})$ obtenus par l'échantillonneur de Gibbs.

Remarque 8 *L'un des avantages de l'algorithme* MCMC *est qu'il donne aussi la possibilité de calculer numériquement les caractéristiques marginales de la distribution a posteriori (II.22). Ainsi, les espérances marginales* $\mathrm{E}_{\boldsymbol{A} \mid .}\big[h_1(\boldsymbol{A})\big]$ *et* $\mathrm{E}_{\boldsymbol{s}_{1..T} \mid .}\big[h_2(\boldsymbol{s}_{1..T})\big]$ *sont approximées par :*

$$\begin{cases} \mathrm{E}_{\boldsymbol{A} \mid .}\big[h_1(\boldsymbol{A})\big] & \approx \frac{1}{M} \sum_{m=1}^{M} h_1(\boldsymbol{A}^{(m)}) \\ \mathrm{E}_{\boldsymbol{s}_{1..T} \mid .}\big[h_2(\boldsymbol{s}_{1..T})\big] & \approx \frac{1}{M} \sum_{m=1}^{M} h_2(\boldsymbol{s}_{1..T}^{(m)}) \end{cases}$$

Introduction des variables cachées : Le principal reproche à la méthode du maximum de vraisemblance qu'on trouve dans la littérature de la séparation de sources[15] est le choix de la densité *a priori* des sources $\mathrm{p}_s(\boldsymbol{s}_{1..T} \mid \boldsymbol{I})$. Ce choix doit tenir compte de trois impératifs :

1. La forme de la densité des sources doit être assez générale pour s'adapter à plusieurs types de sources.
2. L'expression de cette densité ne doit pas être compliquée pour permettre une implémentation efficace des algorithmes de séparation.
3. Il faut garantir l'identifiabilité du modèle (entre autres celle de la matrice de mélange). Par exemple, choisir une densité gaussienne i.i.d. pour les sources rend la matrice de mélange non identifiable.

L'introduction d'un modèle hiérarchique rentre bien dans l'approche bayésienne et apporte une solution flexible au choix de la densité des sources. Un modèle hiérarchique d'ordre 1 consiste à introduire une couche de variables cachées $\boldsymbol{z}_{1..T}$ (discrètes ou continues) expliquant la génération *a priori* des sources :

$$(\boldsymbol{I} \longrightarrow \boldsymbol{s}_{1..T}) \rightsquigarrow (\boldsymbol{I} \longrightarrow \boldsymbol{z}_{1..T} \longrightarrow \boldsymbol{s}_{1..T}).$$

La distribution des sources s'écrit alors sous une forme intégrale :

$$\mathrm{p}_s(\boldsymbol{s}_{1..T} \mid \boldsymbol{I}) = \int \mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{z}_{1..T}, \boldsymbol{I}) \mathrm{p}(\boldsymbol{z}_{1..T} \mid \boldsymbol{I}) d\, \boldsymbol{z}_{1..T}, \tag{II.23}$$

où la densité conditionnelle $\mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{z}_{1..T}, \boldsymbol{I})$ appartient à une famille paramétrique de dimension k : $\{\phi(. \mid \boldsymbol{\eta}), \boldsymbol{\eta} \in \mathbb{R}^k\}$. On choisit la forme de la fonction ϕ de manière à avoir des expressions simples à manipuler lors de l'implémentation des algorithmes de séparation.

Exemple 1 *Le mélange de gaussiennes est un exemple de modèle hiérarchique utilisé avec succès dans les problèmes de séparation de sources i.i.d. [Bermond, 2000; Attias, 1999; Snoussi et Mohammad-Djafari, 2000]. Dans ce modèle, les variables cachées* $\boldsymbol{z}_{1..T}$ *sont discrètes i.i.d. et les densités conditionnelles paramétriques* $\phi(. \mid \boldsymbol{\eta})$ *sont des gaussiennes.*

Outre son importance au niveau de la modélisation des sources, le modèle hiérarchique (II.23) est bien adapté au problème de séparation de sources. En effet,

[15] On retrouve aussi dans les méthodes d'analyse en composantes indépendantes le problème du choix de la distribution des sources. Le mérite de l'approche bayésienne est qu'elle rend explicite ce problème.

1. concernant l'estimation de la matrice de mélange $\boldsymbol{A}$, les sources $\boldsymbol{s}_{1..T}$ forment une première couche de variables cachées. La méthode de séparation se basant sur cette structure cachée (EM, MCMC) est alors flexible à l'introduction d'autres couches de variables cachées comme les $\boldsymbol{z}_{1..T}$ qui forment une deuxième couche de variables cachées pour l'estimation des paramètres $\boldsymbol{\eta}$ de la densité des sources.
2. Les variables cachées peuvent donner à la distribution *a posteriori* une interprétation facilitant la discussion sur l'identifiabilité de la matrice de mélange $\boldsymbol{A}$. Comme nous allons le voir dans les chapitres suivants, en prenant des variables $\boldsymbol{z}_{1..T}$ discrètes, des distributions conditionnelles $\mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{z}_{1..T}, \boldsymbol{I})$ gaussiennes et un bruit blanc gaussien, la séparation va se baser sur des statistiques d'ordre deux en exploitant la non stationnarité des sources.

Prédiction : Le problème d'inférence est défini par $\mathcal{I} := (\boldsymbol{x}_{1..T} \wedge \boldsymbol{I} \longrightarrow \mathrm{p}(\boldsymbol{x}_{t+1}))$. Connaissant les données $\boldsymbol{x}_{1..T}$, notre objectif est de prédire la densité de l'observation $\boldsymbol{x}_{t+1}$ à l'instant $(t+1)$. A titre illustratif, on suppose que les $\boldsymbol{x}_t$ sont i.i.d. de densité commune p. La variable à manipuler est donc une densité de probabilité. La méthodologie bayésienne se généralise facilement à l'espace $\mathcal{P} = \{\mathrm{p}, \int \mathrm{p} = 1\}$ des distributions de probabilités. En effet, on peut écrire la densité *a posteriori* de p avec la règle de Bayes :

$$\mathrm{P}_r(\mathrm{p} \mid \boldsymbol{x}_{1..T}) \propto \mathrm{P}_r(\boldsymbol{x}_{1..T} \mid \mathrm{p})\, \mathrm{P}_r(\mathrm{p}),$$

où $\mathrm{P}_r(\boldsymbol{x}_{1..T} \mid \mathrm{p}) = \prod_{t=1}^{T} \mathrm{p}(\boldsymbol{x}_t)$ est la vraisemblance de la *d.d.p* p. $\mathrm{P}_r(\mathrm{p})$ représente l'information *a priori* sur p. Le coût d'estimation $\mathrm{C}(\mathrm{p}, \mathrm{q})$ peut être défini comme une mesure de divergence d [16] entre la vraie distribution inconnue p et une distribution q [Zhu et Rohwer, 1995]. Le critère d'estimation (espérance *a posteriori* du coût $\mathrm{C}(\mathrm{p}, \mathrm{q})$) s'écrit :

$$\mathcal{J}(\mathrm{q}) = \int_{\mathrm{p}} \mathrm{d}(\mathrm{p}, \mathrm{q}) \mathrm{P}_r(\mathrm{p} \mid \boldsymbol{x}_{1..T}). \tag{II.24}$$

En considérant la famille D_δ des δ-divergences [Amari, 1985],

$$\mathrm{D}_\delta(\mathrm{p}, \mathrm{q}) = \frac{1}{\delta(1-\delta)} \left(1 - \int \mathrm{p}^\delta\, \mathrm{q}^{1-\delta}\right),$$

le minimiseur $\hat{\mathrm{q}}$ du critère (II.24) vérifie la relation suivante :

$$\hat{\mathrm{q}}^\delta =< \mathrm{p}^\delta >= \int_{\mathrm{p}} \mathrm{p}^\delta \mathrm{P}_r(\mathrm{p} \mid \boldsymbol{x}_{1..T}). \tag{II.25}$$

Dans le cas où la distribution p appartient à une famille paramétrique $\mathcal{Q}$:

$$\mathcal{Q} = \{\mathrm{p}(. \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \mathbb{R}^k\}$$

on peut avoir une expression analytique de l'estimateur $\hat{\mathrm{q}}$ (II.25) :

$$\hat{\mathrm{q}}(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_{1..T})^\delta = \int_{\boldsymbol{\theta}} \mathrm{p}(\boldsymbol{x}_{t+1} \mid \boldsymbol{\theta})^\delta \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}) d\boldsymbol{\theta}. \tag{II.26}$$

Exemple 2 *Si $\delta = 1$ (D_1 est la divergence de Kullback-Leibler), $\hat{\mathrm{q}}$ est l'estimateur* EAP *de la distribution* p *:*

$$\hat{\mathrm{q}} = \mathrm{E}_{\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}} \big[\mathrm{p}(\boldsymbol{x}_{t+1} \mid \boldsymbol{\theta})\big]$$

Dans le problème de séparation de sources, on se trouve dans le cas paramétrique avec $\boldsymbol{\theta}$ représentant soit :

[16] Ceci nécessite la manipulation des outils de la géométrie différentielle. Une grande partie du chapitre (VII) est consacrée à ces notions.

1. la matrice de mélange $\boldsymbol{A}$: la famille paramétrique $\mathcal{Q}$ est $\{\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{A}), \boldsymbol{A} \in \mathbb{R}^{m \times n}\}$,
2. les sources $\boldsymbol{s}_{1..T}$: la famille paramétrique $\mathcal{Q}$ est $\{\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{s}), \boldsymbol{s} \in \mathbb{R}^{n}\}$,
3. la matrice de mélange et les sources : la famille paramétrique $\mathcal{Q}$ est $\{\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{A}, \boldsymbol{s}), (\boldsymbol{A}, \boldsymbol{s}_{1..T}) \in \mathbb{R}^{m \times n} \times \mathbb{R}^{n}\}$.

On note que l'ensemble $\mathcal{Q}$ change selon le choix du paramètre $\boldsymbol{\theta}$.

II.7 Conclusion

Nous avons essayé dans ce chapitre de décrire les fondements de la méthode bayésienne au niveau fondamental et au niveau applicatif. D'un point de vue théorique, l'approche bayésienne se distingue de l'approche classique fréquentiste en considérant les probabilités comme une extension du raisonnement logique. Les probabilités représentent une mesure de l'incertitude de l'implication entre deux propositions et non la fréquence d'un événement dans une infinité de réalisations. Cet aspect de l'approche bayésienne lui donne une consistance avec le raisonnement logique qui a une conséquence directe sur sa mise en œuvre dans les problèmes d'inférence. En effet, la méthode bayésienne ne donne pas des solutions *ad hoc* mais offre une méthodologie unique. Cette méthodologie se résume ainsi,

1. définir le problème d'inférence logique,
2. construire la distribution *a posteriori* (avec les règles de calcul de probabilités comme la règle de Bayes...) contenant toute l'information sur le paramètre à estimer,
3. choisir les distributions de probabilité intervenant dans l'expression de la densité *a posteriori*,
4. résumer l'information *a posteriori* par une des caractéristiques de la distribution *a posteriori*.

Des techniques de calcul comme l'algorithme EM ou les méthodes MCMC assurent l'implémentation efficace de la méthode bayésienne lorsque l'étape 4 est difficile, voire impossible à mener.

Nous avons essayé d'illustrer la méthodologie bayésienne dans le problème de séparation de sources dans sa généralité. Les problèmes d'inférence pratiques dans les chapitres suivants illustrent mieux la faisabilité et le mérite de cette approche. En particulier, l'introduction des variables cachées, naturellement incorporée par l'approche bayésienne au niveau pratique et technique, va jouer un rôle important pour simplifier l'implémentation des algorithmes de séparation et pour garantir l'identifiabilité du mélange.

Bibliographie

[Amari, 1985] S. Amari. *Differential-Geometrical Methods in Statistics.* Volume 28 of Springer Lecture Notes in Statistics, Springer-Verlag, New York, 1985.

[Attias, 1999] H. Attias. Blind separation of noisy mixture : An EM algorithm for independent factor analysis. *Neural Computation*, 11 : 803–851, 1999.

[Bermond, 2000] O. Bermond. *Méthodes statistiques pour la séparation de sources.* thèse de doctorat, Ecole Nationale Supérieure des Télécommunications, 2000.

[Boyles, 1983] R. A. Boyles. On the convergence of the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 45 : 47–50, 1983.

[Cardoso et Labeld, 1996] J. Cardoso et B. Labeld. Equivariant adaptative source separation. *Signal Processing*, 44 : 3017–3030, 1996.

[Cox, 1946] R. Cox. Probability, frequency and reasonable expectation. *Am. J. Physics*, 14 : 1–13, 1946.

[Cox, 1961] R. Cox. *The Algebra of Probable Inference.* Johns Hopkins University Press, Baltimore, MD, USA, 1961.

[Cox, 1979] R. Cox. On inference and inquiry. In *Proc. Maximum Entropy Formalism Conference, MIT Press*, pages 119–167, 1979.

[De Finetti, 1974] B. De Finetti. *Theory of Probability,* 2 *vols.* Translated by A. Machi and A.F.M. Smith, Wiley, London, 1974.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird et D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39 : 1–38, 1977.

[Feller, 1968] W. Feller. *An introduction to probability theory and its applications, Volume 1, 3rd edition.* Wiley, New York, NY, USA, 1968.

[Geweke, 1989] J. Geweke. Bayesian inference in econometric models using monte carlo integration. *Econometrika*, 57 : 1317–1339, 1989.

[Green, 1990] P. J. Green. Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Trans. Medical Imaging*, 9 (1) : 84–93, mars 1990.

[Hastings, 1970] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57 : 97, janvier 1970.

[Hero et Fessler, 1993] A. O. Hero et J. A. Fessler. Asymptotic convergence properties of EM-type algorithms. Preprints 85-T-21, Dept. of Electrical Engineering and Computer Science, University of Michigan, 1993.

[Kass *et al.*, 1988] M. Kass, A. P. Witkin et D. Terzopoulos. Snakes : Active contour models. *Int. J. Computer Vision*, 1 (4) : 321–331, 1988.

[Kass et Wasserman, 1994] R. E. Kass et L. Wasserman. Formal rules for selecting prior distributions : A review and annotated bibliography. Technical report no. 583, Department of Statistics, Carnegie Mellon University, 1994.

[Lindley, 1980] D. Lindley. Apporximate Bayesian methods (with discussion). In *Bayesian Statistics, J.M. Bernardo et al. editor. Valencia University Press*, pages 223–245, 1980.

[McLachlan et Krishnan, 1997] G. J. McLachlan et T. Krishnan. *The EM Algorithm and Extensions.* Wiley series in probability and statistics. John Wiley and Sons, Inc., 1997.

[Metropolis *et al.*, 1953] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller et E. Teller. Equations of state calculations by fast computing machines. *Journal of chemical physics*, 21 : 1087–1092, juin 1953.

[Meyn et Tweedie, 1993] S. Meyn et R. Tweedie. *Markov chains and stochastic stability.* Springer-Verlag, London, 1993.

[Müller, 1991] P. Müller. A generic approach to posterior integration and Gibbs sampling. Rapport technique 91-09, Purdue Uni. West Lafayette, Indiana, Indiana, 1991.

[Müller, 1992] P. Müller. Alternatives to the Gibbs sampling scheme. Rapport technique, Institute of Statistics and Decision Sciences, Duke Uni., 1992.

[Robert, 1996] C. Robert. *Méthodes de Monte-Carlo par chaînes de Markov.* Economica, Paris, 1996.

[Rodríguez, 1991] C. Rodríguez. Entropic priors. *Tech. rep. Electronic form* `http : omega.albany.edu :8008/entpriors.ps`, 1991.

[Rubinstein, 1981] R. Rubinstein. *Simulation and the Monte Carlo Method.* J. Wiley, New York, 1981.

[Senecal, 2002] S. Senecal. *Méthodes de simulation Monte-Carlo par chaînes de Markov pour l'estimation de modèles. Applications en séparation de sources et en égalisation.* thèse de doctorat, INPG (Grenoble), 2002.

[Snoussi et Mohammad-Djafari, 2000] H. Snoussi et A. Mohammad-Djafari. Bayesian source separation with mixture of Gaussians prior for sources and Gaussian prior for mixture coefficients. In A. Mohammad-Djafari, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 388–406, Gif-sur-Yvette, juillet 2000. Proc. of MaxEnt, Amer. Inst. Physics.

[Snoussi et Mohammad-Djafari, 2002] H. Snoussi et A. Mohammad-Djafari. Information Geometry and Prior Selection. In C. Williams, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 307–327. MaxEnt Workshops, Amer. Inst. Physics, août 2002.

[Tierney, 1994] L. Tierney. Markov chain for exploring posterior distribution. *Annals Statist.*, 22 (4) : 1701–1762, décembre 1994.

[Tierney et Kadane, 1986] L. Tierney et J. Kadane. Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Ass.*, (81) : 82–86, 1986.

[Tierney *et al.*, 1986] L. Tierney, R. Kass et J. Kadane. Approximate marginal densities of nonlinear functions. *Biometrika*, (76) : 425–433, 1986.

[Wu, 1983] C. F. J. Wu. On the convergence of the EM algorithm. *Ann. Statist.*, 11 (1) : 95–103, 1983.

[Zhu et Rohwer, 1995] H. Zhu et R. Rohwer. Bayesian invariant measurements of generalisation. In *Neural Proc. Lett.*, volume 2 (6), pages 28–31, 1995.

Chapitre III

SÉPARATION DE SOURCES MONO-VARIÉES : NON STATIONNARITÉ TEMPORELLE

Dans ce chapitre, on considère le problème de séparation de sources dans le cas d'un mélange bruité instantané. La méthode du maximum de vraisemblance a été considérée dans [Attias, 1999; Bermond, 2000] en modélisant les sources par un mélange de gaussiennes. Nous allons étendre ces travaux à plusieurs niveaux.

1. *En interprétant le mélange de gaussiennes comme étant un modèle hiérarchique, on peut donner aux étiquettes du mélange une structure markovienne afin de tenir compte de la corrélation temporelle des sources. L'effet de ce modèle markovien, au niveau de l'estimation, peut être considéré comme une régularisation pour la classification des sources.*
2. *Dans une approche bayésienne, on peut incorporer des informations a priori sur la matrice de mélange et les autres paramètres intervenant dans la modélisation des sources et du bruit. L'introduction des distributions a priori présente aussi d'autres avantages comme l'élimination de la dégénérescence de la vraisemblance et de la non-identifiabilité du problème de séparation.*
3. *Au niveau algorithmique, nous allons décrire l'implémentation de l'algorithme* EM *en utilisant la procédure de Baum-Welsh [Rabiner et Juang, 1986]. Nous présentons aussi des versions de l'*EM *sous optimales mais moins coûteuses en temps et en mémoire : Viterbi-EM et Gibbs-EM.*

III.1 Introduction

On considère le mélange linéaire instantané bruité :

$$\boldsymbol{x}(t) = \boldsymbol{A}\,\boldsymbol{s}(t) + \boldsymbol{\epsilon}(t),\ t = 1..T \tag{III.1}$$

où $\boldsymbol{x}(t)$ est le vecteur $(m \times 1)$ des observations, $\boldsymbol{s}(t)$ est le vecteur $(n \times 1)$ des sources, $\boldsymbol{\epsilon}(t)$ est un bruit additif blanc gaussien de matrice de covariance $\boldsymbol{R}_\epsilon$ et $\boldsymbol{A}$ est la matrice $(m \times n)$ de mélange.

Seules les observations $\boldsymbol{x}_{1..T}$ sont connues. La présence du bruit fait que le problème de séparation de sources est composé de deux sous-problèmes : reconstruction des sources et identification de la matrice de mélange.

Composition du chapitre

Ce chapitre est organisé en deux parties :

1. Dans la première partie, on définit le problème d'inférence et on décrit la méthodologie bayésienne pour le résoudre. On présente les lois *a priori* des sources, de la matrice de mélange et des paramètres de ces mêmes lois. Les sources suivent un modèle de Markov caché (HMM). Les variables cachées représentent les étiquettes des sources et forment une chaîne de Markov. Conditionnellement à ces étiquettes, les sources sont gaussiennes indépendantes. Cette modélisation est convenable dans la mesure où elle constitue une alternative intéressante à la modélisation non paramétrique et prend en compte une éventuelle corrélation temporelle. Le cas du mélange de gaussiennes (étiquettes i.i.d.) a été étudié dans [Attias, 1999; Bermond, 2000; Snoussi et Mohammad-Djafari, 2000] et représente un cas particulier de la modélisation HMM. L'estimation des variances d'un mélange de gaussiennes par la méthode du maximum de vraisemblance, en observant directement les sources, souffre d'un problème de dégénérescence. En effet, la vraisemblance n'est pas bornée et tend vers l'infini quand les variances tendent vers zero. Une solution proposée dans [Hathaway, 1986] est de contraindre les variances à appartenir à un intervalle strictement positif. Cette contrainte complique la mise en œuvre de l'estimation des variances. Récemment, une solution se basant sur l'approche bayésienne a été proposée dans [Ridolfi et Idier, 1999] afin d'éliminer la dégénérescence de la vraisemblance dans le cas où les sources sont directement observées. Elle consiste à pénaliser la vraisemblance avec un *a priori* inverse gamma. Dans [Snoussi et Mohammad-Djafari, 2001], on montre que cette dégénérescence se produit aussi dans le cas de la séparation de sources et qu'elle peut être éliminée en choisissant un *a priori* inverse gamma pour les variances[1]. Le chapitre (VI) est entièrement consacré à l'étude de ces dégénérescences dans le cas des sources multivariables directement observées (non mélangées) et dans le cas où les sources sont mélangées.
   Les coefficients de la matrice de mélange suivent des lois gaussiennes.

   Concernant l'aspect algorithmique, la structure à variables cachées suggère l'utilisation des algorithmes de restauration-maximisation.

2. La deuxième partie est consacrée à l'implémentation des algorithmes de restauration-maximisation. Nous commençons par établir les équations de ré-estimation de l'algorithme EM en utilisant la procédure de Baum-Welsh et discuter le coût de son implémentation. Ensuite, nous présentons d'autres types d'algorithmes de restauration-maximisation en modifiant l'étape $\boldsymbol{E}$ (*Expectation*) de l'algorithme EM.
   - Les algorithmes V-EM (*Viterbi-EM*) et G-EM (*Gibbs-EM*) : l'étape $\boldsymbol{E}$ est remplacée respectivement par une maximisation et par un échantillonnage. Ces modifications visent à réduire le coût de l'EM dû à la structure temporelle de la chaîne de Markov.

[1] Cette dégénérescence se produit aussi dans le cas où les étiquettes forment une chaîne de Markov.

- Les algorithmes F-V-EM (*Fast-Viterbi-EM*) et F-G-EM (*Fast-Gibbs-EM*) : en reprenant les algorithmes V-EM et G-EM , on introduit une étape de relaxation afin de réduire le coût de calcul dû à la structure spatiale qui provient du mélange.

3. Dans la troisième partie, on étudie les performances numériques des algorithmes proposés.

Placement du travail

L'algorithme EM a été utilisé dans [Attias, 1999; Bermond, 2000] pour séparer des sources modélisées par des mélanges de gaussiennes. Dans ces travaux, on peut montrer que :

- l'algorithme EM n'arrive pas à estimer conjointement les variances des sources et la variance du bruit du fait de la dégénérescence de la vraisemblance.
- Le coût de l'implémentation de l'EM est très important.
- L'algorithme est sensible aux conditions initiales.
- On ne tient pas compte des connaissances *a priori* qu'on peut avoir sur la matrice de mélange et sur les divers paramètres intervenant dans le problème de séparation.

Par rapport à ces travaux, nous avons apporté les contributions suivantes :

- introduire des *a priori* sur les variances afin d'éliminer la dégénérescence mentionnée plus haut,
- introduire un *a priori* sur $\boldsymbol{A}$ afin d'exprimer d'éventuelles connaissances sur les coefficients du mélange,
- donner une structure markovienne aux étiquettes du mélange (régulariser la classification des sources),
- proposer des algorithmes de séparation moins optimaux mais plus rapides.

III.2 Méthodologie bayésienne

III.2.1 Distribution A Posteriori

Dans ce chapitre, le problème d'inférence est $\mathcal{I} := (\boldsymbol{x}_{1..T} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{A})$. Autrement dit, notre objectif est l'inférence sur la matrice de mélange $\boldsymbol{A}$ connaissant les données $\boldsymbol{x}_{1..T}$ observées et toute l'information *a priori* $\boldsymbol{I}$ qu'on possède sur le problème. L'information *a priori* $\boldsymbol{I}$ indique par exemple que les données $\boldsymbol{x}_{1..T}$ sont liées à la matrice $\boldsymbol{A}$ par le modèle (III.1) ainsi que les formes des lois choisies pour toutes les variables qui vont intervenir dans le problème d'inférence.

La distribution *a posteriori* de la matrice $\boldsymbol{A}$ (degré d'incertitude de la proposition $\mathcal{I}$) s'écrit, selon la règle de Bayes,

$$\mathrm{p}(\boldsymbol{A} \mid \boldsymbol{x}_{1..T}, \boldsymbol{I}) \propto \mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{A}, \boldsymbol{I})\, \mathrm{p}(\boldsymbol{A} \mid \boldsymbol{I})$$

où $\mathrm{p}(\boldsymbol{A} \mid \boldsymbol{I})$ est la distribution *a priori* de la matrice de mélange. $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{A}, \boldsymbol{I})$ est la vraisemblance de $\boldsymbol{A}$. Selon le modèle de mélange (III.1), la vraisemblance peut se mettre sous une forme marginale qui fait apparaître les lois *a priori* du bruit $\boldsymbol{\epsilon}_{1..T}$ et des sources $\boldsymbol{s}_{1..T}$:

$$\begin{aligned}
\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{A}, \boldsymbol{I}) &= \int \mathrm{p}(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T} \mid \boldsymbol{A}, \boldsymbol{I})\, \mathrm{d}\boldsymbol{s}_{1..T} \\
&= \int \mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{s}_{1..T}, \boldsymbol{A}, \boldsymbol{I})\, \mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{A}, \boldsymbol{I})\, \mathrm{d}\boldsymbol{s}_{1..T} \\
&= \int \underbrace{\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{s}_{1..T}, \boldsymbol{A}, \boldsymbol{I})}_{\text{loi du bruit } \boldsymbol{\epsilon}_{1..T}} \underbrace{\mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{I})}_{\text{loi des sources } \boldsymbol{s}_{1..T}} \mathrm{d}\boldsymbol{s}_{1..T}
\end{aligned}$$

Le fait d'estimer la matrice de mélange $\boldsymbol{A}$ et non son inverse présente au moins deux avantages : (**i**) $\boldsymbol{A}$ n'est pas forcément carrée ($n \neq m$), (**ii**) naturellement, on possède une information *a priori* sur $\boldsymbol{A}$ et non sur son inverse (qui peut ne pas exister !).

On choisit les lois de probabilités appartenant à des familles paramétriques :

$$\begin{cases} \text{bruit } \boldsymbol{\epsilon}_{1..T} & \longrightarrow \quad \mathrm{p}_\epsilon(. \mid \boldsymbol{\eta}_\epsilon) \\ \text{sources } \boldsymbol{s}_{1..T} & \longrightarrow \quad \mathrm{p}_s(. \mid \boldsymbol{\eta}_s) \\ \text{matrice } \boldsymbol{A} & \longrightarrow \quad \mathrm{p}_A(. \mid \boldsymbol{\eta}_a) \end{cases}$$

Le paramètre $\boldsymbol{\eta} = (\boldsymbol{\eta}_\epsilon, \boldsymbol{\eta}_s, \boldsymbol{\eta}_a)$ n'est pas en général connu. La distribution *a posteriori* de la matrice de mélange s'écrit (on a omis l'information *a priori* $\boldsymbol{I}$ pour alléger l'expression) :

$$\mathrm{p}(\boldsymbol{A} \mid \boldsymbol{x}_{1..T}) \propto \int_{\boldsymbol{\eta}} \left\{ \left[\int_{\boldsymbol{s}_{1..T}} \mathrm{p}_\epsilon(\boldsymbol{x}_{1..T} - \boldsymbol{A}\,\boldsymbol{s}_{1..T} \mid \boldsymbol{\eta}_\epsilon) \mathrm{p}_s(\boldsymbol{s}_{1..T} \mid \boldsymbol{\eta}_s)\, \mathrm{d}\boldsymbol{s}_{1..T} \right] \mathrm{p}_A(\boldsymbol{A} \mid \boldsymbol{\eta}_a) \right\} \mathrm{p}(\boldsymbol{\eta})\, \mathrm{d}\boldsymbol{\eta} \qquad \text{(III.2)}$$

où $\mathrm{p}(\boldsymbol{\eta})$ est la loi *a priori* des paramètres des distributions *a priori*. On doit aussi choisir cette distribution.

On note que l'expression (III.2) nécessite deux intégrations. Une intégration pour marginaliser par rapport aux sources et une intégration pour marginaliser par rapport au paramètre $\boldsymbol{\eta}$. Dans la suite, on modifie le problème d'inférence en incluant le paramètre $\boldsymbol{\eta}$ parmi les paramètres d'intérêt à identifier :

$$\mathcal{I} := (\boldsymbol{x}_{1..T} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{A}) \quad \rightsquigarrow \quad \mathcal{I} := (\boldsymbol{x}_{1..T} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{A} \wedge \boldsymbol{\eta})$$

La distribution *a posteriori* du paramètre $(\boldsymbol{A}, \boldsymbol{\eta})$ ne contient plus qu'une seule intégration par rapport aux sources :

$$\mathrm{p}(\boldsymbol{A}, \boldsymbol{\eta} \mid \boldsymbol{x}_{1..T}) \propto \left[\int_{\boldsymbol{s}_{1..T}} \mathrm{p}_\epsilon(\boldsymbol{x}_{1..T} - \boldsymbol{A}\,\boldsymbol{s}_{1..T} \mid \boldsymbol{\eta}_\epsilon) \mathrm{p}_s(\boldsymbol{s}_{1..T} \mid \boldsymbol{\eta}_s)\, \mathrm{d}\boldsymbol{s}_{1..T} \right] \mathrm{p}_A(\boldsymbol{A} \mid \boldsymbol{\eta}_a) \mathrm{p}(\boldsymbol{\eta}) \qquad \text{(III.3)}$$

III.2.2 Choix des lois de probabilité

[A] Distribution des sources

Chaque composante source s^j suit un modèle de Markov caché. Ce modèle peut être interprété comme un processus doublement stochastique :

1. un processus stochastique continu $(s_1^j, s_2^j, ..., s_T^j)$ à valeurs dans $\mathbb{R}$,
2. un processus stochastique discret caché $(z_1^j, z_2^j, ..., z_T^j)$ à valeurs dans $\{1..K_j\}$.

La suite $(z_t^j)_{t=1..T}$ forme une chaîne de Markov homogène de distribution initiale $\left[p_l = P(z_1^j = l)\right]_{l=1..K_j}$ et de matrice de transition $P_{lk} = \left[P(z_{t+1}^j = k \mid z_t^j = l)\right]_{l,k=1..K_j}$. Conditionnellement à cette chaîne, la source s^j est temporellement blanche :

$$p(s_{1..T}^j \mid z_{1..T}^j) = \prod_{t=1}^{T} p(s_t^j \mid z_t^j) \qquad \text{(III.4)}$$

avec une distribution gaussienne $p(s_t^j \mid z_t^j = l) = \mathcal{N}(m_{jl}, \sigma_{jl})$.

Cette modélisation présente plusieurs avantages. Parmi lesquels, on peut citer :

- Elle appartient à une famille paramétrique. Par conséquent, la possibilité d'estimer ses paramètres la rend flexible et applicable dans les situations réelles. Sa structure cachée, similaire à la structure cachée du problème de séparation de sources, facilite l'intégration de l'identification de ses paramètres dans les algorithmes de séparation.

- Elle présente une bonne alternative à la modélisation non paramétrique. En effet, en augmentant le nombre d'états K_j de la chaîne de Markov cachée, on peut atteindre n'importe quelle distribution de probabilité.
- Elle garantit l'identifiabilité (ou facilite son étude) de la matrice de mélange.

Des modèles de HMM plus élaborés peuvent être trouvés dans [Ghahramani et Jordan, 1997].

La loi *a priori* de la $j^{\text{ème}}$ source s'écrit alors :

$$\mathrm{p}_s(s^j_{1..T} \mid \boldsymbol{\eta}^j_s) = \sum_{z^j_{1..T}} Pr(z^j_{1..T} \mid \boldsymbol{\eta}^j_p) \prod_{t=1}^{t=T} p(s^j_t \mid z^j_t, \boldsymbol{\eta}^j_g)$$

où on a décomposé le paramètre $\boldsymbol{\eta}^j_s = (\boldsymbol{\eta}^j_p, \boldsymbol{\eta}^j_g)$ avec $\boldsymbol{\eta}^j_p$ contenant la probabilité initiale et la matrice de transition de la chaîne $\boldsymbol{z}^j_{1..T}$ et $\boldsymbol{\eta}^j_g$ contenant les moyennes et variances des gaussiennes.

[B] Modélisation de la matrice de mélange

Pour la matrice de mélange, on choisit une distribution gaussienne :

$$p(\boldsymbol{A}_{ij}) = \mathcal{N}(\boldsymbol{M}_{ij}, \sigma^2_{a,ij}). \tag{III.5}$$

Ce choix est motivé par les raisons suivantes :

- On peut interpréter facilement cette loi : on connaît la valeur M_{ij} du coefficient A_{ij} de la matrice de mélange avec une incertitude $\sigma^2_{a,ij}$.
- La distribution gaussienne est un *a priori* conjugué de la vraisemblance complète $\mathrm{p}(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T} \mid \boldsymbol{A})$ dans le cas d'un bruit gaussien (ce qui va être supposé dans la suite de ce chapitre). L'*a priori* conjugué garantit (par sa définition) que la distribution *a posteriori* reste dans la même famille que la distribution *a priori*.
- L'*a priori* conjugué trouve une justification basée sur la géométrie de l'information. Le chapitre (VII) est entièrement consacré au choix de l'*a priori* avec les outils de la géométrie de l'information.

Exemple 3 *Dans certaines applications [Snoussi et al., 2001], on connaît a priori certains éléments de la matrice de mélange. Au lieu de fixer ces éléments, on peut leur attribuer des distributions gaussiennes avec des moyennes M_{ij} égales aux valeurs connues et des variances $\sigma^2_{a,ij}$ très faibles.*

Cependant, en prenant $\boldsymbol{M}_{ij} = 0$ et des valeurs très grandes pour $\sigma^2_{a,ij}$, on s'approche du cas classique où on ne possède pas d'information a priori sur la valeur du coefficient A_{ij}.

[C] Distribution *a priori* des variances

On attribue un *a priori* gamma inverse $\mathcal{IG}(a, b)$ ($a > 0$ and $b > 1$) pour les variances du bruit et des composantes gaussiennes de l'*a priori* des sources. On montre dans [Snoussi et Mohammad-Djafari, 2001] que cette *a priori* est nécessaire pour éliminer la dégénérescence de la vraisemblance quand l'une des variances tend vers zero (ou l'une des matrices de covariance tend vers une matrice singulière). Le chapitre (VI) est entièrement consacré à l'étude de cette dégénérescence et à la manière de l'éliminer.

III.2.3 Coût d'estimation et interprétation du critère

En prenant la distance $d(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 1 - \delta_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*)$, la minimisation du coût moyen d'estimation donne le MAP comme estimateur :

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T})$$

avec $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{\eta})$.

L'introduction des variables cachées $\boldsymbol{z}_{1..T}$ rend le problème doublement caché nécessitant ainsi deux intégrations dans l'expression de la distribution *a posteriori* de $\boldsymbol{\theta}$: une intégration par rapport à $\boldsymbol{s}_{1..T}$ et une intégration (sommation) par rapport à $\boldsymbol{z}_{1..T}$:

$$\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}) \propto \left[\sum_{\boldsymbol{z}_{1..T}} \left\{ \int \mathrm{p}(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T} \mid \boldsymbol{z}_{1..T}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{s}_{1..T} \right\} Pr(\boldsymbol{z}_{1..T}) \right] \mathrm{p}(\boldsymbol{\theta}) \tag{III.6}$$

D'après les modèles choisis pour les sources et pour le bruit, on peut intégrer analytiquement par rapport aux sources connaissant les étiquettes $\boldsymbol{z}_{1..T}$:

$$\begin{aligned} \mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{z}_{1..T}, \boldsymbol{\theta}) &= \int_T \mathrm{p}(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T} \mid \boldsymbol{z}_{1..T}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{s}_{1..T} \\ &= \prod_{t=1}^{T} \int \mathrm{p}(\boldsymbol{x}_t, \boldsymbol{s}_t \mid \boldsymbol{z}_t, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{s}_t \\ &= \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{x}_t \; ; \; \boldsymbol{A}\boldsymbol{m}_{\boldsymbol{k}}, \; \boldsymbol{A}\boldsymbol{\Gamma}_{\boldsymbol{k}}\boldsymbol{A}^* + \boldsymbol{R}_\epsilon) \Big|_{\boldsymbol{k}=\boldsymbol{z}_t} \end{aligned} \tag{III.7}$$

où $\boldsymbol{m}_{\boldsymbol{k}}$ et $\boldsymbol{\Gamma}_{\boldsymbol{k}}$ contiennent les moyennes et les variances *a priori* et $\boldsymbol{k}$ est l'étiquette vectorielle :

$$\boldsymbol{k} = \begin{pmatrix} k_1 \\ \vdots \\ k_n \end{pmatrix}, \quad \begin{matrix} k_1 = 1..K_1 \\ \vdots \\ k_n = 1..K_n \end{matrix}, \quad \boldsymbol{m}_{\boldsymbol{k}} = \begin{pmatrix} m_{k_1} \\ \vdots \\ m_{k_n} \end{pmatrix}, \quad \boldsymbol{\Gamma}_{\boldsymbol{k}} = \begin{pmatrix} \sigma^2_{k_1} & \dots & \\ \vdots & \ddots & \\ & & \sigma^2_{k_n} \end{pmatrix}.$$

Le processus $\boldsymbol{z}_{1..T}$ peut être interprété comme un processus de classification. La vraisemblance (III.7) de $\boldsymbol{\theta}$ connaissant une classification particulière $\boldsymbol{z}_{1..T}$ peut être ré-écrite en réarrangeant les termes selon les classes auxquels ils appartiennent et en définissant les ensembles $\mathcal{T}_k = \{t \mid \boldsymbol{z}_t = \boldsymbol{k}\}$:

$$\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{z}_{1..T}, \boldsymbol{\theta}) = \prod_{k=1}^{K} |2\Pi \boldsymbol{R}_k|^{-\frac{T_k}{2}} \exp\left[-\frac{1}{2} \mathrm{Tr}\left(\boldsymbol{R}_k^{-1} \sum_{t \in \mathcal{T}_k} (\boldsymbol{x}_t - \boldsymbol{A}\boldsymbol{m}_k)(\boldsymbol{x}_t - \boldsymbol{A}\boldsymbol{m}_k)^* \right) \right] \tag{III.8}$$

où $T_k = |\mathcal{T}_k|$ est le cardinal de la classe $\mathcal{T}_k$ et $\boldsymbol{R}_k = \boldsymbol{A}\boldsymbol{\Gamma}_{\boldsymbol{k}}\boldsymbol{A}^* + \boldsymbol{R}_\epsilon$ la matrice de covariance de $\boldsymbol{x}$ conditionnellement à la classe $\boldsymbol{k}$.

Afin de faciliter l'interprétation du critère, on suppose dans ce paragraphe que les moyennes $\boldsymbol{m}_k$ sont nulles. L'opposé du logarithme de la vraisemblance (III.8) normalisé se met, à une constante additive près, sous la forme d'une somme pondérée de divergences de Kullback-Leibler entre les matrices de covariance théoriques $\boldsymbol{R}_k$ et les matrices de covariance empiriques $\hat{\boldsymbol{R}}_k = \sum_{\mathcal{T}_k} \boldsymbol{x}_t \boldsymbol{x}_t^* / T_k$,

$$\begin{aligned} -\mathcal{L}_T(\boldsymbol{\theta} \mid \boldsymbol{z}_{1..T}, \boldsymbol{x}_{1..T}) &= \frac{-\log \mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{z}_{1..T}, \boldsymbol{\theta})}{T} \\ &= -\sum_{k=1}^{K} \alpha_k \left(\frac{1}{2} \log |\boldsymbol{R}_k^{-1} \hat{\boldsymbol{R}}_k| - \frac{1}{2} \mathrm{Tr}\left(\boldsymbol{R}_k^{-1} \hat{\boldsymbol{R}}_k \right) + \frac{m}{2} \right) + cte \\ &= \sum_{k=1}^{K} \alpha_k D_{KL}(\boldsymbol{R}_k, \hat{\boldsymbol{R}}_k) + cte \end{aligned}$$

où les α_k représentent les proportions des classes.

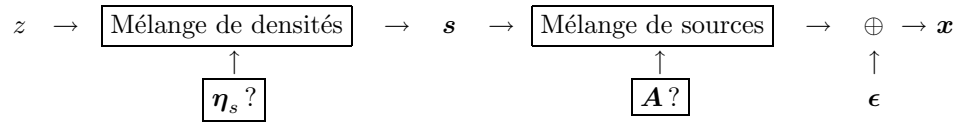
L'opposé du logarithme normalisé de la distribution *a posteriori* de $\boldsymbol{\theta}$ connaissant $\boldsymbol{z}_{1..T}$ est alors une forme régularisée du critère d'ajustement des matrices de covariance (statistiques d'ordre deux) :

$$\frac{-\log \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{z}_{1..T})}{T} = \underbrace{\sum_{k=1}^{K} \alpha_k D_{KL}(\boldsymbol{R}_k, \hat{\boldsymbol{R}}_k)}_{\text{Ajustement des covariances}} - \underbrace{\log \mathrm{p}(\boldsymbol{\theta}) \,/\, T}_{\text{Terme de régularisation}}$$

Le critère se met sous la forme d'un ajustement de matrices de covariance connaissant la classification. La distribution *a posteriori* $\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T})$ est par conséquent interprétée comme un moyennage d'ajustements de statistiques d'ordre deux relatifs à toutes les classifications possibles. On peut aussi changer le problème d'inférence en $\mathcal{I} := (\boldsymbol{x}_{1..T} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{A} \wedge \boldsymbol{\eta} \wedge \boldsymbol{z}_{1..T})$. Autrement dit, on effectue conjointement la classification et l'ajustement des statistiques d'ordre deux. Le modèle markovien sur la chaîne $\boldsymbol{z}_{1..T}$ est une régularisation de l'opération de classification.

III.3 Algorithmes de restauration-maximisation

Les sources $(\boldsymbol{s}_t)_{t=1..T}$, n'étant pas directement observées, forment une deuxième couche de variables cachées. La première couche est formée par les étiquettes $(z_t^j)_{t=1..T}$ des mélanges de densités. Le problème de séparation de sources contient donc deux opérations de mélange : (**i**) un mélange de densités qui est une représentation mathématique de la distribution *a priori* avec des paramètres inconnus $\boldsymbol{\eta}_s$, (**ii**) un mélange réel physique de sources avec une matrice inconnue $\boldsymbol{A}$.



Nous avons un problème à données incomplètes. Les observations $\boldsymbol{x}_{1..T}$ sont les données incomplètes. Les sources $\boldsymbol{s}_{1..T}$ et les étiquettes $\boldsymbol{z}_{1..T}$ sont les données manquantes. Les paramètres à estimer sont $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{\eta})$ avec $\boldsymbol{\eta}$ contenant les paramètres inconnus des lois de probabilité intervenant dans le problème de séparation. La structure à variables cachées suggère l'utilisation des algorithmes de restauration-maximisation dont le principe est le suivant : partant d'un point initial $\tilde{\boldsymbol{\theta}}^{(0)}$, la mise à jour, à l'itération k, de $\tilde{\boldsymbol{\theta}}^{(k)}$ en $\tilde{\boldsymbol{\theta}}^{(k+1)}$ s'effectue en deux étapes :

1. **Restauration** : dans cette étape, connaissant la valeur de $\tilde{\boldsymbol{\theta}}^{(k-1)}$, on attribue à toute fonction $f(\boldsymbol{s}, \boldsymbol{z})$ des variables manquantes intervenant dans l'expression du logarithme de la vraisemblance complète une valeur f^k.
2. **Maximisation** : $\boldsymbol{\theta}^{k+1}$ est alors la valeur qui maximise la vraisemblance complète pénalisée $\log p(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{z} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$.

Il existe plusieurs stratégies de restauration :

1. f^k est l'espérance de $f(\boldsymbol{s}, \boldsymbol{z})$ conditionnellement à la valeur courante $\boldsymbol{\theta}^{(k-1)}$ estimée à l'itération précédente :
$$f^k = \int_{\boldsymbol{s},\boldsymbol{z}} f(\boldsymbol{s}, \boldsymbol{z})\, p(\boldsymbol{s}, \boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(k-1)})\, d\,\boldsymbol{s}\, d\,\boldsymbol{z}. \tag{III.9}$$
C'est exactement le principe de l'algorithme EM [Dempster *et al.*, 1977]. La propriété fondamentale de cet algorithme est qu'il assure la croissance monotone de la distribution *a posteriori* incomplète. Toute valeur $\boldsymbol{\theta}$ faisant croître l'espérance du logarithme de la distribution *a posteriori* complète, fait

aussi croître le logarithme de la distribution *a posteriori* incomplète. En plus, un point critique de la distribution *a posteriori* incomplète est un point fixe de la transformation associée à l'algorithme EM. Plus de détails sur les propriétés de l'algorithme EM sont donnés dans le chapitre (II) ou [McLachlan et Krishnan, 1997].

2. Les variables cachées sont remplacées par leur maximum *a posteriori*. La distribution *a posteriori* est construite connaissant le paramètre $\boldsymbol{\theta}^{(k-1)}$ et les données $\boldsymbol{x}_{1..T}$. D'après les modélisations choisies dans la section précédente, la distribution *a posteriori* des sources est une gaussienne dont on sait calculer analytiquement la moyenne et la covariance. On envisage alors d'estimer d'abord les étiquettes $\boldsymbol{z}_{1..T}$ (la classification) et puis, comme dans l'algorithme EM, remplacer toute fonction de $\boldsymbol{s}$ par son espérance *a posteriori*.
3. On suit le même schéma que la stratégie précédente mais, au lieu de résumer la distribution *a posteriori* des étiquettes par son maximum, on l'échantillonne. Cet algorithme est une version hybride de l'algorithme EM et de l'algorithme SEM (Stochastic EM [Celeux et Diebolt, 1985]). En effet, c'est un algorithme EM (vis-à-vis des sources $\boldsymbol{s}_{1..T}$) et un algorithme SEM (vis-à-vis des étiquettes $\boldsymbol{z}_{1..T}$).

Dans la suite, nous allons détailler ces trois stratégies.

III.3.1 Algorithme EM exact

La fonctionnelle $\mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^k) = \mathrm{E}\left[\log p(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{z} \,|\, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \,|\, \boldsymbol{x}, \boldsymbol{\theta}^k\right]$, calculée dans la première étape de l'algorithme EM, est séparable en trois fonctionnelles $\mathcal{Q}_a$, $\mathcal{Q}_{\boldsymbol{\eta}_g}$ et $\mathcal{Q}_{\eta_p}$,

$$\mathcal{Q} = \mathcal{Q}_a + \mathcal{Q}_{\boldsymbol{\eta}_g} + \mathcal{Q}_{\eta_p}.$$

- La première fonctionnelle $\mathcal{Q}_a$ dépend de $\boldsymbol{A}$ et $\boldsymbol{R}_\epsilon$.
- La deuxième fonctionnelle $\mathcal{Q}_{\boldsymbol{\eta}_g}$ dépend de $\boldsymbol{\eta}_g = (m_{lk}, \sigma_{lk})_{l=1..n, k=1..K_l}$: moyennes et variances des mélanges de densités.
- La troisième fonctionnelle $\mathcal{Q}_{\eta_p}$ dépend de $\eta_p = (\boldsymbol{p}_l, \boldsymbol{P}_l)_{l=1..n}$: probabilités initiales et matrices de transitions des chaînes de Markov.

Maximisation de $\mathcal{Q}_a$: La fonctionnelle à maximiser à chaque itération est :

$$\begin{aligned} \mathcal{Q}(\boldsymbol{A}, \boldsymbol{R}_\epsilon|\boldsymbol{\theta}^0) \;=\; & -\tfrac{T}{2}\log|2\pi\boldsymbol{R}_\epsilon| - \tfrac{T}{2}\mathrm{Tr}\left(\boldsymbol{R}_\epsilon^{-1}(\boldsymbol{R}_{xx} - \boldsymbol{A}\boldsymbol{R}_{sx} - \boldsymbol{R}_{sx}^*\boldsymbol{A}^* + \boldsymbol{A}\boldsymbol{R}_{ss}\boldsymbol{A}^*)\right) \\ & + \log p(\boldsymbol{A}) + \log p(\boldsymbol{R}_\epsilon) \end{aligned} \tag{III.10}$$

où $(*)$ désigne la transposé d'une matrice.

En définissant les statistiques suivantes :

$$\begin{cases} \boldsymbol{R}_{xx} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t\boldsymbol{x}_t^* \\ \boldsymbol{R}_{sx} = \frac{1}{T}\sum_{t=1}^{T} E[\boldsymbol{s}_t|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0]\boldsymbol{x}_t^* \\ \boldsymbol{R}_{ss} = \frac{1}{T}\sum_{t=1}^{T} E[\boldsymbol{s}_t\boldsymbol{s}_t^*|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0] \end{cases} \tag{III.11}$$

la mise à jour de $\boldsymbol{A}$ et $\boldsymbol{R}_\epsilon$ devient :

$$\begin{cases} \mathbf{Vec}(\boldsymbol{A}^{(k+1)}) = \left[T\widehat{\boldsymbol{R}}_{ss}^* \otimes \boldsymbol{R}_\epsilon^{-1} + diag(Vec(\boldsymbol{\Gamma}))\right]^{-1} Vec(T\boldsymbol{R}_\epsilon^{-1}\widehat{\boldsymbol{R}}_{xs} + \boldsymbol{\Gamma} \odot \boldsymbol{M}) \\ \boldsymbol{R}_\epsilon^{(k+1)} = \boldsymbol{R}_{xx} - \boldsymbol{A}^{(k+1)}\boldsymbol{R}_{sx} - \boldsymbol{R}_{xs}(\boldsymbol{A}^{(k+1)})^* + \boldsymbol{A}^{(k+1)}\boldsymbol{R}_{ss}(\boldsymbol{A}^{(k+1)})^* \end{cases} \tag{III.12}$$

où $\otimes$ est le produit de Kronecker [Brewer, 1978], $\odot$ est le produit terme à terme de deux matrices, $\mathbf{Vec}(.)$ est la présentation vectorielle d'une matrice et $\boldsymbol{\Gamma}$ est la matrice $(1/\sigma_{a,ij}^2)$.

On doit alors calculer les espérances conditionnelles $E[\boldsymbol{s}_t|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0]$ et $E[\boldsymbol{s}_t\boldsymbol{s}_t^*|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0]$. En général,

$$E[f(\boldsymbol{s}_t)|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0] = \sum_{\boldsymbol{i}} E[f(\boldsymbol{s}_t)|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0, \boldsymbol{z}_t = \boldsymbol{i}]\ p(\boldsymbol{z}_t = \boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0) \tag{III.13}$$

Le vecteur $\boldsymbol{i} = [i_1, ..., i_n]$ appartient à $\mathcal{Z}_1 \times \mathcal{Z}_2 \times \ldots \mathcal{Z}_n$ avec $\mathcal{Z}_l = \{1..K_l\}$. K_l est le nombre de gaussiennes de la distribution de la $l^{\text{ème}}$ source. On a donc $K = \prod_{l=1}^{n} K_l$ éléments dans la somme (III.13).

Connaissant la variable $\boldsymbol{z}_t = \boldsymbol{i}$, les espérances *a posteriori* sont facilement calculées :

$$\begin{cases} E[\boldsymbol{s}_t|\boldsymbol{x}_t, \boldsymbol{\theta}^0, \boldsymbol{z}_t = \boldsymbol{i}] = [\boldsymbol{A}^*\boldsymbol{R}_\epsilon^{-1}\boldsymbol{A} + \boldsymbol{\Gamma}_i^{-1}]^{-1}[\boldsymbol{A}^*\boldsymbol{R}_\epsilon^{-1}\boldsymbol{x}_t + \boldsymbol{\Gamma}_i^{-1}\boldsymbol{m}_i] = M_{ti}. \\ E[\boldsymbol{s}_t\boldsymbol{s}_t^*|\boldsymbol{x}_t, \boldsymbol{\theta}^0, \boldsymbol{z}_t = \boldsymbol{i}] = [\boldsymbol{A}^*\boldsymbol{R}_\epsilon^{-1}\boldsymbol{A} + \boldsymbol{\Gamma}_i^{-1}]^{-1} + M_{ti}M_{ti}^*. \end{cases} \tag{III.14}$$

Cependant, le coût de calcul des probabilités $p(\boldsymbol{z}_t = \boldsymbol{i} \mid \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)$ en tant que probabilités marginales de $p(\boldsymbol{z}_{1..T}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)$ est très élevé. La procédure de Baum-Welsh [Rabiner et Juang, 1986] peut être étendue au cas où les sources ne sont pas directement observées. On définit les variables $\mathcal{F}_t(\boldsymbol{i})$ (Forward) et les variables $\mathcal{B}_t(\boldsymbol{i})$ (Backward) par :

$$\begin{cases} \mathcal{F}_t(\boldsymbol{i}) &= P(\boldsymbol{z}_t = \boldsymbol{i} \,|\, \boldsymbol{x}_{1..t},\, \boldsymbol{\theta}) \\ \mathcal{B}_t(\boldsymbol{i}) &= \dfrac{p(\boldsymbol{x}_{t+1..T} \,|\, \boldsymbol{z}_t = \boldsymbol{i},\, \boldsymbol{\theta})}{p(\boldsymbol{x}_{t+1..T} \,|\, \boldsymbol{x}_{1..t},\, \boldsymbol{\theta})} \end{cases} \tag{III.15}$$

Le calcul de ces variables est réalisé par les formules de récurrrence suivantes :

$$\begin{cases} \mathcal{F}_1(\boldsymbol{i}) = M_1\, p_{\boldsymbol{i}}\ \mathcal{N}_{(\boldsymbol{A}\boldsymbol{m}_i, \boldsymbol{A}\boldsymbol{\Gamma}_i\boldsymbol{A}^* + \boldsymbol{R}_\epsilon)}[\boldsymbol{x}_1] \\ \mathcal{F}_t(\boldsymbol{i}) = M_t\ \sum_{\boldsymbol{j}} \mathcal{F}_{t-1}(\boldsymbol{j})\ P_{ji}\ \mathcal{N}_{(\boldsymbol{A}\boldsymbol{m}_i, \boldsymbol{A}\boldsymbol{\Gamma}_i\boldsymbol{A}^* + \boldsymbol{R}_\epsilon)}[\boldsymbol{x}_t] \end{cases} \tag{III.16}$$

$$\begin{cases} \mathcal{B}_T(\boldsymbol{i}) = 1 \\ \mathcal{B}_t(\boldsymbol{i}) = M_{t+1} \sum_{\boldsymbol{j}} \mathcal{B}_{t+1}(\boldsymbol{j}) P_{ij} \mathcal{N}_{(\boldsymbol{A}\boldsymbol{m}_j, \boldsymbol{A}\boldsymbol{R}_j\boldsymbol{A}^* + \boldsymbol{R}_\epsilon)}[\boldsymbol{x}_{t+1}] \end{cases}$$

où M_t est une constante de normalisation :

$$\begin{cases} M_1 = \left[\sum_{\boldsymbol{i}} p_{\boldsymbol{i}} \mathcal{N}_{(\boldsymbol{A}\boldsymbol{m}_i, \boldsymbol{A}\boldsymbol{\Gamma}_i\boldsymbol{A}^* + \boldsymbol{R}_\epsilon)}[\boldsymbol{x}_1]\right]^{-1} \\ M_t = \left[\sum_{\boldsymbol{i}} \sum_{\boldsymbol{j}} \mathcal{F}_{t-1}(\boldsymbol{j}) P_{ji} \mathcal{N}_{(\boldsymbol{A}\boldsymbol{m}_i, \boldsymbol{A}\boldsymbol{\Gamma}_i\boldsymbol{A}^* + \boldsymbol{R}_\epsilon)}[\boldsymbol{x}_t]\right]^{-1} \end{cases}$$

et

$$\boldsymbol{m}_i = \begin{pmatrix} m_{i_1} \\ \vdots \\ m_{i_n} \end{pmatrix} \quad , \quad \boldsymbol{\Gamma}_i = \begin{pmatrix} \sigma_{i_1}^2 & 0 & \ldots & 0 \\ 0 & \sigma_{i_2}^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \\ \ldots & & & \sigma_{i_n}^2 \end{pmatrix}$$

Les quantités $p(\boldsymbol{z}_t = \boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)$ sont alors simplement obtenues par :

$$p(\boldsymbol{z}_t = \boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0) = \mathcal{F}_t(\boldsymbol{i})\mathcal{B}_t(\boldsymbol{i})$$

L'indépendance spatiale des sources ou plus précisément celle des étiquettes implique :

$$\begin{cases} p_{\boldsymbol{i}} = \prod_{l=1}^{n} p_{i_l} = p_{i_1} \times p_{i_2} \ldots p_{i_n} \\ P_{\boldsymbol{ij}} = \prod_{l=1}^{n} P_{i_l j_l}^l \end{cases}$$

où p_{i_l} est la probabilité initiale de la chaîne de Markov de la source l et P^l est sa matrice de transition.

La complexité de la procédure Forward-Backward est de l'ordre de $K^2\,T$ avec $K = \prod_{l=1}^{n} K_l$ le nombre des étiquettes vectorielles. Si on choisit le même nombre $K_l = k$ de gaussiennes pour toutes les sources, la complexité $k^{2*n}\,T$ croît exponentiellement avec le nombre de sources.

Maximisation de $\mathcal{Q}_{\boldsymbol{\eta}_g}$: Afin d'établir la connection avec l'estimation des paramètres d'un modèle de Markov caché quand les sources sont directement observées et afin d'élucider le coût de calcul important de la ré-estimation des hyperparamètres, on commence par établir les formules pour le cas vectoriel suivi du cas scalaire qui nous intéresse.

Le vecteur $\boldsymbol{i}$ désigne l'étiquette vectorielle $(i_1, i_2 ... i_n)$. Le vecteur $\boldsymbol{m_i}$ désigne $(m_{i_1}, m_{i_2} ... m_{i_n})^*$. $\boldsymbol{\Gamma_i}$ désigne la matrice diagonale $diag\left(\sigma^2_{i_1}, \sigma^2_{i_2} \ldots \sigma^2_{i_n}\right)$.

La ré-estimation des moyennes vectorielles et des covariances donne :

$$\left\{\begin{array}{l} \boldsymbol{m_i} = \frac{\sum_{t=1}^{T} E[\boldsymbol{s}_t|\boldsymbol{x}_t, \boldsymbol{z}_t=\boldsymbol{i}, \boldsymbol{\theta}^0]\, P(\boldsymbol{z}_t=\boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)}{\sum_{t=1}^{T} P(\boldsymbol{z}_t=\boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)} \\ \\ \boldsymbol{\Gamma_i} = \frac{\sum_{t=1}^{T} [E(\boldsymbol{s}_t \boldsymbol{s}_t^*) - M_{t\boldsymbol{i}} \boldsymbol{m}_{\boldsymbol{i}}^* - \boldsymbol{m_i} M_{t\boldsymbol{i}}^* + \boldsymbol{m_i} \boldsymbol{m}_{\boldsymbol{i}}^*]\, P(\boldsymbol{z}_t=\boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0) + 2b\, \boldsymbol{I}}{\sum_{t=1}^{T} P(\boldsymbol{z}_t=\boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0) + 2\,(a-1)} \end{array}\right. \tag{III.17}$$

où $M_{t\boldsymbol{i}} = E[\boldsymbol{s}_t|\boldsymbol{x}_t, \boldsymbol{z}_t = \boldsymbol{i}, \boldsymbol{\theta}^0]$.

La ré-estimation des moyennes scalaires et des variances est obtenue par une marginalisation spatiale sur les étiquettes vectorielles des expressions (III.17) :

$$\left\{\begin{array}{l} m_{lk} = \frac{\sum_{t=1}^{T} \sum_{(\boldsymbol{i}|\boldsymbol{i}(l)=k)} [E(\boldsymbol{s}_t|\boldsymbol{x}_t, \boldsymbol{z}_t = \boldsymbol{i}, \boldsymbol{\theta}^0)]_l\, P(\boldsymbol{z}_t = \boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)}{\sum_{t=1}^{T} \sum_{(\boldsymbol{i}|\boldsymbol{i}(l)=k)} P(\boldsymbol{z}_t = \boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)} \\ \\ \sigma^2_{lk} = \frac{\sum_{t=1}^{T} \sum_{(\boldsymbol{i}|\boldsymbol{i}(l)=k)} \left([E(\boldsymbol{s}_t \boldsymbol{s}_t^* \,|\, \boldsymbol{x}_t, \boldsymbol{z}_t = \boldsymbol{i})]_{l,l} - m_{lk}[E(\boldsymbol{s} \,|\, \boldsymbol{x}_t, \boldsymbol{z}_t = \boldsymbol{i})]_l + m^2_{lk}\right) P(\boldsymbol{z}_t = \boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0) + 2b}{\sum_{t=1}^{T} \sum_{(\boldsymbol{i}|\boldsymbol{i}(l)=k)} P(\boldsymbol{z}_t = \boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0) + 2\,(a-1)} \end{array}\right. \tag{III.18}$$

Dans la deuxième expression de (III.18), on note que la pénalisation de la vraisemblance par un inverse gamma de paramètres (a, b) n'a pas changé la forme des équations de ré-estimation et qu'il a suffit de rajouter les termes $2b$ et $2\,(a-1)$ respectivement dans le numérateur et dans le dénominateur.

On constate qu'en plus de la marginalisation sur le temps pour calculer les probabilités $P(\boldsymbol{z}_t = \boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)$, il a fallu effectuer une autre marginalisation au niveau spatial.

Maximisation de $\mathcal{Q}_{\boldsymbol{\eta}_p}$: La ré-estimation des probabilités initiales et des matrices stochastiques pour le cas vectoriel donne :

$$\left\{\begin{array}{l} p(\boldsymbol{i}) = P(\boldsymbol{z}_1 = \boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0) \\ \\ P(\boldsymbol{i}\,\boldsymbol{j}) = \frac{\sum_{t=2}^{T} P(\boldsymbol{z}_{t-1} = \boldsymbol{i},\, \boldsymbol{z}_t = \boldsymbol{j}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)}{\sum_{t=2}^{T} P(\boldsymbol{z}_{t-1} = \boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)} \end{array}\right. \tag{III.19}$$

De la même manière, les probabilités relatives aux étiquettes scalaires sont obtenues par une marginalisation spatiale :

$$\begin{array}{l} p(\boldsymbol{i}(l) = k) = \sum_{(\boldsymbol{i}|\boldsymbol{i}(l)=k)} P(\boldsymbol{z}_1 = \boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0) \\ \\ P(\boldsymbol{i}(l) = r,\, \boldsymbol{j}(l) = s) = \frac{\sum_{t=2}^{T} \sum_{(\boldsymbol{i},\boldsymbol{j}|\boldsymbol{i}(l)=r, \boldsymbol{j}(l)=s)} P(\boldsymbol{z}_{t-1} = \boldsymbol{i},\, \boldsymbol{z}_t = \boldsymbol{j}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)}{\sum_{t=2}^{T} \sum_{(\boldsymbol{i}|\boldsymbol{i}(l)=r)} P(\boldsymbol{z}_{t-1} = \boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)} \end{array} \tag{III.20}$$

Les expressions de $P(\boldsymbol{z}_{t-1} = \boldsymbol{i},\, \boldsymbol{z}_t = \boldsymbol{j}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)$ sont obtenues directement à partir des variables Forward et Backward (III.15) :

$$P(\boldsymbol{z}_{t-1} = \boldsymbol{i},\, \boldsymbol{z}_t = \boldsymbol{j} | \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0) = \mathcal{F}^0_{t-1}(\boldsymbol{i})\, P^0(\boldsymbol{i}, \boldsymbol{j})\, \mathcal{N}_{(\boldsymbol{A}\boldsymbol{m}_j, \boldsymbol{A}\boldsymbol{\Gamma}_j\boldsymbol{A}^* + \boldsymbol{R}_\epsilon)}[\boldsymbol{x}_t]\, \mathcal{B}^0_t(\boldsymbol{j})\, M_t.$$

III.3.2 Algorithme Viterbi-EM

Quand le nombre total des étiquettes $K = \prod_{l=1}^n K_l$ croît, le coût de calcul des probabilités marginales $P(\boldsymbol{z}_t = \boldsymbol{i} | \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)$ et des marginalisations spatiales nécessaires pour la ré-estimation des paramètres des sources devient assez important. Afin de réduire ce coût, on va modifier la stratégie de restauration. Les étiquettes sont remplacées par leur maximum *a posteriori*. Ce qui revient à faire une classification. Ceci est réalisé avec une procédure de relaxation visant à rompre la dépendance temporelle de la chaîne de Markov : à l'itération k, $\widehat{z}_t^k$ maximise $p(z_t \,|\, \boldsymbol{x}_{1..T},\, \widehat{z}^k_{i<t},\, \widehat{z}^{k-1}_{i>t})$, ce qui donne pour $t = 1..T$:

$$z_t^k = \underset{l=1..K}{\arg\max}\, \boldsymbol{T}_{[z^k_{t-1}, l]} \phi(\boldsymbol{x}_t \,|\, \boldsymbol{\theta}_l, \boldsymbol{A}^k) \boldsymbol{T}_{[l, z^{k-1}_{t+1}]}$$

et

$$\begin{cases} z_1^k = \underset{l=1..K}{\arg\max}\, \phi(\boldsymbol{x}_1 \,|\, \boldsymbol{\theta}_l, \boldsymbol{A}^k) \boldsymbol{T}_{[l, z^{k-1}_2]} \\ \\ z_T^k = \underset{l=1..K}{\arg\max}\, \boldsymbol{T}_{[z^k_{T-1}, l]} \phi(\boldsymbol{x}_T \,|\, \boldsymbol{\theta}_l, \boldsymbol{A}^k) \end{cases}$$

où $\boldsymbol{T}$ est la matrice de transition multidimensionnelle et $\phi(\boldsymbol{x} \,|\, \boldsymbol{\theta}_l, \boldsymbol{A}^k)$ est la distribution marginale (on intègre par rapport à $\boldsymbol{s}$) de $\boldsymbol{x}$ conditionnellement à $\boldsymbol{z} = \boldsymbol{l}$:

$$\begin{aligned} \phi(\boldsymbol{x} \,|\, \boldsymbol{\theta}_l, \boldsymbol{A}^k) &= \int_{\boldsymbol{s}} p(\boldsymbol{x}, \boldsymbol{s} \,|\, \boldsymbol{z} = \boldsymbol{l}, \boldsymbol{\theta}_l)\, d\,\boldsymbol{s} \\ &= \mathcal{N}(\boldsymbol{x}\,;\, \boldsymbol{A}\boldsymbol{m}_l,\, \boldsymbol{A}\boldsymbol{\Gamma}_l\boldsymbol{A}^* + \boldsymbol{R}_\epsilon) \end{aligned}$$

Ensuite, toutes les espérances intervenant dans l'EM sont simplement remplacées par une seule espérance conditionnelle :

$$\begin{aligned} E[f(\boldsymbol{s}_t) | \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0] &= \textstyle\sum_{\boldsymbol{i}} E[f(\boldsymbol{s}_t) | \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0, \boldsymbol{z}_t = \boldsymbol{i}]\; p(\boldsymbol{z}_t = \boldsymbol{i} | \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0) \\ &\approx E[f(\boldsymbol{s}_t) | \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0, \widehat{z}_t]. \end{aligned}$$

III.3.3 Algorithme Gibbs-EM

Dans cet algorithme, on simule les variables cachées z_t selon leurs distributions *a posteriori*. L'avantage de cette procédure est double : réduction du coût de calcul et la possibilité d'éviter les maxima locaux. Les étiquettes sont simulées avec la procédure de Gibbs : à l'itération k, $\widehat{z}_t^k \;\sim\; p(z_t \,|\, \boldsymbol{x}_{1..T},\; \widehat{z}^k_{i<t},\; \widehat{z}^{k-1}_{i>t})$, ce qui donne pour $t = 1..T$:

$$z_t \;\sim\; T_{z_{t-1} z_t} \phi(\boldsymbol{x}_t \,|\, \boldsymbol{\theta}_z, \boldsymbol{A}^k) T_{z_t z_{t+1}}$$

et

$$\begin{cases} z_1 \;\sim\; \phi(\boldsymbol{x}_1 \,|\, \boldsymbol{\theta}_z, \boldsymbol{A}^k) T_{z_1 z_2} \\ \\ z_T \;\sim\; T_{z_{T-1} z_T} \phi(\boldsymbol{x}_T \,|\, \boldsymbol{\theta}_z, \boldsymbol{A}^k) \end{cases}$$

On se contente d'un seul cycle de l'échantillonneur de Gibbs. En effet, le paramètre d'intérêt $\boldsymbol{\theta}$ varie au cours des itérations de l'algorithme et on n'a pas ainsi besoin d'avoir un échantillon exact de $p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1..T}, \boldsymbol{\theta})$.

Le coût de calcul d'une itération de cette version de l'algorithme est approximativement le même que celui de l'algorithme Viterbi-EM puisqu'à chaque instant t on a besoin de calculer tout le vecteur $[p(z_t = i \,|\, \boldsymbol{x}_{1..T},\, z_{s\neq t})]_{i=1..K}$.

Le but des versions Viterbi et Gibbs de l'algorithme EM est de réduire la partie du coût de calcul due à la structure temporelle des chaînes de Markov discrètes $(z_t^j)_{t=1..T}^{j=1..n}$. La complexité $K^2\,T$ ($K = \prod_{l=1}^{n} K_l$) de la procédure Forward-Backward est réduite à $K\,T$ (réduction par un facteur K). Cependant, il existe une autre source qui ralentit l'algorithme : le nombre de toutes les étiquettes vectorielles $K = |\mathcal{Z}_1 \times \mathcal{Z}_2 \times \ldots \mathcal{Z}_n|$. Son impact apparaît à deux niveaux :

- au niveau du calcul des K quantités $P(\boldsymbol{z}_t = \boldsymbol{i} \,|\, \boldsymbol{x}_{1..T}, \boldsymbol{\theta})$ dans les trois algorithmes EM, Viterbi-EM et Gibbs-EM, nécessaire pour respectivement calculer les espérances (III.13), estimer les variables cachées $\boldsymbol{z}_{1..T}$ et les simuler selon leur distribution *a posteriori*.
- au niveau de la marginalisation spatiale dans la ré-estimation des paramètres $\boldsymbol{\eta}_g$ et $\boldsymbol{\eta}_p$ dans les expressions (III.18) et (III.20).

Nous introduisons dans le paragraphe suivant une procédure de relaxation afin de réduire le coût dû au nombre exponentiel des étiquettes vectorielles.

III.3.4 Versions accélérées

[A] Algorithme Fast-Viterbi-EM

La distribution *a posteriori* du vecteur $\boldsymbol{z}$ s'écrit :

$$\begin{aligned} p(\boldsymbol{z} \,|\, \boldsymbol{x}, \boldsymbol{\theta}) &= \int_{\boldsymbol{s}} p(\boldsymbol{z}, \boldsymbol{s} \,|\, \boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{s} \\ &\propto p(\boldsymbol{z}) \int_{\boldsymbol{s}} p(\boldsymbol{x} \,|\, \boldsymbol{s}, \boldsymbol{\theta}) p(\boldsymbol{s} \,|\, \boldsymbol{z}, \boldsymbol{\theta})\, d\boldsymbol{s}. \end{aligned} \tag{III.21}$$

On note, d'après la deuxième ligne de l'équation plus haut que c'est la distribution $p(\boldsymbol{x} \,|\, \boldsymbol{s}, \boldsymbol{\theta})$ qui donne aux composantes z^j du vecteur $\boldsymbol{z}$ une dépendance spatiale *a posteriori* ce qui n'était pas le cas *a priori* ($p(\boldsymbol{z}) = \prod p(z^j)$). Par conséquent, afin d'estimer ou de simuler les étiquettes z^j, on a besoin de manipuler tout le vecteur $\boldsymbol{z}$. C'est le cas, par exemple, quand on veut calculer les probabilités marginales des composantes z^j qui nécessite une sommation sur toutes les étiquettes vectorielles ayant z^j comme $j^{\text{ème}}$ composante :

$$p(z_j(t) \,|\, \boldsymbol{x}, \boldsymbol{\theta}) = \sum_{\boldsymbol{z}\in\mathcal{Z} | \boldsymbol{z}(j)=z_j(t)} p(\boldsymbol{z}(t) \,|\, \boldsymbol{x}(t), \boldsymbol{\theta}). \tag{III.22}$$

Afin de réduire le coût dû à cette dépendance spatiale, on propose d'introduire une relaxation sur les composantes du vecteur des sources. L'expression (III.22) est remplacée par :

$$p(z_j(t) \,|\, \boldsymbol{x},\, \boldsymbol{\theta}', \widehat{s}_{l\neq j})$$

qui est obtenue en intégrant seulement par rapport à s_j. Les autres composantes $s_l\,(l \neq j)$ sont fixées à leurs estimées MAP ou simulées selon leurs distributions *a posteriori*. Fixer les composantes $s_{l\neq j}$ évite la structure vectorielle du mélange et réduit ainsi considérablement le coût de calcul. Au lieu de calculer, à chaque instant t, k^n ($k = K_1 = \cdots = K_n$) probabilités $p(\boldsymbol{z}_t \,|\, \boldsymbol{x}_t, \boldsymbol{\theta})$ dans la version Viterbi ou Gibbs, nous avons, avec la stratégie de relaxation, seulement $n \times k$ probabilités $\left(p(z_j(t) \,|\, \boldsymbol{x},\, \boldsymbol{\theta}', \widehat{s}_{l\neq j})\right)_{z=1..k}^{j=1..n}$ à calculer. En plus, en fixant les composantes $s_{l\neq j}$, la distribution *a posteriori* de la composante s_j est un mélange de K_j gaussiennes mono-variées et donc son estimation ou son échantillonnage est plus facile que dans le cas vectoriel où $\boldsymbol{s}$ suit une distribution *a posteriori* mélange de $\prod_{l=1}^{n} K_l$ gaussiennes multivariables.

La version *Fast-Viterbi-EM* contient donc une relaxation spatiale (en fixant $s_{l\neq j}$) en plus de la relaxation temporelle (en fixant $z_{i\neq t}$) :

Algorithme *Fast-Viterbi-EM*

$$\left\{\begin{array}{l} 1.\ \ z_j(t)^k = \underset{l=1..K_j}{\arg\max}\, \boldsymbol{T}_{[z^k_{j,t-1},l]}\phi(\boldsymbol{x}_t \mid s_{l\neq j}, \boldsymbol{\theta}_l, \boldsymbol{A}^k)\boldsymbol{T}_{[l,z^{k-1}_{j,t+1}]} \\ \\ 2.\ \ s_j \sim p(s_j \mid \boldsymbol{x}_t, z_j(t)^k, \boldsymbol{\theta}) \\ \\ j = 1..n,\ \ t = 1..T \end{array}\right. \quad (\text{III.23})$$

$$et$$

$$\left\{\begin{array}{l} z_j(1)^k = \underset{l=1..K_j}{\arg\max}\, \phi(\boldsymbol{x}_1 \mid s_{l\neq j}, \boldsymbol{\theta}_l, \boldsymbol{A}^k)\boldsymbol{T}_{[l,z^{k-1}_{j,2}]} \\ \\ z_j(T)^k = \underset{l=1..K_j}{\arg\max}\, \boldsymbol{T}_{[z^k_{j,T-1},l]}\phi(\boldsymbol{x}_T \mid s_{l\neq j}, \boldsymbol{\theta}_l, \boldsymbol{A}^k) \end{array}\right.$$

où $\boldsymbol{T}$ est la matrice de transition de la composante j. On note qu'après chaque estimation de l'étiquette $z_j(t)^k$, on remet à jour la source s_j.

[B] Algorithme Fast-Gibbs-EM

Dans cet algorithme, l'étiquette $z_j(t)$ est simulée selon sa distribution *a posteriori* :

Algorithme *Fast-Gibbs-EM*

$$\left\{\begin{array}{l} 1.\ \ z_j(t) \sim \boldsymbol{T}_{z_{t-1}z_t}\phi(\boldsymbol{x}_t \mid s_{l\neq j}, \boldsymbol{\theta}_z, \boldsymbol{A}^k)\boldsymbol{T}_{z_t z_{t+1}} \\ \\ 2.\ \ s_j \sim p(s_j \mid \boldsymbol{x}_t, z_j(t)^k, \boldsymbol{\theta}) \\ \\ j = 1..n,\ \ t = 2..T-1 \end{array}\right. \quad (\text{III.24})$$

$$et$$

$$\left\{\begin{array}{l} z_j(1) \sim \phi(\boldsymbol{x}_1 \mid s_{l\neq j}, \boldsymbol{\theta}_z, \boldsymbol{A}^k)\boldsymbol{T}_{z_1 z_2} \\ \\ z_j(T) \sim \boldsymbol{T}_{z_{T-1}z_T}\phi(\boldsymbol{x}_T \mid s_{l\neq j}, \boldsymbol{\theta}_z, \boldsymbol{A}^k) \end{array}\right.$$

où $\boldsymbol{T}$ est la matrice de transition de la composante j.

La complexité du calcul concernant la remise à jour des probabilités discrètes est ainsi réduite d'un facteur $\frac{\prod_{l=1}^n K_l}{\sum_{l=1}^n K_l}$. Si le nombre des composantes des mélanges est le même pour toutes les sources ($k = K_1 = \cdots = K_l$), la complexité est réduite de k^n à $n \times k$.

III.4 Simulations numériques

Afin de montrer les performances des algorithmes proposés, on considère un mélange de deux sources :

– **Source** 1 **:** La distribution *a priori* est un mélange de 4 gaussiennes $(m,\, \sigma^2) \in \{(-3, 0.1), (-1, 0.1), (1, 0.1), (3, 0.1)\}$ avec une matrice de transition $\boldsymbol{T}_1$:

$$\boldsymbol{T}_1 = \begin{pmatrix} 0.9 & 0.05 & 0.03 & 0.02 \\ 0.8 & 0.1 & 0.05 & 0.05 \\ 0.7 & 0.02 & 0.08 & 0.2 \\ 0.5 & 0.2 & 0.2 & 0.1 \end{pmatrix}$$

– **Source** 2 **:** La distribution *a priori* est un mélange de 4 gaussiennes $(m,\, \sigma^2) \in \{(-3, 0.1), (-1, 0.1), (1, 0.1), (3, 0.1)\}$ avec une matrice de transition $\boldsymbol{T}_2$:

$$\boldsymbol{T}_2 = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix}$$

La première colonne de $\boldsymbol{T}_1$ est dominante. Ce qui signifie que les étiquettes z_t ont une grande probabilité de rester dans l'état 1. Cependant, la matrice $\boldsymbol{T}_2$ possède la même ligne. Ce qui signifie que le mélange est i.i.d.. La figure (III.1) montre des simulations de ces signaux.

Les deux sources sont mélangées avec la matrice $\boldsymbol{A} = \begin{pmatrix} 1 & 0.6 \\ -0.5 & 1 \end{pmatrix}$. Un bruit gaussien de covariance $\boldsymbol{R}_\epsilon = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ (SNR= 8dB) est ajouté au mélange. Le nombre d'échantillons est $T = 1000$. La figure (III.2) illustre les graphes des sources mélangées $(x_1(t))_{t=1..T}$ et $(x_2(t))_{t=1..T}$.

Afin d'évaluer les performances de l'identification de la matrice de mélange, on utilise l'indice suivant [Moreau et Macchi, 1996] :

$$ind(S = \widehat{\boldsymbol{A}}^{-1}\, \boldsymbol{A}) = \frac{1}{2}\left[\sum_i \left(\sum_j \frac{|S_{ij}|^2}{\max_l |S_{il}|^2} - 1\right) + \sum_j \left(\sum_i \frac{|S_{ij}|^2}{\max_l |S_{lj}|^2} - 1\right)\right]$$

La figure (III.3) illustre l'évolution, au cours des itérations, des estimées par l'algorithme EM exact des coefficients de la matrice de mélange. La ligne horizontale indique la vraie valeur du coefficient. On note la convergence de l'algorithme vers la bonne valeur après 20 itérations. Dans ces simulations, on fixe les valeurs des hyperparamètres et on se concentre sur l'estimation de la matrice de mélange pour pouvoir comparer plus facilement les performances des algorithmes proposées. En effet, l'estimation des hyperparamètres avec l'algorithme EM exact est très coûteuse. Tandis qu'avec les versions Viterbi/Gibbs, l'estimation des hyperparamètres est simple à réaliser et ne ralentit pas d'une manière significative la convergence de l'algorithme (convergence après 100 itérations au lieu de 20 itérations comme l'illustre la figure (III.7)). La figure (*b*) de (III.3) illustre la convergence de l'indice de performance de l'algorithme EM vers une valeur satisfaisante de -31 dB. La figure (III.4) montre les résultats de la reconstruction des sources en traçant sur le même graphe les sources originales et les sources reconstruites. On note la bonne qualité de la reconstruction.

Les figures (III.5) et (III.6) montre les résultats de l'algorithme *Viterbi-EM* sur le même exemple de simulation. On note un petit biais concernant l'estimation de la matrice de mélange. On peut expliquer ce biais par le fait qu'on estime conjointement les variables cachées $\boldsymbol{z}_t$ au lieu de les intégrer hors problème. L'estimé est donc biaisé par rapport au maximum de vraisemblance. On note cependant que la consistance et l'efficacité de la méthode du maximum de vraisemblance ne sont garanties que dans le régime asymptotique. Avec un nombre modéré d'échantillons, on perd ces propriétés. Par conséquent, une estimation conjointe des variables cachées n'est pas nécessairement plus mauvaise que la maximisation de la vraisemblance incomplète (voir le biais de l'estimée par l'algorithme EM sur la figure gauche de (III.3)). On note la convergence de l'indice de performance vers la valeur de -24 dB. Le coût de calcul est réduit d'un facteur de $K = 16$ par rapport à l'algorithme EM.

Les figures (III.7) et (III.8) illustrent les résultats de l'algorithme *Gibbs-EM*. On note les fluctuations dues à l'aspect stochastique de l'algorithme. On peut rajouter une procédure de recuit simulé qui fait transformer l'algorithme vers l'EM au cours des itérations [Celeux et Diebolt, 1990]. L'extension naturelle de

l'algorithme *Gibbs-EM* est l'échantillonnage de Gibbs où on échantillonne aussi le paramètre $\boldsymbol{\theta}$ selon sa distribution *a posteriori* complète. On obtient ainsi une chaîne de Markov $(\boldsymbol{z}^k, \boldsymbol{\theta}^k)$. Les échantillons $\boldsymbol{\theta}^k$ suivent asymptotiquement la distribution *a posteriori* $p(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T})$.

Les figures (III.9) et (III.10) illustrent les résultats de l'algorithme *Fast-Viterbi-EM*. Les figures (III.11) et (III.12) illustrent les résultats de l'algorithme *Fast-Gibbs-EM*. On note que les versions rapides ont presque les mêmes performances que les algorithmes Viterbi/Gibbs mais avec une durée moins importante par itération.

III.5 Conclusion

L'estimation des paramètres d'un modèle de Markov caché est un problème à données incomplètes. Les données manquantes sont les étiquettes du mélange. En généralisant ce problème à la séparation aveugle de sources modélisées par des modèles de Markov cachés, on fait apparaître une deuxième couche de variables manquantes formée par les sources. Les algorithmes de restauration-maximisation représentent un outil efficace et naturel pour l'estimation conjointe de la matrice de mélange et des paramètres des HMM. On propose trois stratégies différentes pour l'étape de restauration, qui se distinguent par leurs complexités et leurs propriétés de convergence :

- L'algorithme EM exact : la fonctionnelle est séparable en trois quantités qui correspondent à trois ensembles de paramètres : les paramètres de $p(\boldsymbol{x} \mid \boldsymbol{s}, \boldsymbol{z})$, ceux de $p(\boldsymbol{s} \mid \boldsymbol{z})$ et ceux de $p(\boldsymbol{z})$.
- L'algorithme *Viterbi-EM* : les étiquettes sont remplacées par leur maximum *a posteriori* MAP.
- L'algorithme *Gibbs-EM* : les étiquettes sont échantillonnées selon leur distribution *a posteriori*.

Une relaxation spatiale est proposée afin d'accélérer les algorithmes ci-dessus. Cette modification vise à réduire la partie du coût de calcul due à l'aspect vectoriel du mélange qui varie exponentiellement avec le nombre des sources et le nombre des densités constituants le mélange.

Bibliographie

[Attias, 1999] H. Attias. Blind separation of noisy mixture : An EM algorithm for independent factor analysis. *Neural Computation*, 11 : 803–851, 1999.

[Bermond, 2000] O. Bermond. *Méthodes statistiques pour la séparation de sources.* thèse de doctorat, Ecole Nationale Supérieure des Télécommunications, 2000.

[Brewer, 1978] J. W. Brewer. Kronecker products and matrix calculus in system theory. *IEEE Trans. Circ. Syst.*, CS-25 (9) : 772–781, 1978.

[Celeux et Diebolt, 1985] G. Celeux et J. Diebolt. The SEM algorithm : A probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Comput. Statist. Quat.*, 2 : 73–82, 1985.

[Celeux et Diebolt, 1990] G. Celeux et J. Diebolt. Une version de type recuit simulé de l'algorithme EM. *Compte-rendus de l'académie des sciences*, 310 : 2, 1990.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird et D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39 : 1–38, 1977.

[Ghahramani et Jordan, 1997] Z. Ghahramani et M. Jordan. Factorial Hidden Markov Models. *Machine Learning*, (29) : 245–273, 1997.

[Hathaway, 1986] R. J. Hathaway. A constrained EM algorithm for univariate normal mixtures. *J. Statist. Comput. Simul.*, 23 : 211–230, 1986.

[McLachlan et Krishnan, 1997] G. J. McLachlan et T. Krishnan. *The EM Algorithm and Extensions.* Wiley series in probability and statistics. John Wiley and Sons, Inc., 1997.

[Moreau et Macchi, 1996] E. Moreau et O. Macchi. High-order contrasts for self-adaptative source separation. In *Adaptative Control Signal Process. 10*, pages 19–46, 1996.

[Rabiner et Juang, 1986] L. R. Rabiner et B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Mag.*, pages 4–16, 1986.

[Ridolfi et Idier, 1999] A. Ridolfi et J. Idier. Penalized maximum likelihood estimation for univariate normal mixture distributions. In *Actes 17^e coll. GRETSI*, pages 259–262, Vannes, septembre 1999.

[Snoussi et Mohammad-Djafari, 2000] H. Snoussi et A. Mohammad-Djafari. Bayesian source separation with mixture of Gaussians prior for sources and Gaussian prior for mixture coefficients. In A. Mohammad-Djafari, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 388–406, Gif-sur-Yvette, juillet 2000. Proc. of MaxEnt, Amer. Inst. Physics.

[Snoussi et Mohammad-Djafari, 2001] H. Snoussi et A. Mohammad-Djafari. Penalized maximum likelihood for multivariate gaussian mixture. In R. L. Fry, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 36–46. MaxEnt Workshops, Amer. Inst. Physics, août 2001.

[Snoussi *et al.*, 2001] H. Snoussi, G. Patanchon, J. Macías-Pérez, A. Mohammad-Djafari et J. Delabrouille. Bayesian blind component separation for cosmic microwave background observations. In R. L. Fry, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 125–140. MaxEnt Workshops, Amer. Inst. Physics, août 2001.

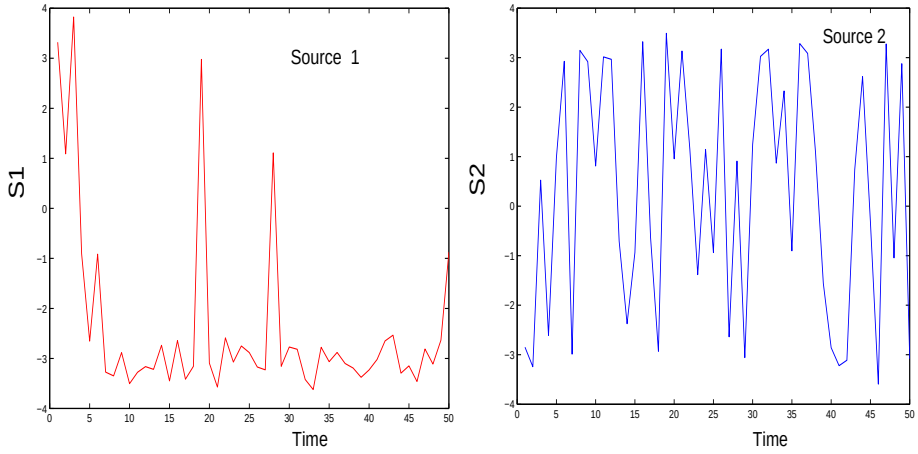


FIG. III.1: Graphes des sources s_1 et s_2. Seuls les 50 premiers échantillons sont montrés.

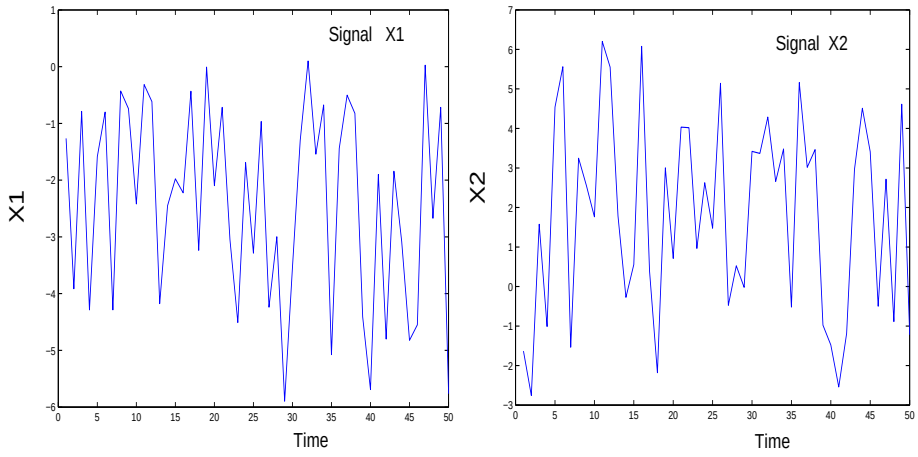
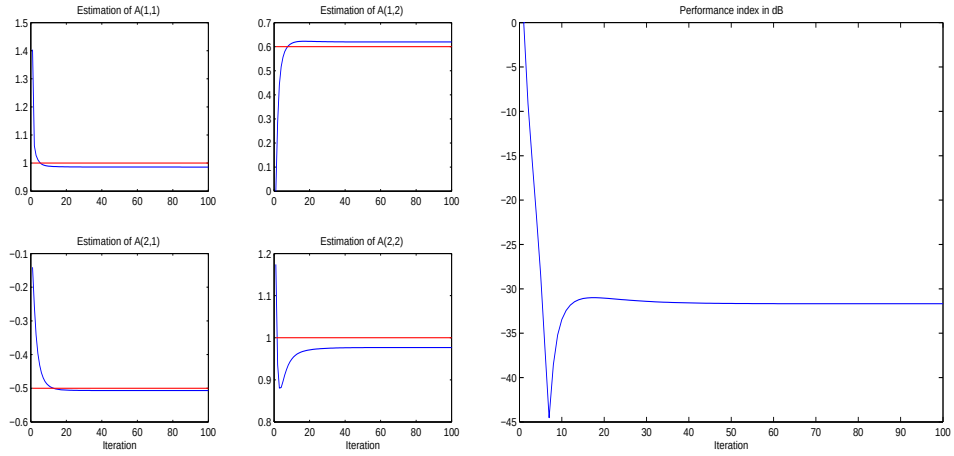


FIG. III.2: Graphes des sources mélangées $X_1 = a_{11}S_1 + a_{12}S_2$ et $X_2 = a_{21}S_1 + a_{22}S_2$



a b

FIG. III.3: (a) Evolution des estimés des coefficients de mélange avec l'algorithme EM au cours des itérations, (b) évolution de l'indice de performance de l'algorithme EM.

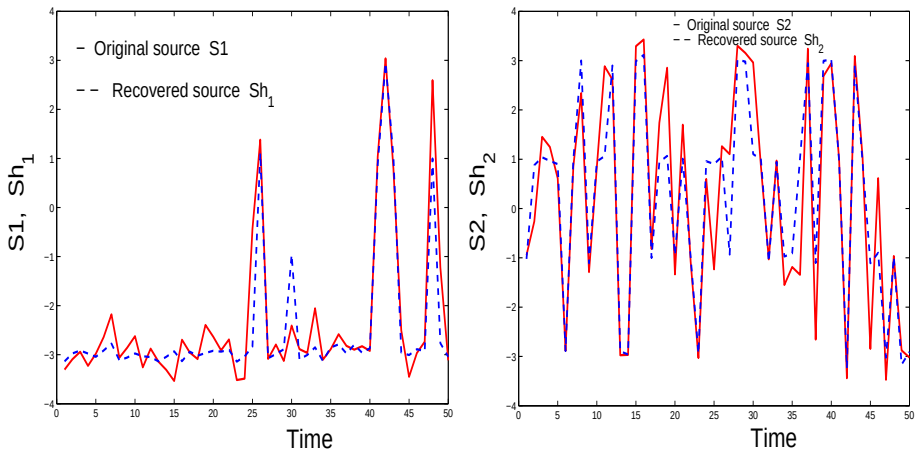
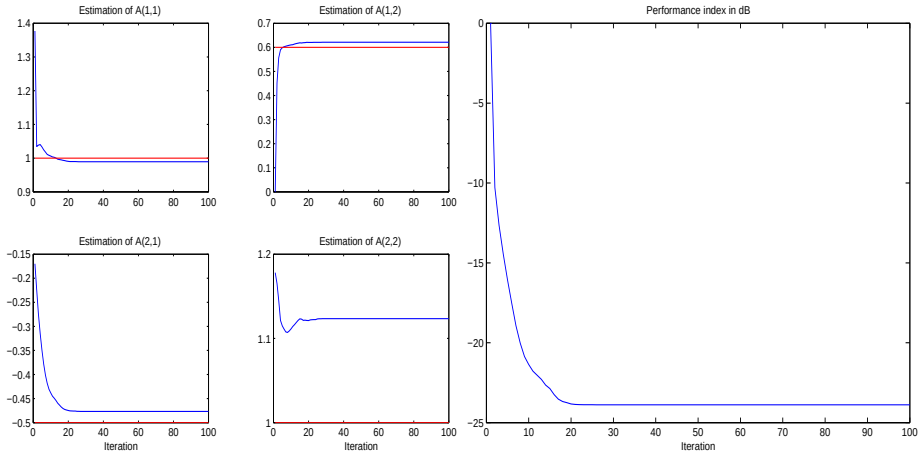


FIG. III.4: Résultats de reconstruction des sources avec l'algorithme EM.



a b

FIG. III.5: (a) Evolution au cours des itérations des estimés des coefficients de mélange avec l'algorithme *Viterbi-EM*, (b) évolution de l'indice de performance avec *Viterbi-EM*.

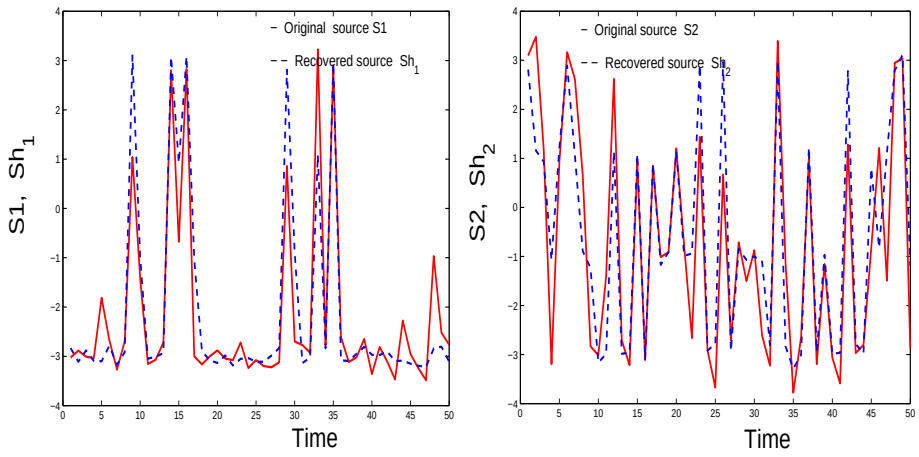
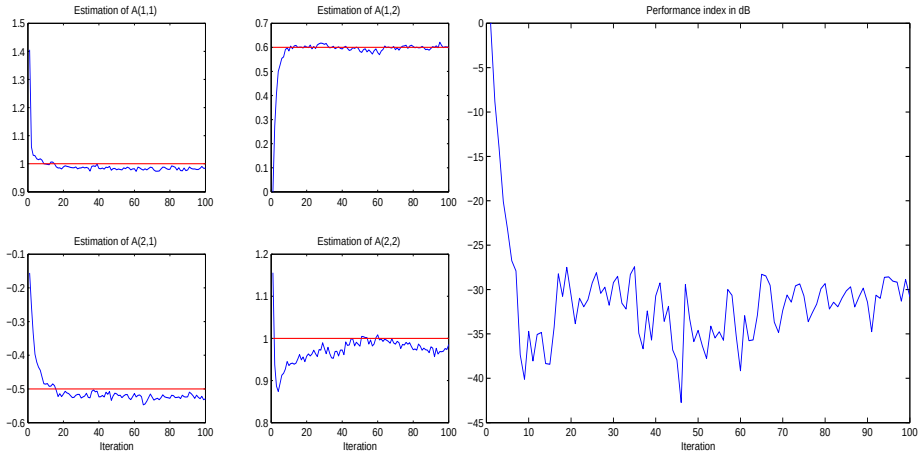


FIG. III.6: Résultats de reconstruction des deux sources avec l'algorithme *Viterbi-EM*.



a b

FIG. III.7: (a) Evolution au cours des itérations des estimés des coefficients de mélange avec l'algorithme *Gibbs-EM*, (b) évolution de l'indice de performance avec *Gibbs-EM*.

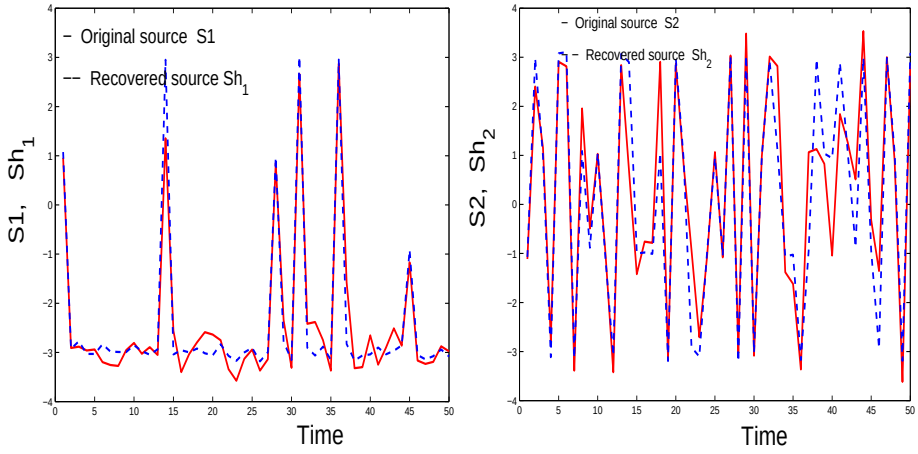
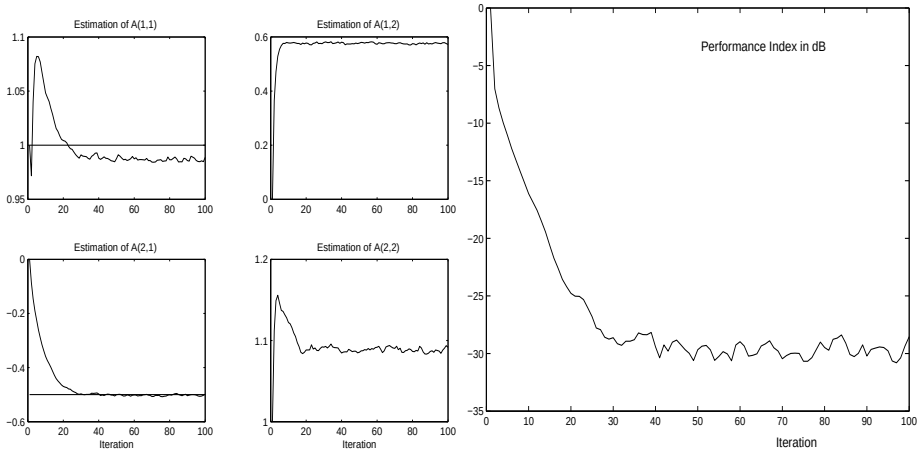


FIG. III.8: Résultats de reconstruction des deux sources avec l'algorithme *Gibbs-EM*.



a b

FIG. III.9: (a) Evolution au cours des itérations des estimés des coefficients de mélange avec l'algorithme *Fast-Viterbi-EM*, (b) évolution de l'indice de performance avec *Fast-Viterbi-EM*.

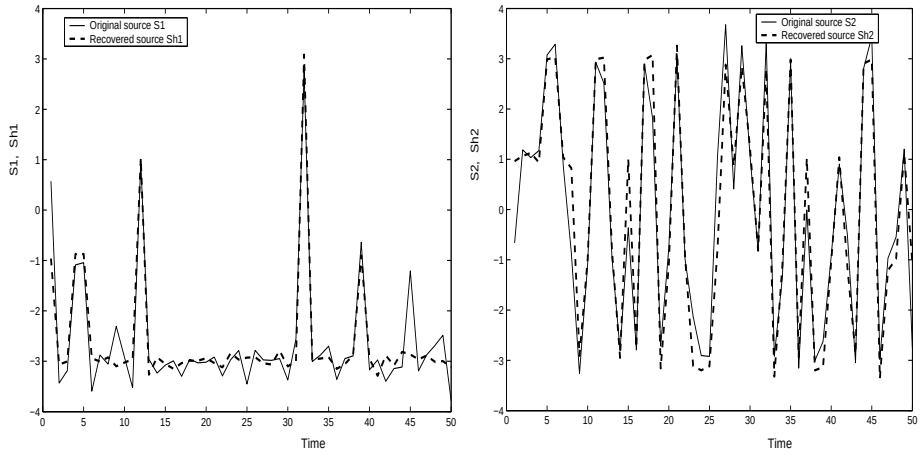
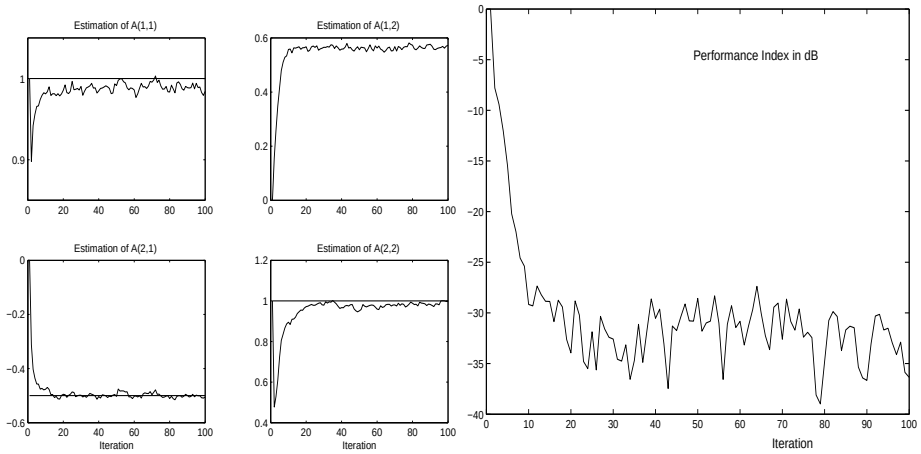


FIG. III.10: Résultats de reconstruction des deux sources avec l'algorithme *Fast-Viterbi-EM*.



a b

FIG. III.11: (a) Evolution au cours des itérations des estimés des coefficients de mélange avec l'algorithme *Fast-Gibbs-EM*, (b) évolution de l'indice de performance avec *Fast-Gibbs-EM*.

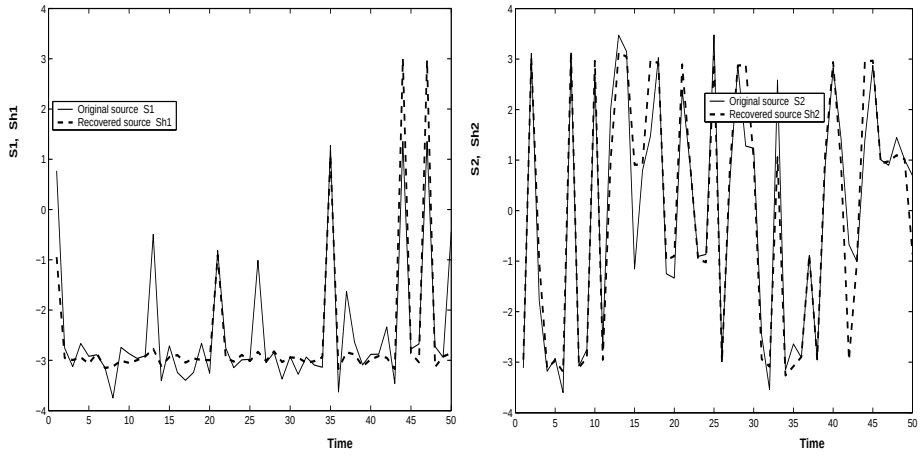
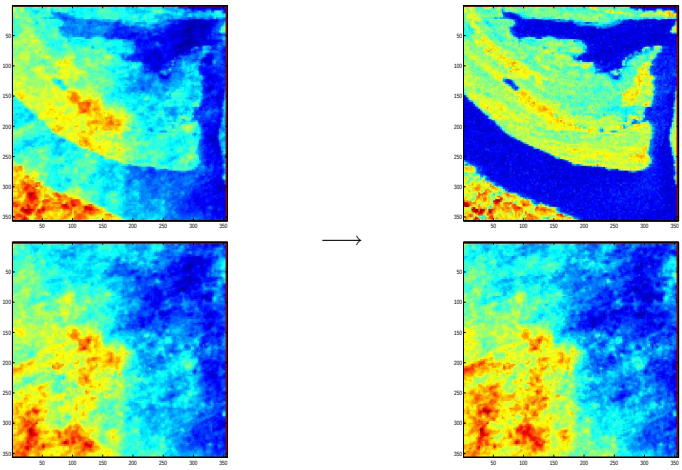


FIG. III.12: Résultats de reconstruction des deux sources avec l'algorithme *Fast-Gibbs-EM*.

Chapitre IV

SÉPARATION DE SOURCES MULTIVARIÉES : NON STATIONNARITÉ SPATIALE



IV.1 Introduction

Les observations sont représentées par m images $(\boldsymbol{X}^i)_{i=1..m}$. Chaque image $\boldsymbol{X}^i$ est définie sur un ensemble de sites $\mathcal{S}$ correspondants aux pixels de l'image[1] : $\boldsymbol{X}^i = (x_r^i)_{r\in\mathcal{S}}$. On suppose que les observations sont le résultat d'un mélange linéaire instantané bruité de n images (sources) $(\boldsymbol{S}^j)_{j=1..n}$ définies sur le même ensemble de sites $\mathcal{S}$:

$$x_r^i = \sum_{j=1}^{n} a_{ij} s_r^j + n_r^i,\ r \in \mathcal{S},\ i = 1, \ldots, m$$

où $\boldsymbol{A} = (a_{ij})$ est la matrice de mélange et $\boldsymbol{N}^i = (n_r^i)_{r\in\mathcal{S}}$ est l'image modélisant le bruit additif sur le $i^{\text{ème}}$ capteur (voir figure (IV.1)).

A chaque pixel $r \in \mathcal{S}$, la notation matricielle est :

$$\boldsymbol{x}_r = \boldsymbol{A}\,\boldsymbol{s}_r + \boldsymbol{n}_r. \tag{IV.1}$$

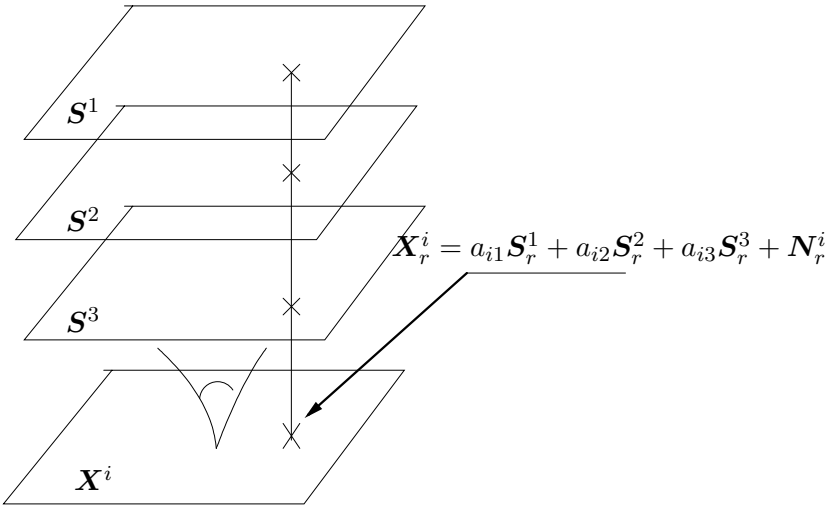


FIG. IV.1: Mélange de sources : l'image observée sur le capteur i est une combinaison linéaire bruitée des images sources. Les cœfficients de la combinaison forment la $i^{\text{ème}}$ ligne de la matrice de mélange $\boldsymbol{A}$.

[A] Modélisation des sources et du bruit

Le bruit est supposé statistiquement indépendant des sources, gaussien, de moyenne nulle et temporellement indépendant de covariance $\boldsymbol{R}_\epsilon$:

$$\mathrm{E}\big[\boldsymbol{n}_r\,\boldsymbol{n}_s^*\big] = \delta(r-s)\,\boldsymbol{R}_\epsilon,$$

où * désigne le transposé d'un vecteur.

La matrice $\boldsymbol{R}_\epsilon$ n'est pas forcément diagonale et on peut ainsi tenir compte d'une éventuelle corrélation entre les bruits des différents capteurs.

La modélisation des sources par des mélanges de gaussiennes dans le problème de séparation de sources a été motivée par les raisons suivantes.

[1] Le pixel est l'équivalent de l'indice temporel et la notation $\boldsymbol{X}$ est équivalente à la notation $\boldsymbol{x}_{1..T}$ définie dans le cas monodimensionnel (voir chapitre (III))

- Le mélange de gaussiennes donne une classe de distributions très riche pouvant atteindre toute distribution de probabilité en jouant sur le nombre de composantes constituant le mélange.
- On peut assurer l'identifiabilité de la matrice de mélange $\boldsymbol{A}$ en garantissant les conditions du théorème de Darmois ([Darmois, 1953; Comon, 1994], chapitre (I)). En effet, sous cette modélisation, les sources ne sont pas gaussiennes.
- On obtient des expressions analytiques explicites lors de l'implémentation de l'algorithme EM.

En plus de ces avantages, la modélisation par des modèles de Markov cachés nous permet de :
- tenir compte d'une structure spatiale,
- mettre l'accent sur la structure cachée de ce modèle qui fait apparaître une étape de classification. En effet, la modélisation des étiquettes par un champ de Markov (dans le cas 2-D) est un moyen de régulariser la classification et de la rendre robuste vis-à-vis du bruit.

Ce modèle est bien approprié en traitement d'images. En effet, les images naturelles sont souvent homogènes par morceaux. Cette homogénéité locale peut être modélisée par un champ d'étiquettes discrètes $\boldsymbol{Z}$ possédant la propriété de Markov. Avant de donner l'expression d'un champ de Markov, on rappelle quelques définitions concernant la notion de voisinage [Winkler, 1995].

Définition 2 *Une collection $\partial = \{\partial(r),\ r \in \mathcal{S}\}$ de sous-ensembles de $\mathcal{S}$ est appelée un système de voisinage, si* (**i**) $r \notin \partial(r)$ *et* (**ii**) $r \in \partial(t)$ *si et seulement si* $t \in \partial(r)$. *Les sites $r \in \partial(t)$ sont appelés les* **voisins** *de t. Un sous ensemble C de $\mathcal{S}$ est appelé une* **clique** *si deux éléments distincts de C sont voisins. L'ensemble des cliques est noté $\mathcal{C}$. On note $r \sim t$ pour r et t voisins.*

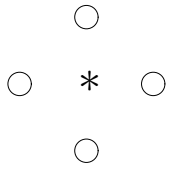
Exemple 4 *On suppose que $\mathcal{S}$ est un sous-graphe de $\mathbb{Z} \times \mathbb{Z}$,*

$$\mathcal{S} = \{(i,j) \in \mathbb{Z} \times \mathbb{Z} \mid -m \leq i,j \leq m\},$$

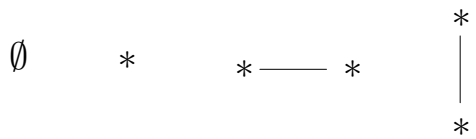
et le système de voisinage est défini par :

$$\partial(i,j) = \{(k,l) \mid 0 < (k-i)^2 + (l-j)^2 \leq c\},$$

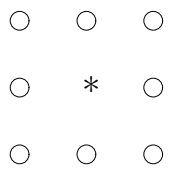
où c est une constante qui mesure l'étendu du voisinage. Pour $c = 1$, chaque site $*$ possède 4 voisins $\circ$ (voisinage d'ordre 1) :



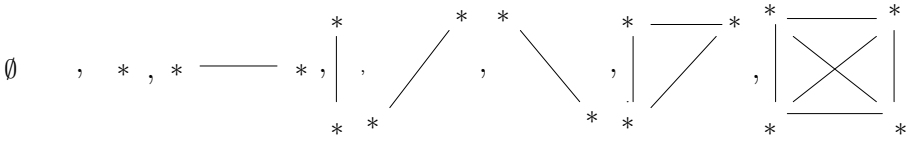
Les cliques correspondants sont :



Pour $c = 2$, chaque site $*$ possède 8 voisins $\circ$:



Les cliques correspondants sont :



Pour un système de voisinage ∂, on peut définir un champ de Markov :

Définition 3 *Un champ aléatoire P_M est un champ de* **Markov** *pour le système de voisinage ∂, si pour tout $\boldsymbol{Z}$,*

$$P_M(Z_r \mid \boldsymbol{Z}_{\mathcal{S}\setminus\{r\}}) = P_M(z_r \mid \boldsymbol{Z}_{\partial(r)}), \tag{IV.2}$$

où la notation $\boldsymbol{Z}_A$ désigne le champ restreint à l'ensemble $A \subset \mathcal{S}$.

On note que cette propriété est plus difficile à caractériser que dans le cas unidimensionnel où la chaîne de Markov est simplement définie par sa probabilité initiale et sa matrice de transition. La règle séquentielle de Bayes :

$$Pr(\boldsymbol{z}_{1..T}) = Pr(z_T \mid z_{T-1})\, Pr(z_{T-1} \mid z_{T-2}) \ldots Pr(z_2 \mid z_1) Pr(z_1)$$

qui permet le calcul de la probabilité conjointe de tout le vecteur $\boldsymbol{z}_{1..T}$, n'a pas d'équivalent simple dans le cas 2-D. Cependant, d'après le théorème de Hammersley-Clifford [Hammersley et Clifford, 1968], on possède une meilleure caractérisation d'un champ de Markov. En effet, un champ aléatoire P_M est un champ de Markov *si et seulement si* P_M est un champ de **Gibbs** dont l'expression est la suivante :

$$P_G(\boldsymbol{Z}) = \frac{\exp - \left(\sum_{C \in \mathcal{C}} U_C(\boldsymbol{Z}) \right)}{\sum_{\boldsymbol{Y}} \exp - \left(\sum_{C \in \mathcal{C}} U_C(\boldsymbol{Y}) \right)} \tag{IV.3}$$

où $\mathcal{C}$ est l'ensemble des cliques correspondant au voisinage ∂ et $U_C(\boldsymbol{Z})$ est la fonction **Potentielle** vérifiant la propriété suivante :

$$(i) \quad U_\emptyset = 0,$$

$$(ii) \quad U_A(\boldsymbol{Z}) = U_A(\boldsymbol{Z}'),\ si\ \boldsymbol{Z}_A = \boldsymbol{Z}'_A$$

Dans ce chapitre, on prend, comme exemple, les champs de **Potts** :

$$P_M(\boldsymbol{Z}) = [W(\alpha)]^{-1} \exp\{\alpha \sum_{r \sim s} I_{z_r = z_s}\},$$

où $r \sim s$ est défini par le système de voisinage choisi, I est la fonction caractéristique et α est un cœfficient qui reflète la dépendance spatiale du champ de Gibbs. α est appelé paramètre de champ et il est supposé connu dans la suite. Un champ de **Ising** est un champ de Potts à deux couleurs.

Chaque source $\boldsymbol{S}^j$ est ainsi modélisée par un champ de Markov caché (HMF) : conditionnellement à un champ de Markov (IV.2) $\boldsymbol{Z}^j$ (équivalent à un champ de Gibbs (IV.3)), la source $\boldsymbol{S}^j$ est un champ à valeurs continues dont les éléments $S_r^j, r \in \mathcal{S}$ sont statistiquement indépendants :

$$p(\boldsymbol{S}^j \mid \boldsymbol{Z}^j, \boldsymbol{\eta}^j) = \prod_{r \in \mathcal{S}} p_r(s_r^j \mid z_r^j, \boldsymbol{\eta}^j)$$

où $\boldsymbol{\eta}^j \in \mathbb{R}^d$ est le vecteur des paramètres des lois conditionnelles $p_r(. \mid z_r)$. Dans la suite, on suppose que $p_r(. \mid z_r)$ est une gaussienne. Dans ce cas, si K_j est le nombre d'étiquettes de la $j^{\text{ème}}$ source, le paramètre $\boldsymbol{\eta}^j = (\mu_{jk}, \sigma_{jk}^2)_{k=1..K_j}$ forme les K_j moyennes et variances de ces gaussiennes.

On note que chaque source possède sa propre classification $\boldsymbol{Z}$ avec son propre paramètre de champ α reflétant l'homogénéité de cette classification et ses propres moyennes et variances (μ_k, σ_k^2) correspondant aux gaussiennes conditionnelles. Les sources se distinguent ainsi statistiquement les unes des autres :

- soit par leurs classifications,
- soit par leurs moyennes et variances,
- soit par les deux simultanément.

[B] Objectif

Connaissant les observations $\boldsymbol{X}^i (i = 1..m)$, on se propose de reconstruire et de segmenter les sources $\boldsymbol{S}^j (j = 1..n)$. Nous avons ainsi un problème inverse à deux niveaux.

1. La reconstruction des sources à partir des observations ne connaissant pas la matrice de mélange est le problème de **séparation de sources**.
2. La classification des sources (estimation des étiquettes $\boldsymbol{Z}^j (j = 1..n)$) ne connaissant pas les paramètres $\boldsymbol{\eta}^j$ est un problème de **segmentation non supervisée**.

La figure (IV.2) illustre ces deux opérations qui ont en commun l'aspect de séparation. En effet, la reconstruction des sources est une séparation le long de la dimension des capteurs et la segmentation est une séparation le long de la dimension spatiale. Un traitement optimal ne consiste pas à effectuer une séparation suivie d'une segmentation mais plutôt à mener simultanément ces deux opérations. Le formalisme bayésien donne un cadre judicieux à cette séparation conjointe et les algorithmes MCMC offre un outil efficace pour son implémentation.

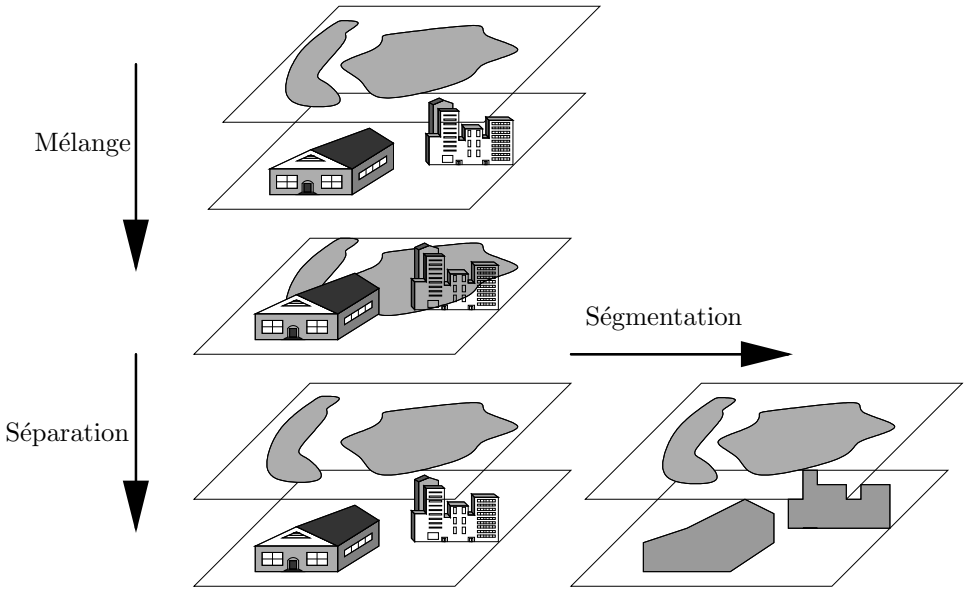


Fig. IV.2: On distingue deux types de séparation : (**i**) une séparation transversale le long des capteurs, (**ii**) une séparation spatiale le long des pixels

[C] Placement du travail

- Par rapport au chapitre (III), ce travail représente une généralisation de la modélisation des sources par des modèles de Markov cachés au cas 2-D. La propriété de Markov dans le cas 2-D est mieux traduite par une distribution de Gibbs (IV.3). Le traitement conjoint de la séparation et de la segmentation est une généralisation à deux sens.

1. Vis-à-vis du problème de séparation de sources, la segmentation peut être considérée comme une étape intermédiaire facilitant la modélisation des sources et l'exploitation de la non stationnarité spatiale pour la séparation.
2. Vis-à-vis de la segmentation, ce travail est une extension de la segmentation non supervisée au cas plus difficile où les images à segmenter ont subi une opération de mélange bruité et ne sont pas ainsi directement accessibles.

- Dans ce chapitre, on applique le critère de sélection d'*a priori* développé dans [Snoussi et Mohammad-Djafari, 2002]. On donne les expressions des distributions δ-*a priori* ainsi que les expressions des distributions *a posteriori* de la matrice de mélange, de la covariance du bruit et des moyennes et variances des gaussiennes constituant l'*a priori* des sources.
- L'échantillonneur de Gibbs présente un outil efficace permettant d'estimer conjointement les sources et leurs classifications. Une répartition particulière du vecteur des paramètres et une implémentation parallèle de l'échantillonnage du champ de Gibbs accélèrent la convergence de l'algorithme de séparation.
- Des simulations sur des données synthétiques et réelles illustrent les performances de l'algorithme proposé.

IV.2 Formulation bayésienne

IV.2.1 Distribution A Posteriori

L'objectif de départ est l'identification[2] des paramètres intervenant dans le problème décrit plus haut, à savoir la matrice de mélange $\boldsymbol{A}$, la covariance du bruit $\boldsymbol{R}_\epsilon$ et les moyennes et variances $(\mu_{jk}, \sigma^2_{jk})_{j=1..n,k=1..K}$ des gaussiennes conditionnelles modélisant l'*a priori* des sources. Le problème d'inférence associé est $\mathcal{I} := (\boldsymbol{X} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{\theta})$ où $\boldsymbol{X} = (\boldsymbol{X}^1, ..., \boldsymbol{X}^m)$ est l'ensemble d'images observées, $\boldsymbol{I}$ est toute l'information *a priori* qu'on possède sur le problème étudié et $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{R}_\epsilon, \mu_{jk}, \sigma^2_{jk})$ représente les paramètres à identifier. La distribution *a posteriori* de $\boldsymbol{\theta}$, contenant toute l'information qu'on peut extraire des données, s'écrit :

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{X}) \propto p(\boldsymbol{X} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta}).$$

Dans le paragraphe suivant, on discute l'attribution des lois *a priori*. Concernant la vraisemblance, elle possède la forme suivante :

$$\begin{aligned} p(\boldsymbol{X} \,|\, \boldsymbol{\theta}) &= \sum_{\boldsymbol{Z}} \int_{\boldsymbol{S}} p(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Z} \,|\, \boldsymbol{\theta}) d\,\boldsymbol{S} \\ &= \sum_{\boldsymbol{Z}} \left\{ \prod_{r \in \mathcal{S}} \mathcal{N}(\boldsymbol{x}_r \,;\, \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{z}_r}, \boldsymbol{A}\boldsymbol{R}_{\boldsymbol{z}_r}\boldsymbol{A}^* + \boldsymbol{R}_\epsilon) \right\} P_M(Z) \end{aligned} \tag{IV.4}$$

où $\mathcal{N}$ est la distribution gaussienne, $\boldsymbol{x}_r$ est le vecteur $(m \times 1)$ des observations au pixel r, $\boldsymbol{z}_r$ est le vecteur des étiquettes, $\boldsymbol{\mu}_{\boldsymbol{z}_r} = [\mu_{1z_1}, ..., \mu_{nz_n}]$ et $\boldsymbol{R}_{\boldsymbol{z}_r}$ est la matrice diagonale $diag[\sigma^2_{1z_1}, ..., \sigma^2_{nz_n}]$. On note que l'expression (IV.4) n'a pas une expression explicite en fonction de $\boldsymbol{\theta}$ à cause de la double intégration par rapport à $\boldsymbol{S}$ et $\boldsymbol{Z}$. Cependant, on peut bénéficier de l'augmentation naturelle des données. Les images $\boldsymbol{X}$ sont les données incomplètes et l'ensemble des sources $\boldsymbol{S}$ et des étiquettes $\boldsymbol{Z}$ représente les données manquantes.

Comme dans le cas mono-dimensionnel, l'expression (IV.4) peut être interprétée comme une moyenne d'un critère d'ajustement de matrices de covariance d'un processus non stationnaire. Le champ d'étiquettes $\boldsymbol{Z}$ est une classification vectorielle des images observées. Sachant cette classification en régions homogènes, l'opposé du logarithme de la vraisemblance complétée $\log \mathrm{p}(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\theta})$ est, à une constante additive près,

[2] La raison pour laquelle on commence par s'intéresser à l'estimation d'un paramètre de dimension finie et fixe est qu'on espère ainsi garantir la consistance et l'efficacité asymptotique quand le nombre de données augmente à l'infini.

une somme pondérée de divergences de Kullback-Leibler entre les covariances empiriques et les covariances théoriques de chaque région :

$$\frac{-\log \mathrm{p}(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\theta})}{|\mathcal{S}|} = \sum_{k=1}^{K} \alpha_k D_{KL}(\boldsymbol{\Gamma}_k, \hat{\boldsymbol{\Gamma}}_k) \; + \; cte,$$

où $\alpha_k = \frac{|\mathcal{S}_k|}{|\mathcal{S}|}$ est la proportion de la région $\mathcal{S}_k$ appartenant à la classe k. $\hat{\boldsymbol{\Gamma}}_k = \sum_{\mathcal{S}_k} \boldsymbol{x}_t \boldsymbol{x}_t^* / |\mathcal{S}_k|$ est la covariance empirique et $\boldsymbol{\Gamma}_k = \boldsymbol{A}\boldsymbol{R}_k\boldsymbol{A}^* + \boldsymbol{R}_\epsilon$ est la covariance théorique de la région k.

La diversité des sources permettant l'identification de la matrice de mélange est assurée par deux configurations.

1. Les sources ont la même classification $\boldsymbol{Z} = \boldsymbol{Z}^1 = \cdots = \boldsymbol{Z}^n$. La classification des observations est alors égale à $\boldsymbol{Z}$. Le nombre total des étiquettes K est égale au nombre des étiquettes du champ $\boldsymbol{Z}$ commun à toutes les sources (voir figure (IV.3)). Dans ce cas, la diversité des sources est assurée par la diversité des moyennes et variances des gaussiennes conditionnelles. Autrement dit, les sources ont des profils $\left[S(k) = (\mu_{jk}, \sigma^2_{jk})\right]_{k=1..K}$ distincts. C'est le principe de l'exploitation de la non stationnarité dans le cas mono-dimensionnel dans [Pham et Cardoso, 2001], sauf que nous supposons que la classification n'est pas connue.
2. Les sources n'ont pas la même classification. La classification $\boldsymbol{Z}$ des observations est alors une classification vectorielle $\boldsymbol{Z} = [\boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^n]$ et le nombre total des étiquettes est égale au produit des nombres des étiquettes de toutes les classifications : $K = \prod_{j=1}^{n} K_j$. Le fait d'avoir des classifications distinctes assure la diversité des sources. Les moyennes et les variances peuvent être les mêmes pour toutes les sources sans altérer les perfomances de la séparation.

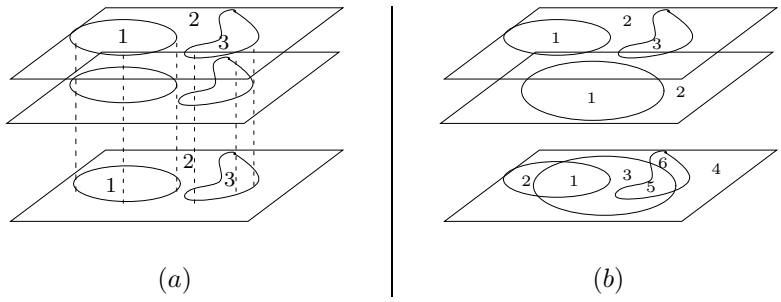


FIG. IV.3: (a)- Même classification : le nombre des étiquettes des observations est égale au nombre des étiquettes communes des sources $K = K_1 = K_2 = 3$, (b)- Classifications différentes : $K = K_1 \times K_2 = 6$

IV.2.2 Sélection d'*a priori*

Le paramètre d'intérêt est décomposé de la manière suivante : $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{R}_\epsilon, \boldsymbol{\eta})$. $\boldsymbol{A}$ est la matrice de mélange, $\boldsymbol{R}_\epsilon$ est la covariance du bruit et $\boldsymbol{\eta}$ contient tous les paramètres de la distribution des sources :

$$\begin{cases} \boldsymbol{\eta}^j = \left(\boldsymbol{\eta}^j_k\right)_{k=1..K_j} \\ \\ \boldsymbol{\eta}^j_k = (\mu^j_k, v^j_k = (\sigma^j_k)^2) \end{cases}$$

où l'indice j est le numéro de la source et k est le numéro de la gaussienne dans le mélange modélisant la distribution de la $j^{\text{ème}}$ source.

Le choix de la distribution *a priori* est fait selon un critère développé dans [Snoussi et Mohammad-Djafari, 2002]. La construction de ce critère est inspirée de la théorie de l'information et fait l'objet du chapitre (VII). On obtient une classe particulière de distributions *a priori* :

$$\Pi(\boldsymbol{\theta}) \propto e^{-\frac{\gamma_e}{\gamma_u} D_\delta(p_\theta, p_0)} \sqrt{\|g(\boldsymbol{\theta})\|} \tag{IV.5}$$

où p_θ est la vraisemblance de $\boldsymbol{\theta}$ et p_0 est une distribution de référence appartenant à l'espace entier des densités de probabilités $\mathcal{P} = \{p \mid \int p = 1\}$. $\frac{\gamma_e}{\gamma_u}$ mesure le compromis entre le degré de confiance γ_e qu'on possède sur la distribution de référence p_0 et le degré d'uniformité γ_u. $g(\boldsymbol{\theta})$ est la matrice d'information de Fisher et D_δ est la δ-divergence [Amari et Nagaoka, 2000] :

$$D_\delta(p, q) = \frac{\int p}{1-\delta} + \frac{\int q}{\delta} - \frac{\int p^\delta q^{1-\delta}}{\delta(1-\delta)}.$$

Dans la suite, l'appellation δ-*a priori* désigne la distribution (IV.5).

On suppose que la distribution de référence p_0 appartient à la famille paramétrique $\{p_\theta\}$ et elle est donc représentée par un paramètre de référence $\boldsymbol{\theta}^0 = (\boldsymbol{A}^0, \boldsymbol{R}_\epsilon^0, \boldsymbol{\eta}^0)$. La mesure de divergence entre les points de $\{p_\theta\}$ et le calcul de la matrice de Fisher sont inextricables à cause de la structure incomplète de la vraisemblance qui fait intervenir deux intégrations. Par conséquent, nous allons approcher l'expression (IV.5) en travaillant directement sur les vraisemblances complétées $\mathrm{p}(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Z} \mid \boldsymbol{\theta})$.

On commence par le calcul de la matrice d'information de Fisher.

[A] Matrice d'information de Fisher

La matrice de Fisher $g(\boldsymbol{\theta})$ est définie par :

$$g_{ij}(\boldsymbol{\theta}) = -\underset{\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T}}{E} \left[\frac{\partial^2}{\partial_i \partial_j} \log p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \mid \boldsymbol{\theta}) \right]$$

La factorisation de la distribution jointe $p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \mid \boldsymbol{\theta})$:

$$p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \mid \boldsymbol{\theta}) = p(\boldsymbol{x}_{1..T} \mid \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T}, \theta)\, p(\boldsymbol{s}_{1..T} \mid \boldsymbol{z}_{1..T}, \theta)\, p(\boldsymbol{z}_{1..T} \mid \theta)$$

et celle des espérances :

$$\underset{\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T}}{E}[.] = \underset{\boldsymbol{z}_{1..T}}{E}[.] \underset{\boldsymbol{s}_{1..T} \mid \boldsymbol{z}_{1..T}}{E}[.] \underset{\boldsymbol{x}_{1..T} \mid \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T}}{E}[.]$$

et en tenant compte des indépendances conditionnelles $((\boldsymbol{x}_{1..T} \mid \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T}) \Leftrightarrow (\boldsymbol{x}_{1..T} \mid \boldsymbol{s}_{1..T})$ et $(\boldsymbol{s}_{1..T} \mid \boldsymbol{z}_{1..T}) \Leftrightarrow \prod \boldsymbol{s}^j_{1..T} \mid \boldsymbol{z}^j_{1..T})$, on arrive à une structure bloc-diagonale de la matrice d'information de Fisher :

$$g(\boldsymbol{\theta}) = \begin{bmatrix} g(\boldsymbol{A}, \boldsymbol{R}_\epsilon) & \dots & [0] \\ \vdots & g(\boldsymbol{\eta}^1) & \\ & & \ddots \\ [0] & & \dots \quad g(\boldsymbol{\eta}^n) \end{bmatrix}$$

[A].1 Bloc $(\boldsymbol{A}, \boldsymbol{R}_\epsilon)$

La matrice d'information de Fisher relative à $(\boldsymbol{A}, \boldsymbol{R}_\epsilon)$,

$$g_{ij}(\boldsymbol{A}, \boldsymbol{R}_\epsilon) = -\underset{\boldsymbol{s}}{E} \underset{\boldsymbol{x} \mid \boldsymbol{s}}{E} \left[\frac{\partial^2}{\partial_i \partial_j} \log p(\boldsymbol{x}_{1..T} \mid \boldsymbol{s}_{1..T}, \boldsymbol{A}, \boldsymbol{R}_\epsilon) \right]$$

est très similaire à la matrice de Fisher de la moyenne et de la covariance d'une gaussienne multivariable. On obtient l'expression suivante :

$$g(\boldsymbol{A}, \boldsymbol{R}_\epsilon) = \begin{bmatrix} \left(\underset{\boldsymbol{s}_{1..T}}{E} \boldsymbol{R}_{ss} \right) \otimes \boldsymbol{R}_\epsilon^{-1} & [0] \\ [0] & -\frac{1}{2} \frac{\partial \boldsymbol{R}_\epsilon^{-1}}{\partial \boldsymbol{R}_\epsilon} \end{bmatrix}$$

où $\boldsymbol{R}_{ss} = \frac{1}{T} \sum \boldsymbol{s}_t \boldsymbol{s}_t^*$ et $\otimes$ est le produit de Kronecker.

On note la bloc-diagonalité de $g(\boldsymbol{A}, \boldsymbol{R}_\epsilon)$. Le terme correspondant à la matrice de mélange $\boldsymbol{A}$ (quantité d'information sur $\boldsymbol{A}$) est le rapport signal à bruit. Le volume induit de $(\boldsymbol{A}, \boldsymbol{R}_\epsilon)$ est alors :

$$|g(\boldsymbol{A}, \boldsymbol{R}_\epsilon)|^{1/2} \, d\,\boldsymbol{A} \, d\boldsymbol{R}_\epsilon = \frac{|\underset{\boldsymbol{\eta}}{E} \boldsymbol{R}_{ss}|^{m/2}}{|\boldsymbol{R}_\epsilon|^{\frac{m+n+1}{2}}} d\,\boldsymbol{A} \, d\boldsymbol{R}_\epsilon \tag{IV.6}$$

[A].2 Bloc $(\boldsymbol{\eta}^j)$

Chaque bloc $g(\boldsymbol{\eta}^j)$ est l'information de Fisher d'une gaussienne scalaire :

$$\left|g(\boldsymbol{\eta}^j)\right|^{1/2} d\,\boldsymbol{\eta}^j = \prod_{k=1}^{K_j} \frac{1}{v_k^{3/2}} d\,\boldsymbol{\eta}^j$$

(regarder [Snoussi et Mohammad-Djafari, 2002] pour plus de détails).

[A].3 δ-Divergence $(\delta = 0)$

Dans ce chapitre, on fixe la valeur de δ à 0. La 0-divergence entre deux paramètres $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{R}_\epsilon, \boldsymbol{\eta})$ et $\boldsymbol{\theta}^0 = (\boldsymbol{A}^0, \boldsymbol{R}_\epsilon^0, \boldsymbol{\eta}^0)$ relativement à la vraisemblance complète $p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta})$ est :

$$D_0(\boldsymbol{\theta} : \boldsymbol{\theta}^0) = \underset{x,s,z|\theta^0}{E} \log \frac{p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta}^0)}{p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta})}$$

Des développements similaires à ceux menés pour le calcul de la matrice de Fisher, en se basant sur les indépendances conditionnelles, font apparaître une forme affine de la divergence qui se met sous la forme d'une somme de la divergence moyennée entre les paramètres $(\boldsymbol{A}, \boldsymbol{R}_\epsilon)$ et de la divergence entre les paramètres des sources $\boldsymbol{\eta}$:

$$D_0(\boldsymbol{\theta} : \boldsymbol{\theta}^0) = \underset{s|\eta^0}{E} \underset{|s}{D_0}(\boldsymbol{A}, \boldsymbol{R}_\epsilon \,:\, \boldsymbol{A}^0, \boldsymbol{R}_\epsilon^0) + D_0(\boldsymbol{\eta} \,:\, \boldsymbol{\eta}^0)$$

où $\underset{|s}{D_0}$ désigne la divergence entre les distributions $p(\boldsymbol{x}_{1..T} \,|\, \boldsymbol{A}, \boldsymbol{R}_\epsilon, \boldsymbol{s}_{1..T})$ et $p(\boldsymbol{x}_{1..T} \,|\, \boldsymbol{A}^0, \boldsymbol{R}_\epsilon^0, \boldsymbol{s}_{1..T})$ en gardant les sources $\boldsymbol{s}_{1..T}$ fixées.

Compte tenu de l'indépendance des sources, la 0-divergence entre $\boldsymbol{\eta}$ et $\boldsymbol{\eta}^0$ est la somme des 0-divergences entre les paramètres $\boldsymbol{\eta}^j$ et $\boldsymbol{\eta}^{0j}$. Dans la suite, on omet l'indice j afin d'alléger les notations.

La divergence entre $\boldsymbol{\eta}$ et $\boldsymbol{\eta}^0$ est obtenue comme un cas particulier $(n = 1)$ de celle calculée dans le cas général d'une gaussienne multivariable [Snoussi et Mohammad-Djafari, 2002]. On obtient ainsi un *a priori* **normal gamma inverse** pour $\boldsymbol{\eta}$:

$$\Pi_0(\boldsymbol{\eta}) \;=\; \prod_{k=1}^K \Pi_0(\boldsymbol{\eta}_k) \;=\; \prod_{k=1}^K \mathcal{N}\left(\mu_k \,;\, \mu^0, \tfrac{v_k}{\nu^0}\right) \mathcal{G}\left(v_k^{-1} \,;\, \tfrac{\nu^0}{2}, \tfrac{\nu^0}{2} v^0\right) \tag{IV.7}$$

avec $\nu^0 = \alpha\, w_i^0$, $\alpha = \frac{\gamma_e}{\gamma_u}$, w_i^0 est la probabilité marginale de référence de l'étiquette k et $\mathcal{G}(.)$ est la distribution **gamma** :

$$\mathcal{G}(x \mid d, \beta) \propto x^{d-1} \exp\left[-\beta\, x\right].$$

La divergence moyennée entre $(\boldsymbol{A}, \boldsymbol{R}_\epsilon)$ et $(\boldsymbol{A}^0, \boldsymbol{R}_\epsilon^0)$ s'écrit :

$$\begin{aligned} \underset{s|\eta^0}{E}\, \underset{|s}{D_0}(\boldsymbol{A}, \boldsymbol{R}_\epsilon : \boldsymbol{A}^0, {\boldsymbol{R}_\epsilon}^0) = \tfrac{1}{2} \Big(\log \Big| \boldsymbol{R}_\epsilon {\boldsymbol{R}_\epsilon^0}^{-1} \Big| + \mathrm{Tr}\left(\boldsymbol{R}_\epsilon^{-1} {\boldsymbol{R}_\epsilon}^0 \right) \\ + \mathrm{Tr}\left(\boldsymbol{R}_\epsilon^{-1} (\boldsymbol{A} - \boldsymbol{A}^0) \underset{s|\eta^0}{E}[\boldsymbol{R}_{ss}] (\boldsymbol{A} - \boldsymbol{A}^0)^* \right) \Big). \end{aligned} \tag{IV.8}$$

En combinant (IV.8) avec (IV.6), on obtient l'expression de la distribution 0-*a priori* de $(\boldsymbol{A}, \boldsymbol{R}_\epsilon)$:

$$\Pi_0(\boldsymbol{A}, \boldsymbol{R}_\epsilon^{-1}) = \mathcal{N}\left(\boldsymbol{A}\,;\, \boldsymbol{A}^0, \tfrac{1}{\alpha} {\boldsymbol{R}_{ss}^0}^{-1} \otimes \boldsymbol{R}_\epsilon\right) \mathcal{W}_m\left(\boldsymbol{R}_\epsilon^{-1}\,;\, \alpha, {\boldsymbol{R}_\epsilon^0}^{-1}\right) |\underset{s|\eta}{E}[\boldsymbol{R}_{ss}]|^{\frac{m}{2}} \tag{IV.9}$$

où $\boldsymbol{R}_{ss}^0 = \underset{s|\eta^0}{E} \boldsymbol{R}_{ss}$ et $\mathcal{W}_n$ est la distribution **wishart** d'une matrice $(n \times n)$:

$$\mathcal{W}_n(\boldsymbol{R}\,;\, \nu, \boldsymbol{\Sigma}) \propto |\boldsymbol{R}|^{\frac{\nu-(n+1)}{2}} \exp\left[-\frac{\nu}{2} \mathrm{Tr}\left(\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\right)\right]$$

La distribution 0-prior est **normale inverse wishart** (*a priori* conjugué). On note que la matrice de mélange et la covariance du bruit ne sont pas *a priori* indépendants. En effet, d'après l'expression de Π_0, la covariance de $\boldsymbol{A}$ est le rapport signal sur bruit $\frac{1}{\alpha} {\boldsymbol{R}_{ss}^0}^{-1} \otimes \boldsymbol{R}_\epsilon$. La précision résultante $\alpha \boldsymbol{R}_{ss}^0 \otimes \boldsymbol{R}_\epsilon^{-1}$ autour de la matrice de référence $\boldsymbol{A}^0$ est le produit du degré de confiance α qu'on possède *a priori* et le rapport signal sur bruit. On note aussi le terme multiplicatif, dans l'expression de Π_0, qui est une puissance du déterminant de l'espérance *a priori* de la matrice de covariance des sources $\underset{s|\eta}{E}[\boldsymbol{R}_{ss}]$. Ce terme peut être injecté dans la distribution *a priori* $p(\boldsymbol{\eta})$ et les deux ensembles de paramètres $(\boldsymbol{A}, \boldsymbol{R}_\epsilon)$ et $\boldsymbol{\eta}$ sont, par conséquent, *a priori* indépendants.

IV.3 Algorithmes stochastiques

Dans le chapitre (III), nous avons considéré la même structure du problème dans le cas $1-D$ en implémentant l'algorithme EM. Cependant, dans le cas 2-D, on n'a pas l'équivalent de la procédure de Baum-Welsh et donc la première étape (*Expectation*) de l'algorithme EM n'est pas implémentable[3]. On s'oriente alors vers les techniques d'échantillonnage en considérant deux types d'algorithmes : des algorithmes de type EM et des algorithmes de type MCMC.

IV.3.1 Approximations stochastiques de l'EM

A chaque itération k, on considère trois étapes :

1. On simule M échantillons $\boldsymbol{Z}^{(m)}$ (M images $\boldsymbol{Z}$) selon la distribution *a posteriori* $\mathrm{p}(\boldsymbol{Z} \mid \boldsymbol{X}, \tilde{\boldsymbol{\theta}}^{(k)})$
2. On construit la fonctionnelle suivante :

$$\tilde{\mathcal{Q}}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}^{(k)}) = \frac{1}{M} \underset{s}{\mathrm{E}} \left[\log\, \mathrm{p}(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Z}^{(m)} \mid \boldsymbol{\theta})\right] + \log \mathrm{p}(\boldsymbol{\theta}) \tag{IV.10}$$

   On a donc une somme empirique sur les $\boldsymbol{Z}$ et une intégration exacte par rapport à $\boldsymbol{S}$.
3. On maximise la fonctionnelle pour remettre à jour le paramètre $\boldsymbol{\theta}$:

$$\tilde{\boldsymbol{\theta}}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} \tilde{\mathcal{Q}}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}^{(k)}).$$

[3] C'est l'intégration par rapport au champ d'étiquettes $\boldsymbol{Z}$ qui complique le calcul

On distingue deux cas selon la valeur de M :

1. $M \longrightarrow \infty$: on obtient un algorithme de type **MCEM** (Monte Carlo EM) qui converge vers un algorithme EM exact.
2. $M < \infty$: on obtient un algorithme de type RB-EM (Rao-Blackwellised EM). Seuls des résultats asymptotiques (lorsque le nombre d'échantillons tend vers l'infini) peuvent être dérivés dans ce cas. Ces propriétés garantissent la consistance et la normalité asymptotiques (avec une variance supérieure à l'inverse de l'information de Fisher).

On constate que dans les deux configurations précédentes, on peut dériver des résultats de convergence mais seulement asymptotiquement. Dans le cas du MCEM, la limite infinie concerne le nombre de simulations M. Dans le cas du RB-EM, la limite infinie concerne plutôt le nombre total d'échantillons.

Avec les algorithmes du type EM, le seul estimateur $\hat{\boldsymbol{\theta}}$ possible à obtenir est l'estimateur MAP du paramètre $\boldsymbol{\theta}$. L'estimation des sources et de leurs étiquettes sont alors effectuées indépendemment après la convergence vers l'estimée $\hat{\boldsymbol{\theta}}$. Ce schéma n'est pas optimal et ne rentre pas dans une méthodologie bayésienne correcte (voir chapitre (II)). L'échantillonneur de Gibbs est par contre bien adapté à ce problème à données manquantes et permet l'estimation conjointe des sources et de leurs classifications.

IV.3.2 Echantillonneur de Gibbs

On partitionne le vecteur des inconnus en deux sous-vecteurs : les variables cachées $(\boldsymbol{Z}, \boldsymbol{S})$ et le paramètre $\boldsymbol{\theta}$. Chaque cycle de l'échantillonnage de Gibbs est composé de deux simulations conditionnelles :

Echantillonneur de Gibbs

répéter jusqu'à convergence,

1. simule $(\tilde{\boldsymbol{Z}}^{(h)}, \tilde{\boldsymbol{S}}^{(h)}) \sim p(\boldsymbol{Z}, \boldsymbol{S} \,|\, \boldsymbol{X}, \tilde{\boldsymbol{\theta}}^{(h-1)})$

2. simule $\tilde{\boldsymbol{\theta}}^{(h)} \sim p(\boldsymbol{\theta} \,|\, \boldsymbol{X}, \tilde{\boldsymbol{Z}}^{(h)}, \tilde{\boldsymbol{S}}^{(h)})$

(IV.11)

Sous des conditions faibles, liées principalement à la connectivité du support de la loi jointe, l'algorithme (IV.11) produit une chaîne de Markov $(\tilde{\boldsymbol{\theta}}^{(h)})$ ergodique de distribution stationnaire $p(\boldsymbol{\theta} \,|\, \boldsymbol{X})$. D'après le théorème (2) du chapitre (II), les sommes empiriques $\sum_{h=1}^{H} f(\tilde{\boldsymbol{\theta}}^{(h)}) \Big/ H$ tendent vers les espérances *a posteriori* $\mathrm{E}\big[f(\boldsymbol{\theta}) \,|\, \boldsymbol{X}\big]$ quand H tend vers l'infini. Cependant, en pratique, on ne peut pas considérer une infinité de termes. Après h_0 itérations (temps de chauffe), on suppose que les échantilllons $(\tilde{\boldsymbol{\theta}}^{(h_0+h)})$ suivent approximativement la loi *a posteriori* $p(\boldsymbol{\theta} \,|\, \boldsymbol{X})$ et on approche les espérances *a posteriori* par :

$$\mathrm{E}\big[f(\boldsymbol{\theta}) \,|\, \boldsymbol{X}\big] \approx \frac{1}{H} \sum_{h=1}^{H} f(\tilde{\boldsymbol{\theta}}^{(h_0+h)}). \tag{IV.12}$$

Echantillonnage de $(\boldsymbol{Z}, \boldsymbol{S})$: D'après la règle séquentielle de Bayes,

$$p(\boldsymbol{Z}, \boldsymbol{S} \,|\, \boldsymbol{X}, \boldsymbol{\theta}) = p(\boldsymbol{S} \,|\, \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\theta})\, p(\boldsymbol{Z} \,|\, \boldsymbol{X}, \boldsymbol{\theta}),$$

l'échantillonnage exact de la distribution *a posteriori* jointe est obtenu par un échantillonnage de la loi marginale $p(\boldsymbol{Z} \,|\, \boldsymbol{X}, \boldsymbol{\theta})$ suivi par un échantillonnage de la distribution conditionnelle $p(\boldsymbol{S} \,|\, \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\theta})$.

1. On simule $\tilde{\boldsymbol{Z}}$ selon sa distribution *a posteriori* marginale (en intégrant par rapport aux sources),

$$p(\boldsymbol{Z} \,|\, \boldsymbol{X}, \boldsymbol{\theta}) \propto p(\boldsymbol{X} \,|\, \boldsymbol{Z}, \boldsymbol{\theta})\, P_M(\boldsymbol{Z}). \tag{IV.13}$$

Dans l'expression (IV.13), on note que le champ $\boldsymbol{Z}$ des étiquettes vectorielles possède *a posteriori* deux sortes de dépendances induites, d'une manière complémentaire, par la vraisemblance et l'*a priori*.

- Une dépendance le long des pixels est induite par la distribution *a priori*. En effet, $p(\boldsymbol{Z}) = \prod_{j=1}^{n} p(\boldsymbol{Z}^j)$ et donc les étiquettes vectorielles $\boldsymbol{Z}$ ont une structure markovienne dont le système de voisinage est l'union des systèmes de voisinage des champs $\boldsymbol{Z}^j$.
- Une dépendance le long des capteurs est induite par la vraisemblance. En effet, conditionnellement à $\boldsymbol{Z}$, le champ observé $\boldsymbol{X}$ est indépendant le long des pixels $p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\theta}) = \prod_{r \in \mathcal{S}} p(\boldsymbol{x}_r \mid \boldsymbol{z}_r, \boldsymbol{\theta})$ mais, à chaque pixel r, ses composantes x_r^i sont dépendantes le long des capteurs à cause de l'opération de mélange,
$$p(\boldsymbol{x}_r \mid \boldsymbol{z}_r, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}_r\,;\, \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{z}_r}, \boldsymbol{A}\boldsymbol{R}_{\boldsymbol{z}_r}\boldsymbol{A}^* + \boldsymbol{R}_\epsilon)$$
où $\boldsymbol{z}_r$ est le vecteur des étiquettes sur le site r, $\boldsymbol{\mu}_{\boldsymbol{z}_r} = [\mu_{1z_1}, ..., \mu_{nz_n}]$ et $\boldsymbol{R}_{\boldsymbol{z}_r}$ est la matrice diagonale $diag[\sigma^2_{1z_1}, ..., \sigma^2_{nz_n}]$.

2. Sachant $\tilde{\boldsymbol{Z}}$, on simule $\tilde{\boldsymbol{S}}$ selon sa loi *a posteriori* conditionnelle :
$$p(\boldsymbol{S} \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}) = \prod_{r \in \mathcal{S}} \mathcal{N}(\boldsymbol{s}_r\,;\, \boldsymbol{m}_r^{apost}, \boldsymbol{V}_r^{apost})$$
où les moyennes et les covariances *a posteriori* sont simples à calculer [Snoussi et Mohammad-Djafari, 2000],
$$\begin{aligned} \boldsymbol{V}_r^{apost} &= \left[\boldsymbol{A}^*\boldsymbol{R}_\epsilon^{-1}\boldsymbol{A} + \boldsymbol{R}_{\boldsymbol{z}_r}^{-1}\right]^{-1} \\ \boldsymbol{m}_r^{apost} &= \boldsymbol{V}_r^{apost}\left(\boldsymbol{A}^*\boldsymbol{R}_\epsilon^{-1}\boldsymbol{x}_r + \boldsymbol{R}_{\boldsymbol{z}_r}^{-1}\boldsymbol{\mu}_{\boldsymbol{z}_r}\right) \end{aligned} \tag{IV.14}$$

Echantillonnage de $\boldsymbol{\theta}$: Connaissant les observations $\boldsymbol{X}$, les sources $\boldsymbol{S}$ et les classifications $\boldsymbol{Z}$ (simulées dans la première étape), l'échantillonnage du paramètre $\boldsymbol{\theta}$ est simple à effectuer (c'est la raison pour laquelle on a introduit les variables cachées $\boldsymbol{S}$ et $\boldsymbol{Z}$). La distribution conditionnelle $p(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{S})$ se factorise en deux termes,
$$p(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{S}) \propto p(\boldsymbol{A}, \boldsymbol{R}_\epsilon \mid \boldsymbol{X}, \boldsymbol{S})\, p(\boldsymbol{\mu}, \boldsymbol{\sigma} \mid \boldsymbol{S}, \boldsymbol{Z}),$$
conduisant à un découplage entre l'échantillonnage de $(\boldsymbol{A}, \boldsymbol{R}_\epsilon)$ et $(\boldsymbol{\mu}, \boldsymbol{\sigma})$. En choisissant les 0-*a priori* développés dans la section précédente, la distribution *a posteriori* de $\boldsymbol{\theta}$ possède la forme suivante :

- **inverse wishart** pour la covariance du bruit et **inverse gamma** pour les variances des sources,
- **gaussienne** pour la matrice de mélange et pour les moyennes des sources.

Les expressions de ces distributions sont développées dans l'annexe 1 page 81. On note que la forme inverse wishart des lois des matrices de covariance élimine le risque de dégénérescence mentionné dans le chapitre précédent (III) et repris en détail dans le chapitre (VI). On donne, ci-après, les expressions des distributions *a posteriori* correspondantes aux paramètres $(\boldsymbol{A}, \boldsymbol{R}_\epsilon)$ dans le cas particulier d'un *a priori* de Jeffreys :
$$\begin{cases} \boldsymbol{R}_\epsilon^{-1} \sim \mathcal{W}_m(\nu_p, \boldsymbol{\Sigma}_P), \;\; \nu_p = \frac{|\mathcal{S}|-n}{2}, \;\; \boldsymbol{\Sigma}_p = \frac{|\mathcal{S}|}{2}(\boldsymbol{R}_{xx} - \boldsymbol{R}_{xs}\boldsymbol{R}_{ss}^{-1}\boldsymbol{R}_{xs}^*) \\ p(\boldsymbol{A} \mid \boldsymbol{R}_\epsilon) \sim \mathcal{N}(\boldsymbol{A}_p, \boldsymbol{\Gamma}_p), \;\; \boldsymbol{A}_p = \boldsymbol{R}_{xs}\boldsymbol{R}_{ss}^{-1}, \;\; \boldsymbol{\Gamma}_p = \frac{1}{|\mathcal{S}|}\boldsymbol{R}_{ss}^{-1} \otimes \boldsymbol{R}_\epsilon \end{cases} \tag{IV.15}$$
où on a défini les sommes empiriques $\boldsymbol{R}_{xx} = \frac{1}{|\mathcal{S}|}\sum_r \boldsymbol{x}_r\boldsymbol{x}_r^*$, $\boldsymbol{R}_{xs} = \frac{1}{|\mathcal{S}|}\sum_r \boldsymbol{x}_r\boldsymbol{s}_r^*$ et $\boldsymbol{R}_{ss} = \frac{1}{|\mathcal{S}|}\sum_r \boldsymbol{s}_r\boldsymbol{s}_r^*$ (les sources $\boldsymbol{S}$ sont générées dans la première étape de l'échantillonneur de Gibbs). On note que la matrice de covariance de la matrice de mélange est proportionnelle à l'inverse du rapport signal à bruit. Ceci peut expliquer une lenteur de convergence dans les conditions d'un fort rapport signal sur bruit.

Remarque 9 *La distribution a posteriori* $\mathrm{p}(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})$ *est un champ de Gibbs avec le même voisinage* ∂ *que celui du champ de Gibbs a priori* $P_G(\boldsymbol{Z})$ *(puisque la vraisemblance n'introduit pas de dépendance spatiale).*

Par conséquent, l'échantillonnage exact de cette loi (dans la première étape des algorithmes stochastiques ou la première étape de l'échantillonneur de Gibbs) n'est pas possible. On peut alors implémenter un échantillonneur de Gibbs (ou un autre algorithme de type MCMC*) à chaque itération des algorithmes décrits plus haut pour obtenir un échantillon de* $\mathrm{p}(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})$*. Cependant, cette procédure est très coûteuse puisque l'obtention d'un échantillon exact n'est garantie qu'asymptotiquement. La solution retenue consiste à se contenter d'un seul cycle de l'échantillonneur de Gibbs à chaque itération. A l'itération* k *de chacun des algorithmes proposés plus haut, l'échantillonnage :*

$$\tilde{\boldsymbol{Z}} \sim \mathrm{p}(\boldsymbol{Z} \mid \boldsymbol{X}, \tilde{\boldsymbol{\theta}}^{(k-1)})$$

est remplacé par :

$$\begin{cases} \text{pour tout } \ r \in \mathcal{S}, \\ \\ Z_r \sim \mathrm{p}(Z_r \mid \boldsymbol{Z}_{\mathcal{S}\setminus r}\boldsymbol{X}, \tilde{\boldsymbol{\theta}}^{(k-1)}) \end{cases} \tag{IV.16}$$

Les points suivants résument l'impact de cette modification sur chacun des algorithmes proposés :

1. MCEM *: l'algorithme* MCEM*(Monte Carlo* EM*) n'est pas affecté par cette limitation. En effet, la première étape de cet algorithme repose sur la simulation d'une infinité de réalisations de* $\boldsymbol{Z}$ *(*$M \longrightarrow \infty$*) et d'approcher la fonctionnelle* $\mathcal{Q}$ *de l'*EM *par une moyenne empirique. Un algorithme* MCMC *garantit les mêmes performances en approchant la fonctionnelle de l'*EM *par une moyenne empirique sur une chaîne de Markov (voir le théorème (2) du chapitre (II)).*
2. RB-EM *: remplacer l'échantillonnage exact dans la première étape de l'algorithme* RB-EM *par un seul cycle (IV.16) de l'échantillonneur de Gibbs modifie l'algorithme. Cependant, en pratique, cette version modifiée garde de bonnes performances. Ce qui peut se comprendre intuitivement. En effet, puisque le paramètre* $\tilde{\boldsymbol{\theta}}^{(k)}$ *change d'une itération à l'autre, on n'a pas vraiment besoin d'un échantillonnage exact de* $\mathrm{p}(\boldsymbol{Z} \mid \boldsymbol{X}, \tilde{\boldsymbol{\theta}}^{(k)})$*. En plus, bien qu'on n'arrive pas encore à prouver la consistance asymptotique de cette modification, on ne peut pas affirmer sa sous-optimalité par rapport à la version exacte.*
3. *Echantillonneur de Gibbs : théoriquement, cette modification rentre dans le principe de l'échantillonneur de Gibbs. En effet, exécuter un seul cycle (IV.16) revient à repartitionner le vecteur des paramètres. Avant, la partition était en deux sous-vecteurs :* $\boldsymbol{V}_1 = (\boldsymbol{Z})$ *et* $\boldsymbol{V}_2 = (\boldsymbol{S}, \boldsymbol{\theta})$*. Avec un seul cycle (IV.16), la partition est en* $|\mathcal{S}| + 1$ *sous-vecteurs :* $\boldsymbol{V}_r = (Z_r), r \in \mathcal{S}$ *et* $\boldsymbol{V}_{|\mathcal{S}|+1} = (\boldsymbol{S}, \boldsymbol{\theta})$*. Concernant les performances, cette modification risque de ralentir l'algorithme de séparation. En plus, on n'a plus la propriété de dualité [Robert, 1996].*

Remarque 10 *Dans le cas d'un voisinage* ∂ *d'ordre* 1*, l'échantillonnage du champ de Gibbs* $\mathrm{p}(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})$ *peut être implémenté en parallèle [Winkler, 1995]. L'ensemble des sites* $\mathcal{S}$ *est partionné en deux sous-ensembles les* **noirs** *et les* **blancs** *(en échiquier, voir la figure (IV.4)). En fixant les noirs (ou les blancs), les blancs (ou les noirs) sont indépendants et peuvent être échantillonnés en parallèle. Le cycle (IV.16) contient désormais uniquement deux étapes. L'algorithme de séparation est le suivant.*

Echantillonneur parallèle de Gibbs

à l'itération h

1. *simule* $\boldsymbol{Z}_N^{(h)} \sim \mathrm{p}(\boldsymbol{Z}_N \mid \boldsymbol{Z}_B^{(h-1)}, \boldsymbol{X}, \boldsymbol{\theta}^{(h-1)})$

   simule $\boldsymbol{Z}_B^{(h)} \sim \mathrm{p}(\boldsymbol{Z}_B \mid \boldsymbol{Z}_N^{(h)}, \boldsymbol{X}, \boldsymbol{\theta}^{(h-1)})$ (IV.17)

   simule $\boldsymbol{S}^{(h)} \sim \mathrm{p}(\boldsymbol{S} \mid \boldsymbol{Z}^{(h)}, \boldsymbol{X}, \boldsymbol{\theta}^{(h-1)})$

2. *simule* $\boldsymbol{\theta}^{(h)} \sim \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{S}^{(h)}, \boldsymbol{Z}^{(h)})$

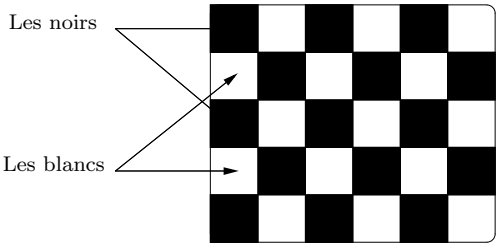


FIG. IV.4: Implémentation parallèle en échiquier

IV.3.3 CONTRÔLE DE CONVERGENCE

Le contrôle de convergence d'une chaîne de Markov est une question délicate [Brooks et Roberts, 1995]. Beaucoup d'outils de contrôle ont été développés dans la litérature des méthodes MCMC. Cependant, aucune méthode n'est préconisée [Robert, 1996]. En effet, avec ces méthodes, on peut détecter la non convergence de la chaîne de Markov mais on ne peut pas affirmer sa convergence. La validité d'une méthode dépend fortement du problème traité. On se contente, dans la suite, de rappeler quelques outils simples de contrôle.

[A] VISUALISATION DE LA CHAÎNE

C'est la méthode la plus simple qui consiste à tracer la série $\tilde{\boldsymbol{\theta}}^{(h)}$ en fonction de h. On essaie de détecter à l'œil si la série tend vers un comportement stationnaire. On constate qu'avec cette méthode on ne peut pas affirmer objectivement la convergence de la chaîne mais on peut détecter un comportement non stationnaire reflétant la non convergence.

[B] SOMMES EMPIRIQUES

On peut aussi tracer les sommes empiriques d'une quantité d'intérêt $f(\boldsymbol{\theta})$:

$$S_H = \frac{1}{H} \sum_{h=1}^{H} f(\tilde{\boldsymbol{\theta}}^{(h)})$$

en fonction de H. La série des sommes cumulées $(S_H)_{H \in \mathbb{N}}$ doit converger vers $\mathrm{E}_g\big[f(\boldsymbol{\theta})\big]$ quand $H \longrightarrow \infty$ avec g la loi stationnaire de la chaîne de Markov $(\tilde{\boldsymbol{\theta}}^{(h)})$.

[C] Rao-Blackwellisation

Si la chaîne d'intérêt $(\tilde{\boldsymbol{\theta}}^{(h)})$ est obtenue à partir d'une autre chaîne $\boldsymbol{\eta}^{(h)}$ (comme c'est le cas dans les échantillonneurs de Gibbs), la quantité $\mathrm{E}_g\big[f(\boldsymbol{\theta})\big]$ peut être approchée par la somme cumulée suivante :

$$S_H^{rb} = \frac{1}{H}\sum_{h=1}^{H} \mathrm{E}\left[f(\boldsymbol{\theta}) \mid \boldsymbol{\eta}^{(h)}\right]$$

qui est une sorte de conditionnement appelée Rao-Blackwellisation par référence au théorème de Rao-Blackwell [Lehmann et Casella, 1996].

Dans le cas de l'augmentation de données :

1. simule $\tilde{\boldsymbol{\theta}}^{(h)} \sim \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{\eta}^{(h-1)})$
2. simule $\boldsymbol{\eta}^{(h)} \sim \mathrm{p}(\boldsymbol{\eta} \mid \tilde{\boldsymbol{\theta}}^{(h)})$

on montre dans [Liu et Pierce, 1994] que l'estimateur S_H^{rb} domine S_H en terme de variance :

$$var(S_H^{rb}) \leq var(S_H)$$

Dans le cas de la séparation de sources, on peut calculer la somme cumulée Rao-Blackwellisée de $\boldsymbol{A}$ et de $\boldsymbol{S}$. En effet, $\boldsymbol{S}$ est un champ gaussien (*a posteriori* connaissant $(\boldsymbol{X}, \tilde{\boldsymbol{A}}, \tilde{\boldsymbol{Z}})$) de moyenne $(\boldsymbol{m}_r^{apost})_{r\in\mathcal{S}}$ (IV.14). La somme Rao-Blackwellisée S_H^{rb} s'écrit, à chaque pixel r,

$$\begin{aligned}
S_H^{rb}(r) &= \tfrac{1}{H}\sum_{h=1}^{H} \mathrm{E}\left[\boldsymbol{s}(r) \mid \boldsymbol{X}, \boldsymbol{Z}^{(h)}, \tilde{\boldsymbol{\theta}}^{(h)}\right] \\
&= \tfrac{1}{H}\sum_{h=1}^{H} \boldsymbol{m}_r^{apost} \\
&= \tfrac{1}{H}\sum_{h=1}^{H} \left[\boldsymbol{A}^* \boldsymbol{R}_\epsilon^{-1} \boldsymbol{A} + \boldsymbol{R}_{\boldsymbol{z}_r}^{-1}\right]^{-1} \left(\boldsymbol{A}^* \boldsymbol{R}_\epsilon^{-1} \boldsymbol{x}_r + \boldsymbol{R}_{\boldsymbol{z}_r}^{-1} \boldsymbol{\mu}_{\boldsymbol{z}_r}\right)
\end{aligned}$$

où les paramètres $\tilde{\boldsymbol{\theta}}^{(h)} = (\boldsymbol{A}, \boldsymbol{R}_\epsilon, \boldsymbol{R}_k, \boldsymbol{\mu}_k)$ et le champ $(\boldsymbol{z}_r)_{r\in\mathcal{S}}$ évoluent à chaque itération h.

Concernant la matrice $\boldsymbol{A}$, en choisissant l'*a priori* Π_0 (voir l'expression (IV.9)), sa distribution *a posteriori* est gaussienne ((IV.18) de l'annexe 1 page (81)). Afin d'alléger les notations, on choisit le cas particulier de $\alpha = 0$ (*a priori* de Jeffreys). La somme cumulée Rao-Blackwellisée de la matrice de mélange s'écrit,

$$\begin{aligned}
S_H^{rb} &= \tfrac{1}{H}\sum_{h=1}^{H} \mathrm{E}\left[\boldsymbol{A} \mid \boldsymbol{X}, \boldsymbol{Z}^{(h)}, \boldsymbol{S}^{(h)}\right] \\
&= \tfrac{1}{H}\sum_{h=1}^{H} \boldsymbol{R}_{xs} \boldsymbol{R}_{ss}^{-1}
\end{aligned}$$

Remarque 11 *Dans le cas de l'implémentation parallèle (IV.17), la chaîne $\tilde{\boldsymbol{\theta}}^{(h)}$ n'est pas obtenue par un algorithme d'augmentation de données. En effet, l'échantillonnage de $\boldsymbol{Z}$ n'est pas exact. Par conséquent, l'estimateur S_H^{rb} ne domine pas forcément S_H. Cependant, S_H^{rb} constitue un autre outil de contrôle de convergence.*

[D] CUSUM Plot

Pour une statistique scalaire $T(\boldsymbol{\theta})$, on considère la série suivante :

$$\hat{S}_H = \sum_{h=h_0+1}^{H} \left[T(\tilde{\boldsymbol{\theta}}^{(h)}) - \hat{\mu} \right], \;\; \hat{\mu} = (H-h_0)^{-1} \sum_{h=h_0+1}^{H} T(\tilde{\boldsymbol{\theta}}^{(h)})$$

où on commence à partir de h_0 ("temps de chauffe") afin d'éliminer le biais initial. On peut estimer grossièrement h_0 en visualisant directement la série $\{T(\tilde{\boldsymbol{\theta}}^{(h)})\}$.

Le graphe CUSUM suggérée par [Yu et Mykland, 1994] consiste à tracer la série $\{\hat{S}_H\}$ en fonction de H et de connecter les points successifs par des segments. La vitesse de convergence de la chaîne de Markov $\{T(\tilde{\boldsymbol{\theta}}^{(h)})\}$ est liée à la douceur du graphe CUSUM. Plus les variations du graphe sont rapides plus la chaîne converge rapidement et plus les variations sont lentes plus la chaîne converge lentement.

IV.4 Résultats de simulation

On commence par illustrer les performances de l'échantillonneur de Gibbs sur des simulations synthétiques. On génère deux champs 64×64 d'étiquettes suivant le modèle de Potts :

$$P_M(\boldsymbol{Z}^j) = [W(\alpha_j)]^{-1} \exp\{\alpha_j \sum_{r \sim s} I_{z_r = z_s}\}, \alpha_j = 2,$$

où le voisinage d'un pixel est formé par les 4 pixels les plus proches. La valeur de $\alpha_j = 2$, supposée connue, implique une structure homogène (voir première ligne de la figure (IV.5)). La première source possède 3 couleurs (3 gaussiennes) tandis que la deuxième source possède deux couleurs (modèle de Ising).

Conditionnellement à $\boldsymbol{Z}$, les sources à valeurs dans $\mathbb{R}$ suivent des lois gaussiennnes de moyennes $\mu_1 = \begin{bmatrix} -3 & 0 & 3 \end{bmatrix}$ et variances $\sigma_1 = \begin{bmatrix} 1 & 0.3 & 0.5 \end{bmatrix}$ pour la première source et $\mu_2 = \begin{bmatrix} -3 & 3 \end{bmatrix}$, $\sigma_2 = \begin{bmatrix} 0.1 & 2 \end{bmatrix}$ pour la deuxième source.

Les sources sont ensuite mélangées avec la matrice $\boldsymbol{A} = \begin{bmatrix} 0.85 & 0.44 \\ 0.50 & 0.89 \end{bmatrix}$. Un bruit gaussien de covariance $\boldsymbol{R}_\epsilon = \begin{bmatrix} 3 & 1 \\ 1 & 5 \end{bmatrix}$ est ajouté au mélange linéaire (RSB= 1 à 3 dB). La figure (IV.5) montre les étiquettes discrètes, les sources originales et les sources mélangées observées sur les détecteurs.

On applique l'échantillonneur de Gibbs décrit dans la section (IV.3.2) pour obtenir la chaîne de Markov $(\boldsymbol{A}^{(h)}, \boldsymbol{R}_\epsilon^{(h)}, \mu_{jk}^{(h)}, \sigma_{jk}^{2\ (h)})$. La figure (IV.6) illustre les histogrammes représentant approximativement les distributions marginales concentrées autour de la vraie valeur de la matrice de mélange. Sur le même graphe, on note la convergence des moyennes empiriques après 2000 itérations de l'algorithme. Les figures (IV.7), (IV.8) et (IV.9) montrent la convergence des moyennes empiriques des paramètres des sources et de la covariance du bruit. On peut noter que la convergence des variances est plus lente que celle des cœfficients du mélange ou des moyennes des sources. Dans la figure (IV.10), on a montré un échantillon de la distribution *a posteriori* des sources et des étiquettes. En les comparant aux valeurs originales, on note le succès de l'algorithme proposé à reconstruire les sources ainsi que leurs classifications.

Nous avons testé l'algorithme proposé sur des images réelles en simulant le mélange. La première source représente une portion de la terre observée par satellite et la deuxième source représente des nuages. La figure (IV.11)) contient :

- les vraies sources sur la première ligne,
- les sources mélangées et bruitées sur la deuxième ligne,
- les sources reconstruites sur la troisième ligne,
- les résultats de la segmentation sur la dernière ligne.

On note la bonne qualité de la séparation des sources. Les résultats de la segmentation sont presque les mêmes que si on segmente directement les sources non mélangées.

IV.5 Conclusion

Dans ce chapitre, nous avons considéré le problème de séparation d'images. Le mélange est linéaire, instantané et bruité. Le point de départ de ce travail est la modélisation des sources par des champs de Markov cachés. Les avantages de cette modélisation sont multiples.

Performances de séparation

Concernant les performances de séparation, l'introduction des champs de variables discrètes $(\boldsymbol{Z}^j)_{j=1..n}$ permet :

1. de tenir compte de la corrélation spatiale des sources via la structure markovienne des champs des étiquettes,
2. d'exploiter la non stationnarité des sources via l'interprétation des champs des étiquettes comme un processus de classification (une classification commune ou plusieurs classifications indépendantes).

Séparation et segmentation simultanées

La non connaissance des champs $(\boldsymbol{Z}^j)_{j=1..n}$ (deuxième attribut des sources) a introduit une deuxième couche de variables cachées (la première est celle des sources recherchées). Par conséquent, le problème d'inférence de départ ($\mathcal{I} := (\boldsymbol{X} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{S})$) inclut désormais le problème de segmentation des sources ($\mathcal{I} := (\boldsymbol{X} \wedge \boldsymbol{I} \longrightarrow \boldsymbol{S} \wedge \boldsymbol{Z})$). On a donc deux problèmes de séparation :

- une séparation spatiale de chaque image, le long des pixels, qui s'appuie sur la diversité des statistiques d'ordre deux (les moyennes et les variances des gaussiennes sont distinctes),
- une séparation le long des capteurs (séparation de sources) qui s'appuie sur la diversité des statistiques multivariables d'ordre deux (éventuellement induite par des classifications distinctes).

La segmentation peut être interprétée comme un artifice pour améliorer la séparation des images (comme on l'a mentionné plus haut). Réciproquement, ce travail peut être considéré comme une généralisation du problème de la segmentation au cas plus difficile où les images à segmenter ne sont pas directement accessibles et ont subi un mélange linéaire bruité.

La formulation bayésienne offre un cadre naturel à la séparation et la segmentation simultanées des sources. En effet, l'introduction des champs cachés d'étiquettes est interprétée comme une représentation hiérarchique (voir chapitre (II)) visant à expliquer logiquement le processus de génération des sources.

Aspect algorithmique

La classification bayésienne des images et l'identification des moyennes et des variances des gaussiennes conditionnelles constituent un problème à variables cachées de même nature que celui de la séparation de sources. Par conséquent, l'aspect algorithmique d'une séparation et d'une segmentation simulanées n'est pas plus compliqué que celui d'une séparation avec une classification connue ou une segmentation directe des sources. A titre d'exemple, l'échantillonneur de Gibbs parallèle (*IV*.17) peut être utilisé exclusivement pour la segmentation d'images (en fixant la matrice de mélange lors de l'échantillonnage de $\boldsymbol{\theta}$) ou exclusivement pour la séparation d'images (en fixant la classification $\boldsymbol{Z}$).

On note que l'algorithme de séparation proposé inclut implicitement le débruitage des images en estimant aussi la matrice de covariance du bruit.

Bibliographie

[Amari et Nagaoka, 2000] S. Amari et H. Nagaoka. *Methods of Information Geometry*, volume 191 of Translations of Mathematical Monographs. AMS, OXFORD, University Press, 2000.

[Brooks et Roberts, 1995] S. Brooks et G. Roberts. Diagnosing convergence of Markov chain Monte Carlo algorithms. Technical report no. 95-12, Stat. Lab., U. of Cambridge, 1995.

[Comon, 1994] P. Comon. Independent Component Analysis, a new concept ? *Signal processing, Special issue on Higher-Order Statistics, Elsevier*, 36 (3) : 287–314, avril 1994.

[Darmois, 1953] G. Darmois. Analyse Générale des Liaisons Stochastiques. *Rev. Inst. Internat. Stat.*, 21 : 2–8, 1953.

[Hammersley et Clifford, 1968] J. M. Hammersley et P. Clifford. Markov fields of finite graphs and lattices. Rapport interne, University of California-Berkeley, preprint, 1968.

[Lehmann et Casella, 1996] E. Lehmann et G. Casella. *Theory of point estimation (revised edition)*. Chapman and Hall, New York, NY, USA, 1996.

[Liu et Pierce, 1994] Q. Liu et D. A. Pierce. A note on Gauss-Hermite quadrature. *J. Amer. Statist. Assoc.*, 81 (3) : 624–629, 1994.

[Pham et Cardoso, 2001] D.-T. Pham et J. Cardoso. Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. Signal Processing*, 49, 9 (11) : 1837–1848, 2001.

[Robert, 1996] C. Robert. *Méthodes de Monte-Carlo par chaînes de Markov*. Economica, Paris, 1996.

[Snoussi et Mohammad-Djafari, 2000] H. Snoussi et A. Mohammad-Djafari. Bayesian source separation with mixture of Gaussians prior for sources and Gaussian prior for mixture coefficients. In A. Mohammad-Djafari, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 388–406, Gif-sur-Yvette, juillet 2000. Proc. of MaxEnt, Amer. Inst. Physics.

[Snoussi et Mohammad-Djafari, 2002] H. Snoussi et A. Mohammad-Djafari. Information Geometry and Prior Selection. In C. Williams, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 307–327. MaxEnt Workshops, Amer. Inst. Physics, août 2002.

[Winkler, 1995] G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*. Springer Verlag, Berlin, Allemagne, 1995.

[Yu et Mykland, 1994] B. Yu et P. Mykland. Looking at Markov samplers through Cusum Path Plots : A simple diagnostic idea. Technical report no. 9413, Department of Statistics, U. of California, Berkeley, 1994.

Annexe 1 : Distributions *a posteriori*

$(\boldsymbol{A}, \boldsymbol{R}_\epsilon)$

Selon la règle de Bayes, la distribution *a posteriori* des paramètres $(\boldsymbol{A}, \boldsymbol{R}_\epsilon)$ s'écrit :

$$\begin{aligned} p(\boldsymbol{A}, \boldsymbol{R}_\epsilon \,|\, \boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Z}) &\propto p(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Z} \,|\, \boldsymbol{A}, \boldsymbol{R}_\epsilon)\Pi_0(\boldsymbol{A}, \boldsymbol{R}_\epsilon) \\ &\propto p(\boldsymbol{X} \,|\, \boldsymbol{S}, \boldsymbol{A}, \boldsymbol{R}_\epsilon)\Pi_0(\boldsymbol{A}, \boldsymbol{R}_\epsilon) \end{aligned}$$

La distribution *a priori* Π_0 présente les mêmes avantages qu'un *a priori* conjugué. Autrement dit, la distribution *a posteriori* appartient à la même famille que celle de la distribution *a priori*. Dans notre cas, c'est la famille **normale inverse wishart** :

$$p(\boldsymbol{A}, \boldsymbol{R}_\epsilon \,|\, \boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Z}) = \mathcal{N}(\boldsymbol{A}\,;\, \boldsymbol{A}_p, \boldsymbol{\Gamma}_p)\mathcal{W}_m(\boldsymbol{R}_\epsilon^{-1}\,;\, \nu_p, \boldsymbol{\Sigma}_p) \tag{IV.18}$$

dont les paramètres sont mis à jour selon les équations suivantes :

$$\begin{cases} \nu_p = K + \alpha,\ (K = |\mathcal{S}|, \alpha = \frac{\gamma_e}{\gamma_u}) \\ Vec(\boldsymbol{A}_p) = \left[\boldsymbol{R}_v^{-1} + \boldsymbol{R}_a^{-1}\right]^{-1}\left[\boldsymbol{R}_v^{-1}\, Vec(\boldsymbol{A}_v) + \boldsymbol{R}_a^{-1}\, Vec(\boldsymbol{A}_0)\right] \\ \boldsymbol{\Gamma}_p^{-1} = \boldsymbol{R}_v^{-1} + \boldsymbol{R}_a^{-1} \\ \boldsymbol{R}_v = K^{-1}\boldsymbol{R}_{ss}^{-1} \otimes \boldsymbol{R}_\epsilon \\ \boldsymbol{R}_a = \alpha^{-1}{\boldsymbol{R}_{ss}^0}^{-1} \otimes \boldsymbol{R}_\epsilon \\ \boldsymbol{A}_v = \boldsymbol{R}_{xs}\,\boldsymbol{R}_{ss}^{-1} \\ \boldsymbol{\Sigma}_p^{-1} = \frac{1}{K+\alpha}\left[k\hat{\boldsymbol{R}}_\epsilon + \alpha\boldsymbol{R}_0 + (\boldsymbol{A}_0 - \boldsymbol{A}_v)(K^{-1}\boldsymbol{R}_{ss}^{-1} + \alpha^{-1}{\boldsymbol{R}_{ss}^0}^{-1})^{-1}(\boldsymbol{A}_0 - \boldsymbol{A}_v)^T\right] \\ \hat{\boldsymbol{R}}_\epsilon = \boldsymbol{R}_{xx} - \boldsymbol{R}_{xs}\,\boldsymbol{R}_{ss}^{-1}\,\boldsymbol{R}_{sx} \end{cases}$$

Les statistiques $\boldsymbol{R}_{xs}$ et $\boldsymbol{R}_{ss}$ sont calculées à partir des sources simulées dans la première étape de l'échantillonneur de Gibbs . $\boldsymbol{R}_{ss}^0$ est l'espérance *a priori* de la matrice $\boldsymbol{R}_{ss}$:

$$\boldsymbol{R}_{ss}^0 = \underset{s|\eta^0}{E}[\boldsymbol{R}_{ss}].$$

$(\mu_k, v = \sigma_k^2)$

Des calculs similaires à ceux menés dans le paragraphe précédent conduisent à une forme **normale gamma inverse** de la loi *a posteriori* des moyennes et variances :

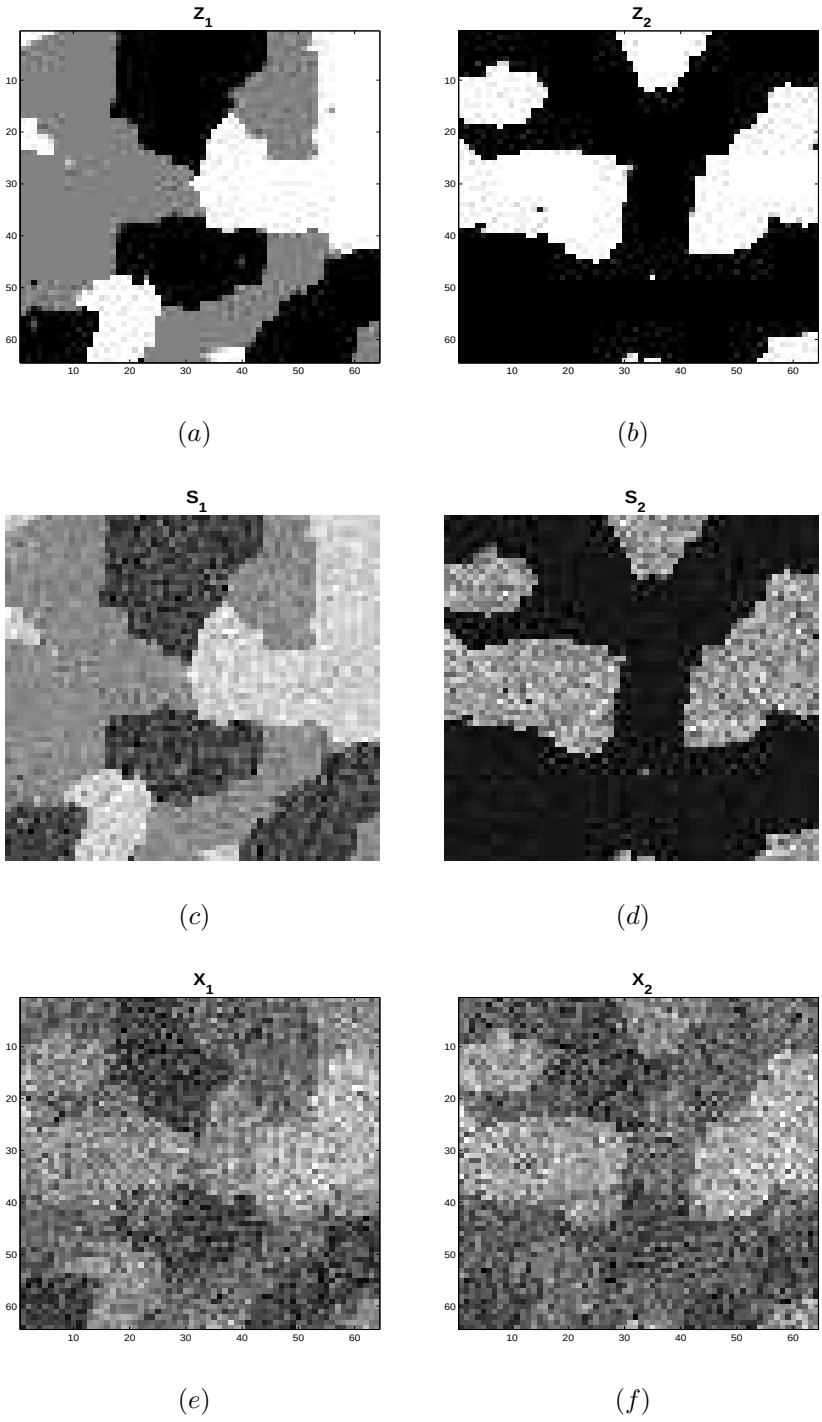
$$p(\mu_k, v_k^{-1} \,|\, \boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Z}) = \mathcal{N}(\mu_k\,;\, \mu_p, v_p)\,\mathcal{G}(v_k^{-1}\,;\, \eta_p, \beta_p)$$

dont les paramètres sont mis à jour, à chaque itération, selon les équations suivantes :

$$\begin{cases} \mu_p = \frac{N_k\bar{s} + \alpha w_i^0\mu_0}{N_k + \alpha w_i^0} \\ v_p = \frac{v_k}{N_k + \alpha w_i^0} \\ \eta_p = \frac{N_k + \alpha w_i^0}{2} \\ \beta_p = \frac{\alpha w_i^0 v_0}{2} + \frac{s^2}{2} + \frac{1}{2}\frac{N_k\alpha w_i^0}{N_k + \alpha w_i^0}(\bar{s} - \mu_0)^2 \\ \bar{s} = \frac{\sum_{r\in\mathcal{S}_k} s(r)}{N_k} \\ s^2 = \sum_{r\in\mathcal{S}_k} s(r)^2 - N_k\bar{s}^2 \end{cases}$$

où $\mathcal{S}_k$ est la région de l'image j appartenant à la classe k :

$$\begin{cases} \mathcal{S}_k = \{r \in \mathcal{S} \,|\, Z(r) = k\} \\ N_k = |\mathcal{S}_k| \end{cases}$$



(a) (b)

(c) (d)

(e) (f)

FIG. IV.5: (a) Classification $\boldsymbol{Z}_1$ de la source 1, (b) Classification $\boldsymbol{Z}_2$ de la source 2, (c) Source originale $\boldsymbol{S}^1$, (d) Source originale $\boldsymbol{S}^2$, (e) Image observée $\boldsymbol{X}^1$, (f) Image observée $\boldsymbol{X}^2$

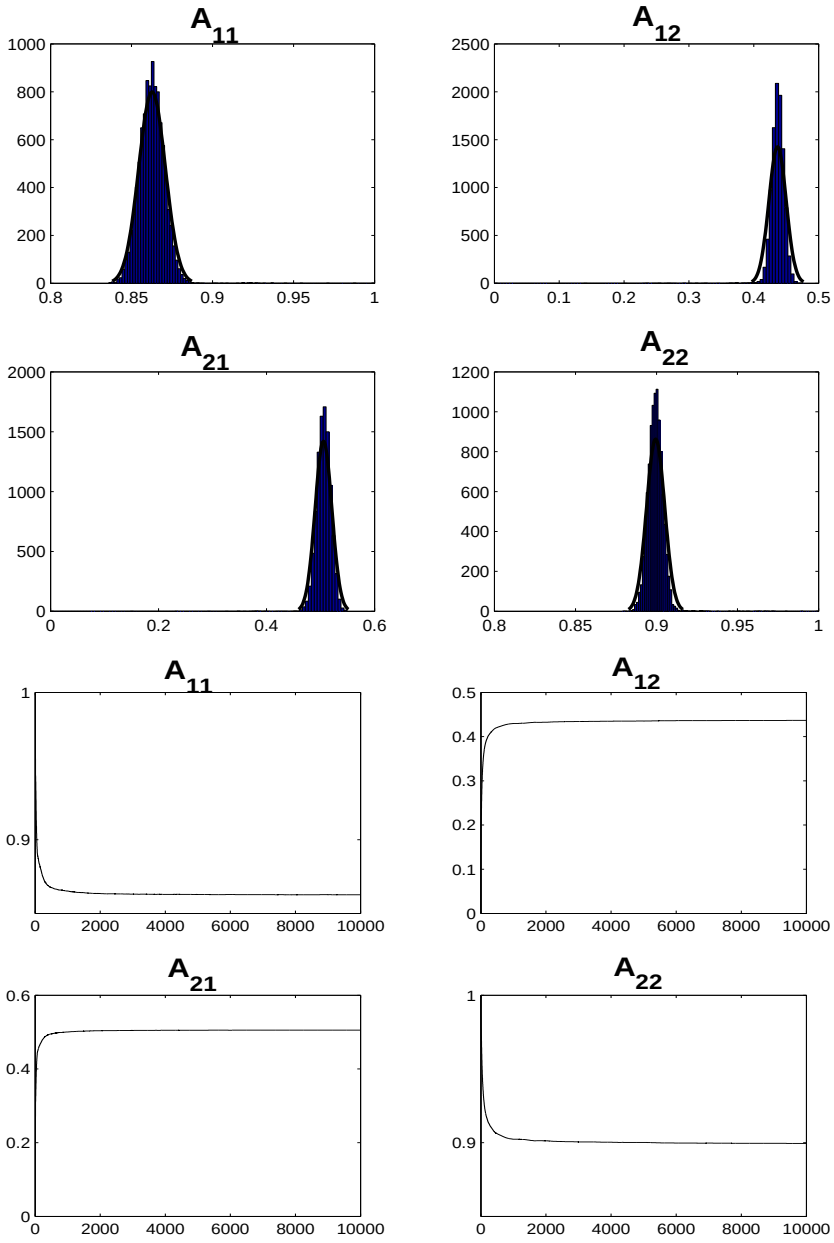


FIG. IV.6: Histogrammes et sommes empiriques des cœfficients de mélange a_{ij}. On note la convergence après 2000 itérations.

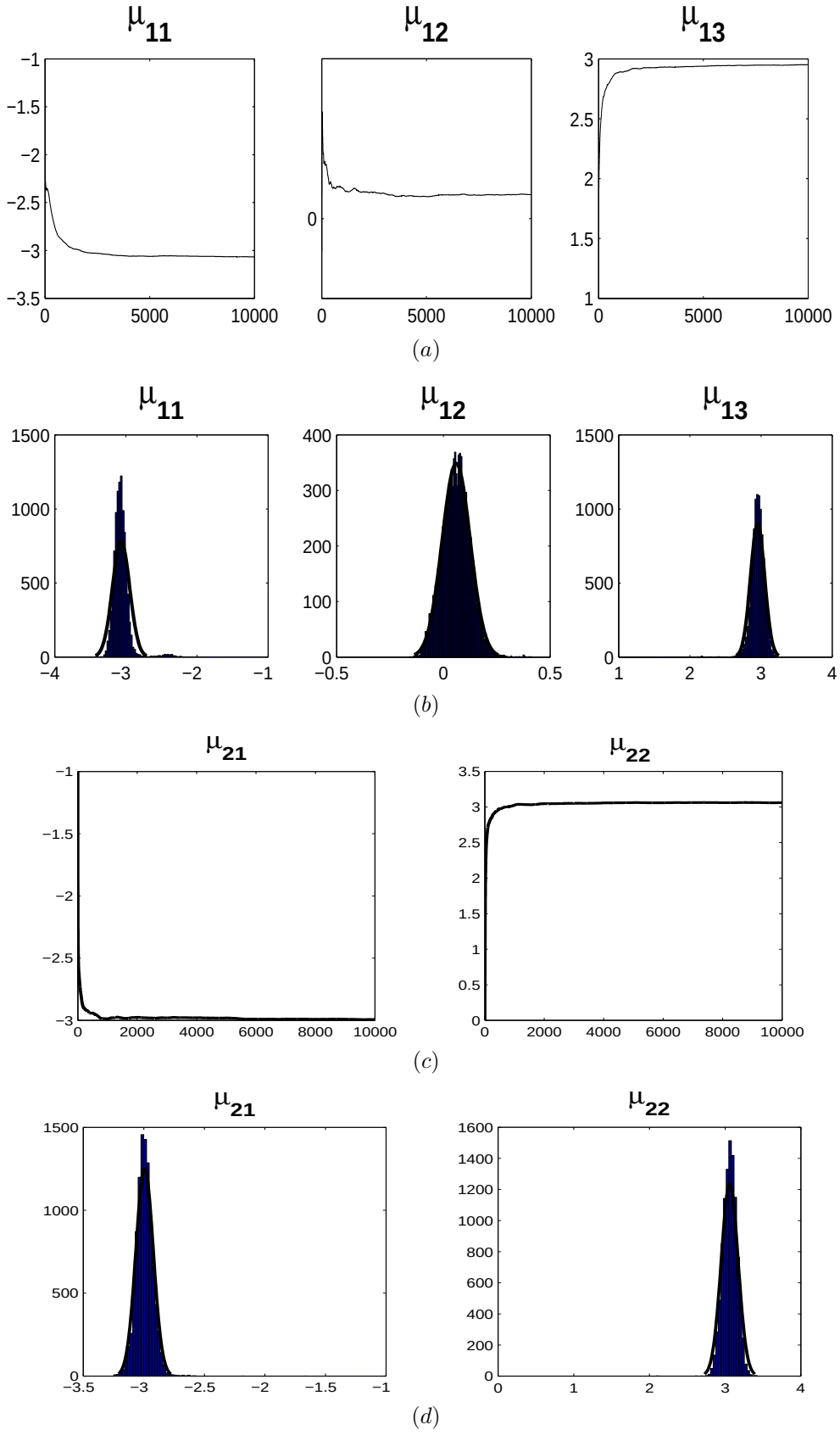


FIG. IV.7: (a)- Convergence des sommes empiriques des moyennes m_{ij} de la source 1 (b)- Histogrammes des moyennes de la source 1 (c)- Convergence des sommes empiriques des moyennes m_{ij} de la source 2 (d)-Histogrammes des moyennes de la source 2

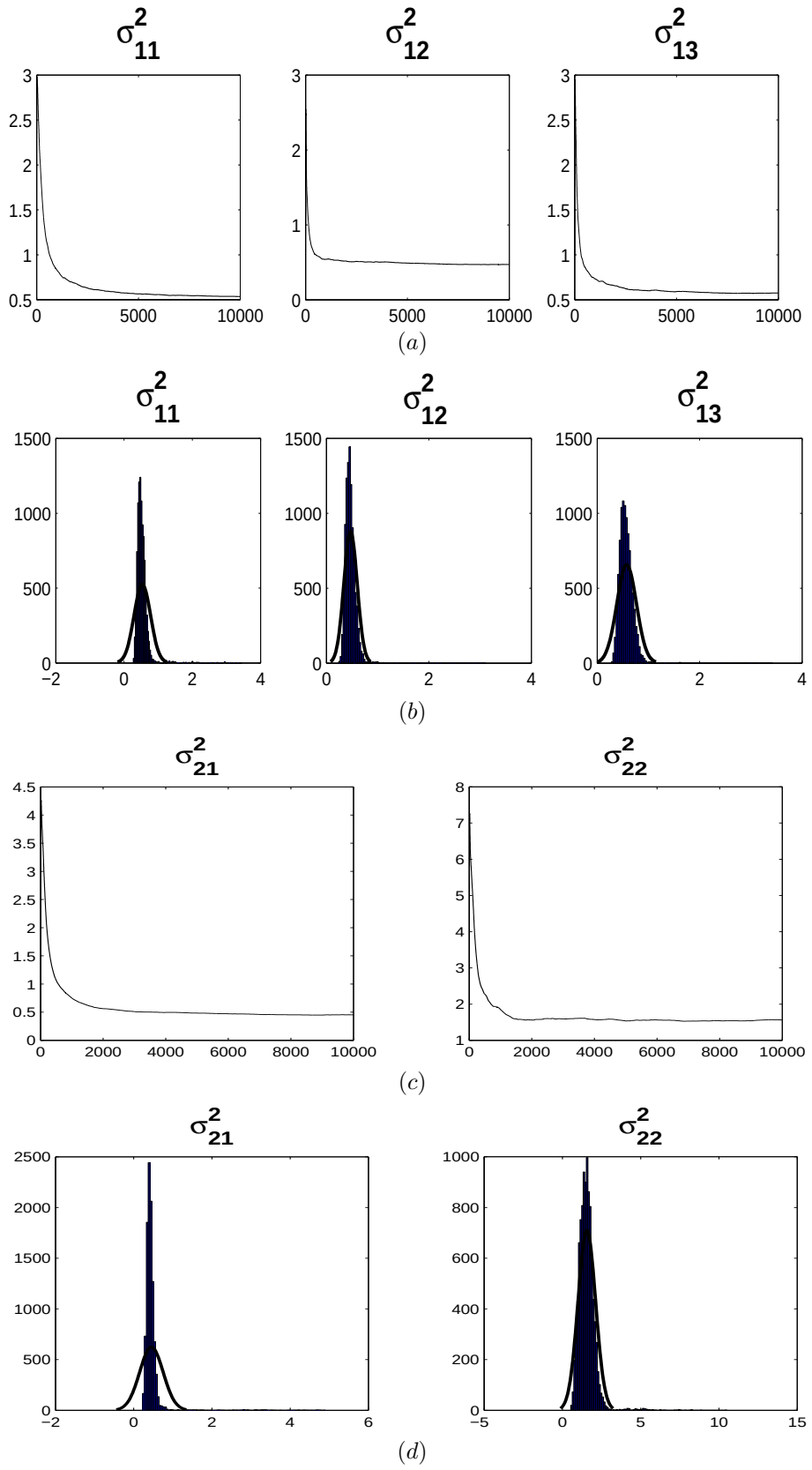
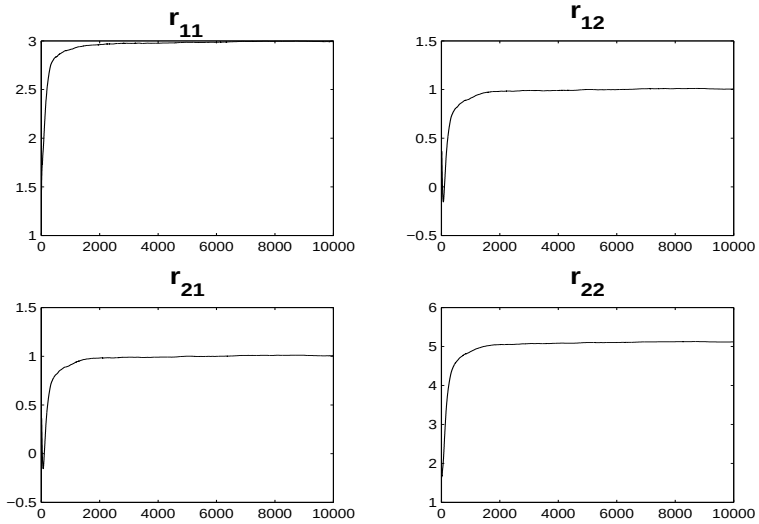
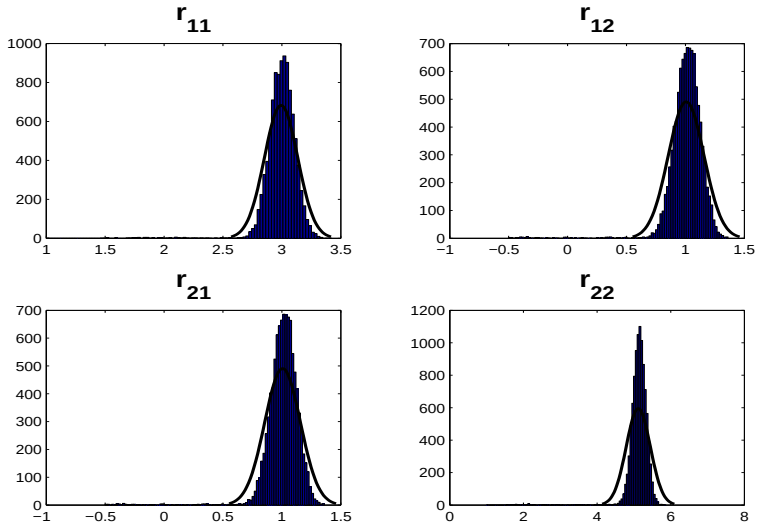


FIG. IV.8: (a)- Convergence des sommes empiriques des variances σ_{ij} de la source 1 (b)- Histogrammes des variances de la source 1 (c)- Convergence des sommes empiriques des variances σ_{ij} de la source 2 (d)- Histogrammes des variances de la source 2



(*a*)



(*b*)

FIG. IV.9: (*a*)- Convergence de la somme empirique de la chaîne des variances du bruit, (*b*) histogrammes des variances du bruit

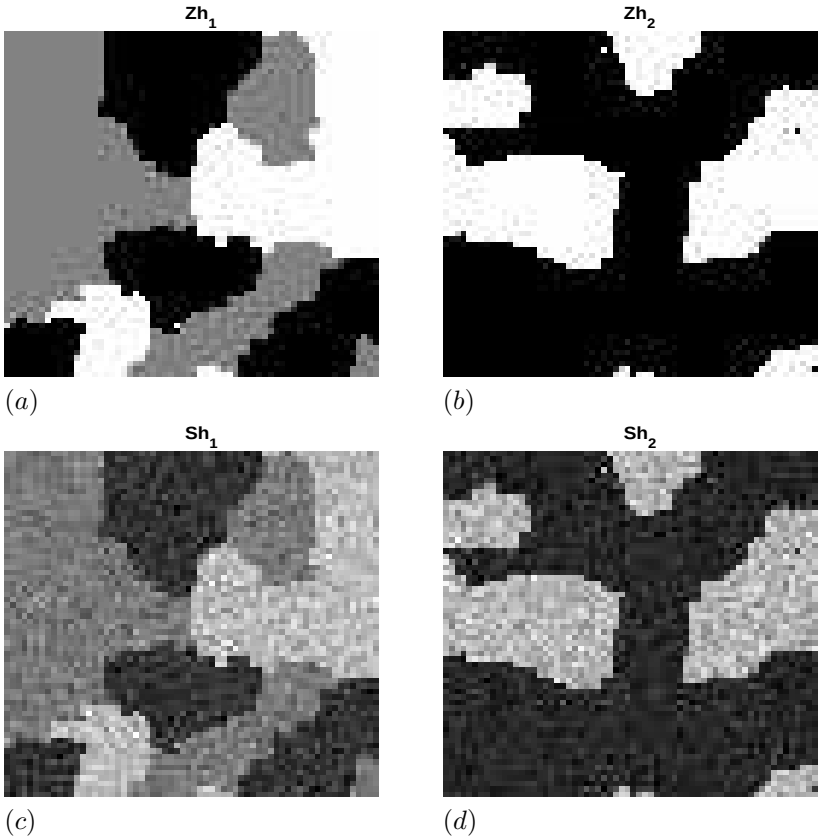


FIG. IV.10: (a)- Estimation de la classification de la source 1, (b)- Estimation de la classification de la source 2, (c)- Reconstruction de la source 1, (d)- Reconstruction de la source 2.

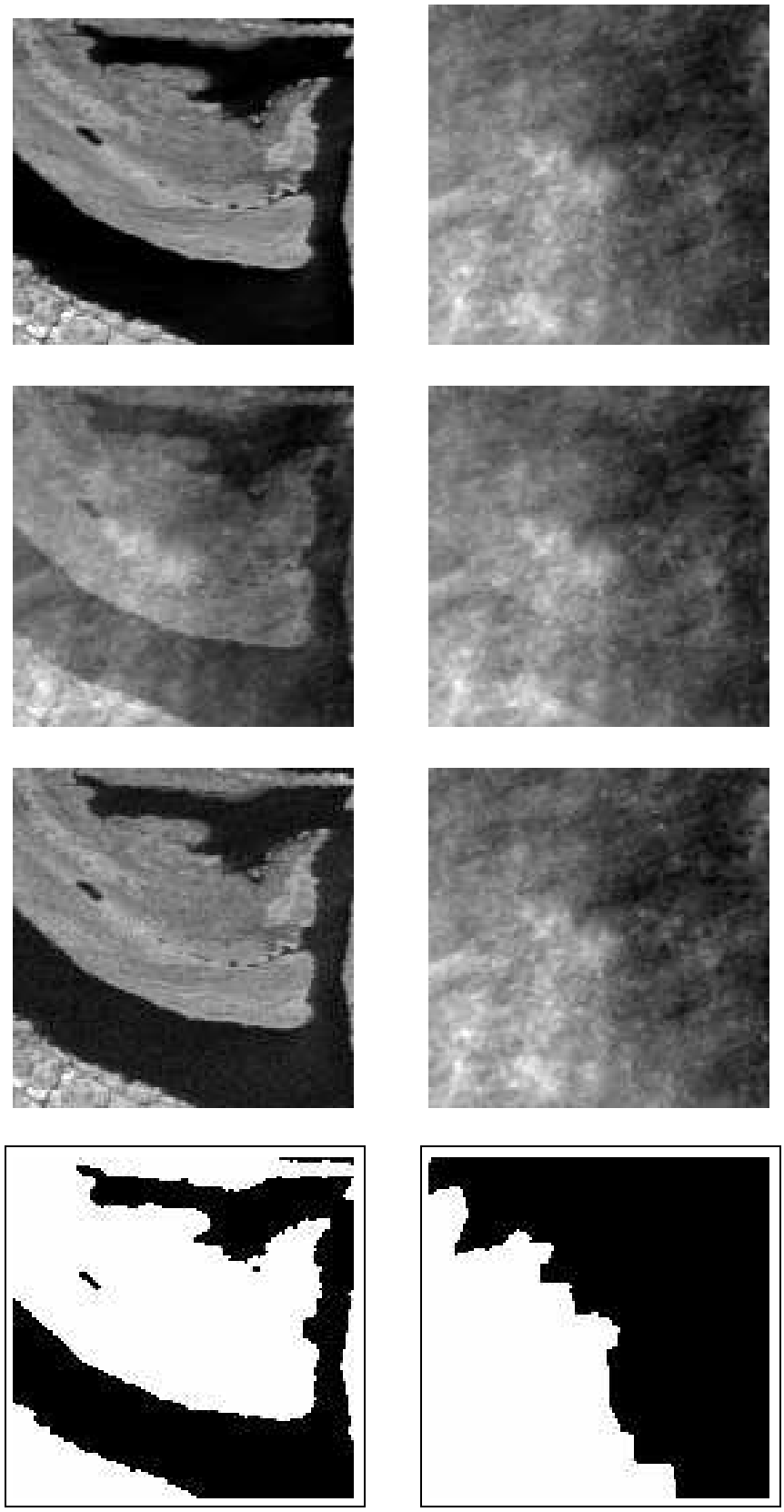
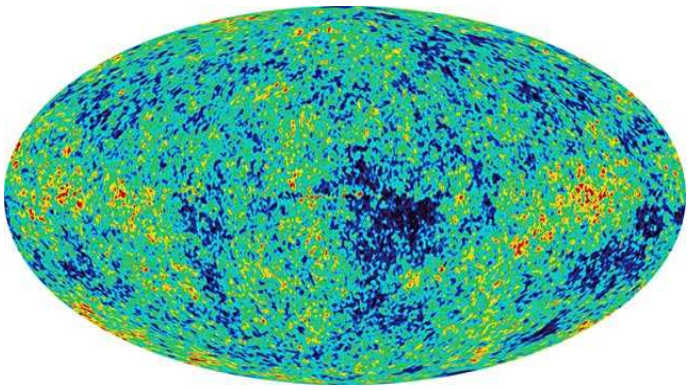


FIG. IV.11: Du haut vers le bas : sources originales, sources mélangées, sources estimées et sources segmentées.

Chapitre V

NON STATIONNARITÉ SPECTRALE : APPLICATION EN COSMOLOGIE OBSERVATIONNELLE



Dans ce chapitre, on illustre l'exploitation de la non stationnarité dans le domaine spectral pour la séparation de composantes astrophysiques dans un mélange bruité. Nous présentons la méthode de séparation basée sur l'algorithme EM pénalisé en modélisant les sources par des processus gaussiens stationnaires et en profitant de l'approximation circulante des matrices de covariance. Le problème de la séparation des composantes astrophysiques et de l'extraction du spectre des fluctuations de la température du rayonnement micro-ondes ***CMB*** (cosmic microwave background) confirme la consistance de la méthodologie bayésienne et le fait qu'elle offre un cadre naturel à l'exploitation de toutes les informations *a priori* émanant des connaissances du physicien.

V.1 Introduction

Dans ce chapitre, nous allons présenter l'approche bayésienne exploitant la non stationnarité dans le domaine spectral et dans le domaine des harmoniques sphériques pour séparer des sources dans un mélange convolutif bruité. Nous allons décrire la méthode de séparation en se basant sur une application en astrophysique. L'objectif de cette application est l'estimation du spectre des fluctuations de température du fond cosmique. En effet, l'observation de ces fluctuations subit des déformations causées par la superposition d'autres émissions astrophysiques d'origines diverses. Les proportions de ce mélange ne sont pas parfaitement connues. Par conséquent, l'emploi des techniques de séparation de sources devient incontournable si on veut mesurer le rayonnement du fond cosmique avec une bonne précision.

Nous avons opté pour une présentation conjointe de la méthode de séparation et du problème physique traîté pour les raisons suivantes.

1. La superposition linéaire bruitée des émissions astrophysiques est un phénomène physique assurant ainsi la validité du modèle de mélange linéaire en séparation de sources.
2. Le développement de la méthode de séparation (bien que générale) a été guidé par cette application physique.
3. La cartographie du rayonnement du fond cosmique est un champ de recherche à part entière. Elle n'est pas une simple application d'une méthode de séparation de sources et mérite donc le développement de méthodes de traitement du signal et de l'image qui lui sont dédiées.

On va commencer par décrire le contexte et les enjeux de la séparation des émissions astrophysiques et placer notre contribution par rapport à d'autres travaux relatifs à ce sujet.

Rayonnement du fond cosmique : de l'observation à l'interprétation

La cartographie des anisotropies du fond diffus cosmique (CMB) et la mesure exacte de son spectre de puissance constituent un des objectifs de la cosmologie observationnelle moderne. L'annonce de ces objectifs scientifiques est rendue possible récemment grâce, en grande partie, au développement des instruments de mesure et à l'élaboration de missions dédiées à la mesure de ces émissions. La mesure des fluctuations de la température primordiale et/ou de la polarisation du rayonnement micro-ondes du fond cosmique (CMB) suscite un intérêt qui ne cesse de croître. En effet, ce rayonnement, émis il y a environ 12 à 15 milliards d'années, contient des informations précieuses sur la physique de notre univers. L'importance de la mesure des anisotropies du rayonnement du fond cosmique (CMB) pour contraindre les modèles cosmologiques est actuellement bien établie. Dans les dix dernières années, beaucoup d'études théoriques ont montré que la mesure des propriétés de ces anisotropies est un outil efficace pour déterminer les paramètres cosmologiques décrivant le contenu de la matière, la géométrie et l'évolution de notre univers [Hu et Sugiyama, 1996; Jungman *et al.*, 1996]. Récemment, des missions comme Boomerang [De Bernardis et *al*, 2000], MAXIMA [Hanany et *al*, 2000], Archéops [Benoît et *al*, 2003], ACBAR [Kuo et *al*, 2002], CBI [Pearson et *al*, 2003] et VSA [Grainge et *al*, 2003] ont fourni des mesures des anisotropies du CMB sur de petites portions du ciel avec une grande résolution angulaire en mettant une forte contrainte sur la quasi-platitude de l'univers.

Des missions satellitaires ont été mises en place pour fournir des mesures des émissions micro-ondes et infra-rouges du ciel à plusieurs fréquences. L'objectif principal de ces missions est la cartographie des fluctuations du CMB sur l'ensemble du ciel avec une grande résolution angulaire et un bon rapport signal à bruit jamais atteint auparavant. Wilkinson Microwave Anisotropy Probe (WMAP) [Bennet et *al*, 2003], l'une de ces missions lancée par la NASA en juin 2001, a déjà fourni des cartes de l'ensemble du ciel avec une bonne résolution de l'ordre de 15 à 30 minutes d'arc avec un grand rapport signal à bruit à chaque pixel. La mission Planck, qui va être lancée par l'agence spatiale européenne ESA en 2007, va fournir des cartes de l'ensemble du ciel avec une résolution angulaire de l'ordre de 5 à 30 minutes d'arc sur 9 fréquences comprises entre 30 et 850 GHz.

L'accomplissement des objectifs théoriques consistant essentiellement à l'étude des modèles cosmologiques et l'évaluation des paramètres cosmologiques exige un niveau assez élevé de précision dans la mesure du CMB

et une erreur de reconstruction beaucoup plus inférieure au niveau des contaminations des autres émissions astrophysiques. En effet, on s'attend à la contribution d'au moins six émissions d'origine diverses dans les observations mesurées par Planck. Par conséquent, le succès des futures missions est étroitement lié à la performance de la séparation du CMB des autres émissions astrophysiques. Les techniques de séparation de sources se trouvent ainsi au cœur du traitement et d'analyse des futures données CMB.

Placement du travail

Deux types d'algorithmes ont été proposés pour séparer le CMB des autres émissions astrophysiques.

1. Filtrage de Wiener et MEM (Maximum Entropy Method) [Bouchet et Gispert, 1999; Tegmark et Estathiou, 1996; Hobson *et al.*, 1998] : on suppose dans ces méthodes que les spectres électromagnétiques des sources (les colonnes de la matrice de mélange) sont connues.
2. ICA (analyse en composantes indépendantes) [Baccigalupi *et al.*, 2000] : aucun *a priori* n'est supposé sur les spectres électromagnétiques.

La première classe d'algorithmes donne des résultats de reconstruction satisfaisantes mais elle est sévèrement limitée par les incertitudes sur les spectres électromagnétiques. En effet, en pratique les spectres électromagnétiques de certaines composantes ne sont pas connues avec une précision satisfaisante. La deuxième classe (ICA) donne de bons résultats dans des cas simples où le mélange n'est ni bruité ni convolutif.

Nous proposons une solution bayésienne pour séparer les sources en se basant sur la diversité spectrale des sources. La méthode proposée exploite la non stationnarité des cœfficients de Fourier pour le traitement des petites portions du ciel (données sous forme de cartes) ou la non stationnarité des cœfficients de la base d'harmoniques sphériques lorsqu'on traîte les données sur l'ensemble du ciel. On estime conjointement les spectres électromagnétiques (la matrice de mélange), les spectres de puissance spatiaux des sources et le niveau du bruit sur chaque capteur. La maximisation de la distribution *a posteriori* de tous ces paramètres est implémentée avec l'algorithme EM en profitant de la structure à variables cachées du problème (les sources sont les variables manquantes). Le critère à maximiser peut être ré-interprété comme un ajustement de matrices de covariance (statistiques d'ordre deux) dans la métrique de Kullback-Leibler et le découpage du domaine spectral en anneaux à spectres constants accélère la mise en œuvre de la méthode proposée. L'approche bayésienne offre un cadre naturel pour incorporer des informations *a priori* sur les spectres électromagnétiques et les spectres spatiaux.

V.2 Modélisation des observations du CMB

Nous classifions les principales composantes astrophysiques, dans le domaine millimétrique, en trois classes. Le CMB, d'origine cosmologique, a été émis avant la formation des objets astrophysiques comme les amas et les galaxies lorsque l'univers a quitté sa forme entièrement ionisée. Les émissions extra-galactiques, plus jeunes que le CMB, sont les émissions provenant de l'extérieur de notre galaxie. Finalement, les composantes galactiques proviennent de notre galaxie et sont très orientées vers le plan galactique. La liste suivante est un exemple d'émissions dans le domaine millimétrique.

1. **Les anisotropies du CMB**.
2. **Les émissions extra-galactiques**
   - sources ponctuelles (radio-galaxies, galaxies infra-rouge, quasars).
   - émissions Sunyaev-Zeldovich (SZ) des amas de galaxies.
3. Les émissions galactiques
   - Poussière : émission thermique des graines de poussière froide intragalactique.
   - Synchrotron : radiation émise par l'intéraction des électrons ultra-relativistes avec les champs magnétiques de la Galaxie.
   - Free-Free (Bremsstrahlung) : rayonnement de freinage des électrons galactiques.

Ces composantes ont des lois d'émission spectrale (en fonction de la fréquence d'observation ν) différentes. La séparation de ces émissions est alors possible à partir des observations prises à différentes longueurs d'onde (différentes fréquences) en se basant sur la diversité des spectres électromagnétiques. Les spectres électromagnétiques du CMB et de l'effet SZ sont connus avec une bonne précision et peuvent être inclus dans les méthodes de séparation [Hobson *et al.*, 1998]. Cependant, pour le reste des sources, on ne dispose, dans les meilleurs des cas, que des spectres extrapolés des fréquences loins de la bande d'observation [De Zotti *et al.*, 1999].

Avant de présenter la technique de séparation, nous allons présenter le modèle décrivant les émissions observées sur le ciel $x_\nu(\boldsymbol{r})$ à la position $\boldsymbol{r}$ et à la fréquence ν. Pour les longueurs d'onde de l'ordre du millimètre et du centimètre, $x_\nu(\boldsymbol{r})$ peut être considéré comme une superposition linéaire du CMB ($\hat{s}_{CMB}(\nu,\boldsymbol{r})$) et des autres émissions galactiques et extra-galactiques ($\hat{s}_f(\nu,\boldsymbol{r})$). Cette somme est convoluée avec un noyau d'observation $b_\nu(\boldsymbol{r})$ qui ne dépend que du détecteur. Un bruit additif $\epsilon_\nu(\boldsymbol{r})$ est présent sur chaque détecteur. Le signal $x_\nu(\boldsymbol{r})$ observé se met alors sous la forme :

$$x_\nu(\boldsymbol{r}) = \hat{s}_{CMB}(\nu,\boldsymbol{r}) * b_\nu(\boldsymbol{r}) + \sum_{f=1}^{N_f} \hat{s}_f(\nu,\boldsymbol{r}) * b_\nu(\boldsymbol{r}) + \epsilon_\nu(\boldsymbol{r}) \tag{V.1}$$

où N_f représente le nombre des émissions galactiques et extra-galactiques considérées, $*$ définit l'opérateur de convolution et $\epsilon_\nu(\boldsymbol{r})$ est le bruit instrumental du détecteur à la fréquence ν.

Concernant le rayonnement du CMB, les réponses spatiale et électromagnétique sont séparables,

$$\hat{s}_{CMB}(\nu,\boldsymbol{r}) = g_{CMB}(\nu) \times s_{CMB}(\boldsymbol{r})$$

où $g_{CMB}(\nu)$ représente le spectre électromagnétique du CMB ne dépendant pas de la position $\vec{r}$ sur le ciel et $s_{CMB}(\boldsymbol{r})$ représente sa distribution spatiale. Pour le reste des émissions, on peut aussi, dans une première approximation, supposer la forme factorisée des réponses fréquentielles et spatiales et l'indépendance du spectre électromagnétique de la position spatiale [Bouchet et Gispert, 1999; Hobson *et al.*, 1998]. L'équation (V.1) s'écrit alors,

$$x_\nu(\boldsymbol{r}) = g_{CMB}(\nu)\ s_{CMB}(\boldsymbol{r}) * b_\nu(\boldsymbol{r}) + \sum_{f=1}^{N_f} g_f(\nu)\ s_f(\boldsymbol{r}) * b_\nu(\boldsymbol{r}) + \epsilon_\nu(\boldsymbol{r})$$

où g_f représente la moyenne du spectre électromagnétique de la composante f. La figure (V.1) montre des distributions spatiales typiques pour le CMB, la poussière galactique (dust) et l'effet SZ. Ces simulations représentent des petites portions du ciel sur des cartes de 300×300 pixels de dimension 2.5 minutes d'arc [Delabrouille *et al.*, 2001]. La figure (V.2) montre les spectres électromagnétiques de ces sources. A titre illustratif, nous avons simulé le mélange sur 6 fréquences entre 100 et 850 GHz correspondant aux 6 détecteurs de l'instrument *HFI* du satellite Planck (figure (V.3)).

De point de vue modélisation, la composante du CMB subit les mêmes transformations que les autres émissions galactiques et extragalactiques et donc on peut écrire le mélange sous la forme suivante,

$$x_\nu(\boldsymbol{r}) = \sum_{i=1}^{N_c} g_i(\nu)\ s_i(\boldsymbol{r}) * b_\nu(\boldsymbol{r}) + \epsilon_\nu(\boldsymbol{r})$$

où $N_c = N_f + 1$ est le nombre total des composantes du mélange. Donc, pour un nombre fini de détecteurs $d = 1, .., N_d$ opérant aux fréquences électromagnétiques ν_d,

$$x_d(\boldsymbol{r}) = \sum_{i=1}^{N_c} A_{di}\ s_i(\boldsymbol{r}) * b_d(\boldsymbol{r}) + \epsilon_d(\boldsymbol{r}) \tag{V.2}$$

où $A_{di} = g_i(\nu_d)$ est une matrice $N_d \times N_c$ qu'on appelle la matrice de mélange.

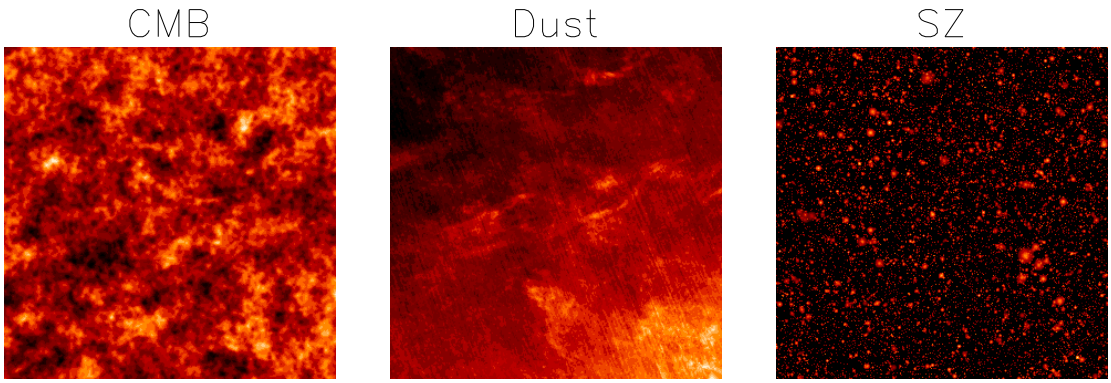


FIG. V.1: Distributions spatiales typiques du CMB, dust et SZ utilisées dans les simulations de ce chapitre. La distribution du SZ est présentée en échelle logarithmique.

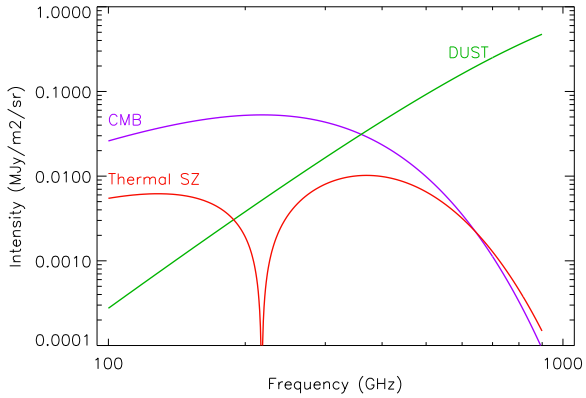


FIG. V.2: Les spectres électromagnétiques relatifs au CMB, à la poussière galactique (dust) et à l'effet SZ. Ces spectres définissent les coefficients du mélange de ces composantes (la matrice de mélange $\boldsymbol{A}$) quand on néglige l'effet de la convolution.

V.3 Méthodologie bayésienne

Le problème d'inférence initial $\mathcal{I} := \boldsymbol{X} \longrightarrow \boldsymbol{A}$ consiste à identifier la matrice de mélange $\boldsymbol{A}$ (les spectres électromagnétiques des sources) à partir des observations $\boldsymbol{X} = \{x_d(\vec{r})\}_{\vec{r}\in\mathcal{D}}^{d=1..N_d}$. N_d est le nombre de capteurs (nombre de fréquences d'observations) et $\mathcal{D}$ est le domaine d'observation (une portion du ciel ou tout le ciel). La méthodologie bayésienne, exposée en détail dans le chapitre (II), consiste à :

1. former la distribution *a posteriori* comme étant la degré d'incertitude de la proposition $\mathcal{I}$,
2. choisir les probabilités intervenant dans l'expression de la distribution *a posteriori*,
3. choisir une fonction coût,
4. implémenter l'algorithme d'optimisation adapté à la structure du problème.

La distribution *a posteriori* de $\boldsymbol{A}$ sachant les observations $\boldsymbol{X}$ s'écrit, selon la règle de Bayes,

$$\mathrm{p}(\boldsymbol{A} \mid \boldsymbol{X}, \boldsymbol{I}) \propto \mathrm{p}(\boldsymbol{X} \mid \boldsymbol{A}, \boldsymbol{I})\, \mathrm{p}(\boldsymbol{A} \mid \boldsymbol{I}) \tag{V.3}$$

où $\boldsymbol{I}$ représente toute l'information *a priori* qu'on possède sur le problème comme le modèle de mélange linéaire, les connaissances *a priori* de certains spectres électromagnétiques...

La vraisemblance $\mathrm{p}(\boldsymbol{X} \mid \boldsymbol{A}, \boldsymbol{I})$ est la modélisation probabiliste du problème direct expliquant l'obtention des observations $\boldsymbol{X}$ à partir de la matrice $\boldsymbol{A}$. En supposant l'invariance par rotation du noyau de convolution,

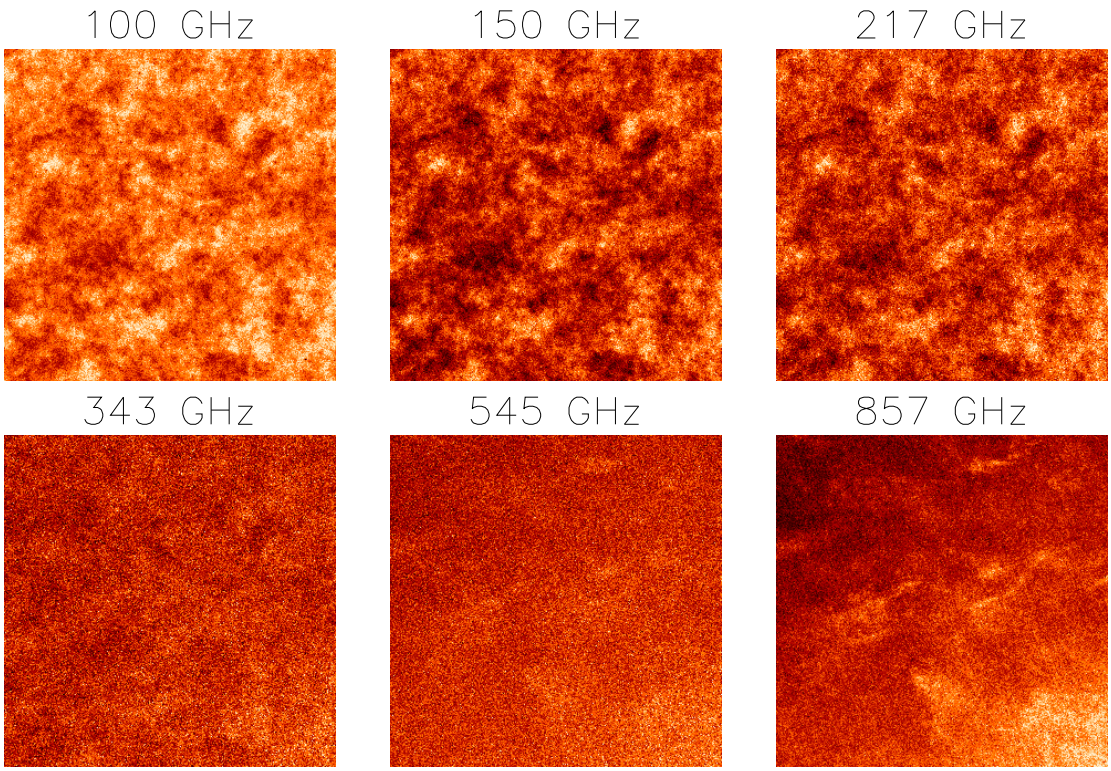


FIG. V.3: Simulation des observations au niveau des six détecteurs de l'instrument HFI de Planck.

le signal sur le détecteur d s'écrit,

$$x_d(\vec{r}) = \sum_{c=1}^{N_c} A_{dc}. \int b_d(|\vec{r} - \vec{r}'|).s_c(\vec{r}')d\,\vec{r}' + \epsilon_d(\vec{r}). \tag{V.4}$$

$\vec{r}$ est la direction d'observation dans le ciel, s_c est l'émission de la source c, ϵ_d est le bruit sur le capteur d et $b_d(|\vec{r} - \vec{r}'|)$ repésente le noyau de convolution et ne dépend que de $|\vec{r} - \vec{r}'|$. Chaque élément A_{dc} de la matrice de mélange est le résultat de l'intégration de la loi d'émission de la source c sur la bande de fréquence du détecteur d.

Le modèle direct fait apparaître naturellement une structure à variables cachées. Les sources représentent les variables manquantes et les observations représentent les données incomplètes. La vraisemblance prend ainsi une forme intégrale,

$$\mathrm{p}(\{\boldsymbol{x}(\vec{r})\}_{\vec{r}\in\mathcal{D}} \mid \boldsymbol{A}, \boldsymbol{I}) \quad = \quad \int \mathrm{p}(\{\boldsymbol{x}(\vec{r})\}_{\vec{r}\in\mathcal{D}} \mid \{\boldsymbol{s}(\vec{r})\}_{\vec{r}\in\mathcal{D}}\boldsymbol{A}, \boldsymbol{I})\, \mathrm{p}(\{\boldsymbol{s}(\vec{r})\}_{\vec{r}\in\mathcal{D}} \mid \boldsymbol{I}) d\,\{\boldsymbol{s}(\vec{r})\}_{\vec{r}\in\mathcal{D}}$$

La corrélation spatiale des sources et la présence de la convolution rendent la manipulation de cette expression, même en faisant appel aux techniques de restauration-maximisation, très délicate. L'un des points clé de notre contribution est le **changement de base**. Nous avons considéré deux bases différentes selon le domaine d'observation $\mathcal{D}$:

1. Si $\mathcal{D}$ est une petite portion du ciel et si elle est quasi plate alors on effectue une transformée de Fourier des observations et on implémente la méthode de séparation dans le domaine spectral.
2. Si $\mathcal{D}$ représente l'ensemble du ciel alors le domaine des harmoniques sphériques représente une base naturelle pour l'implémentation de la méthode de séparation.

V.3.1 Domaine spectral

En passant dans le domaine de Fourier, la convolution dans l'expression (V.4) devient une simple multiplication. A chaque fréquence $\boldsymbol{k}$ du domaine spectral, le mélange s'écrit,

$$x_d(\boldsymbol{k}) = b_d(\boldsymbol{k}) \sum_{c=1}^{N_c} A_{dc} s_c(\boldsymbol{k}) \; + \; \epsilon_d(\boldsymbol{k}).$$

On peut écrire cette expression sous une forme matricielle plus compacte,

$$\boldsymbol{x}(\boldsymbol{k}) = \boldsymbol{B}(\boldsymbol{k})\boldsymbol{A}\boldsymbol{s}(\boldsymbol{k}) + \boldsymbol{\epsilon}(\boldsymbol{k})$$

où $\boldsymbol{x}(\boldsymbol{k})$ est le vecteur $N_d \times 1$ des observations, $\boldsymbol{s}(\boldsymbol{k})$ est le vecteur $N_c \times 1$ des sources, $\boldsymbol{A}$ est la matrice de mélange, $\boldsymbol{B}(\boldsymbol{k})$ est la matrice diagonale contenant les noyaux de convolution : $B_{dd}(\boldsymbol{k}) = b_d(\boldsymbol{k})$ et $\boldsymbol{\epsilon}(\boldsymbol{k})$ est le vecteur contenant les bruits des capteurs. Dans la suite, on désigne $\boldsymbol{x}_{1..K}$ et $\boldsymbol{s}_{1..K}$ l'ensemble des observations et des sources.

Remarque 12 *Nous allons tenir compte du noyau de convolution quand on va exposer le pseudo code de l'algorithme de séparation mais nous allons l'omettre dans la suite afin d'alléger les expressions.*

[A] Modélisation des sources et du bruit

Les sources sont modélisées par un processus gaussien blanc (partie réelle indépendante de la partie imaginaire) non stationnaire [Hobson *et al.*, 1998; Snoussi *et al.*, 2001]. A chaque fréquence $\boldsymbol{k}$, les sources $\boldsymbol{s}_k$ suivent une distribution gaussienne centrée de covariance diagonale $\boldsymbol{P}_k = \mathrm{E}\left[\boldsymbol{s}_k \, \boldsymbol{s}_k^*\right]$ (la diagonalité est due à l'indépendance des sources),

$$\boldsymbol{s}_k \sim \mathcal{N}(0, \boldsymbol{P}_k)$$

Les éléments diagonaux $[\sigma_c^2(k) = P_{cc}(k), k = 1..K]$ des matrices $\boldsymbol{P}_k$ sont les spectres de puissance spatiaux des sources. Nous supposons que les sources sont isotropes et donc que les covariances spectrales sont circulaires. Autrement dit, les spectres $\boldsymbol{P}(\boldsymbol{k})$ ne varient qu'en fonction de la norme $\|\boldsymbol{k}\|$ de la fréquence spatiale.

Le bruit est supposé centré blanc gaussien de spectre constant $\boldsymbol{R}_\epsilon = \mathrm{E}\left[\boldsymbol{\epsilon}_k \boldsymbol{\epsilon}_k^*\right]$. On suppose que la matrice $\boldsymbol{R}_\epsilon$ est diagonale[1] et les éléments diagonaux peuvent avoir des valeurs différentes afin de tenir compte des niveaux de bruit différents sur les capteurs (c'est le cas pour la mission Planck).

Remarque 13 *En pratique, on ne connait pas les spectres des sources et du bruit. Par conséquent, on change le problème d'inférence pour inclure ces spectres dans l'ensemble des paramètres $\boldsymbol{\theta} = \{\boldsymbol{A}, \boldsymbol{P}_k, \boldsymbol{R}_\epsilon\}$ à estimer :*

$$\mathcal{I} := (\boldsymbol{x}_{1..K} \longrightarrow \boldsymbol{A}) \; \rightsquigarrow \mathcal{I} := (\boldsymbol{x}_{1..K} \longrightarrow \boldsymbol{A}, \boldsymbol{P}_k, \boldsymbol{R}_\epsilon)$$

La gaussianité et l'indépendance spectrale des sources et du bruit conduisent à une expression explicite de la vraisemblance :

$$\begin{aligned} \mathrm{p}(\boldsymbol{x}_{1..K} \mid \boldsymbol{\theta}) &= \int \mathrm{p}(\boldsymbol{x}_{1..K} \mid \boldsymbol{s}_{1..K}, \boldsymbol{A}, \boldsymbol{R}_\epsilon)\mathrm{p}(\boldsymbol{s}_{1..K} \mid \boldsymbol{P}_{1..K}) d\,\boldsymbol{s}_{1..K} \\ &= \prod_{\boldsymbol{k}} \int \mathrm{p}(\boldsymbol{x}_k \mid \boldsymbol{s}_k, \boldsymbol{A}, \boldsymbol{R}_\epsilon)\mathrm{p}(\boldsymbol{s}_k \mid \boldsymbol{P}_k) d\,\boldsymbol{s}_k \qquad \text{(V.5)} \\ &= \prod_{\boldsymbol{k}} |2\,\pi\,\boldsymbol{R}_k|^{-1} \exp\left[-\mathrm{Tr}\left(\boldsymbol{R}_k^{-1}\,\boldsymbol{x}_k \boldsymbol{x}_k^*\right)\right] \; ; \; \boldsymbol{R}_k = \boldsymbol{A}\boldsymbol{P}_k\boldsymbol{A}^* + \boldsymbol{R}_\epsilon. \end{aligned}$$

où $\boldsymbol{R}_k$ est la matrice de covariance spectrale des observations $\boldsymbol{x}_k$ à la fréquence $\boldsymbol{k}$.

[1] L'hypothèse de la diagonalité n'est pas nécessaire dans la méthode proposée.

[B] Interprétation du critère

La circularité des spectres de puissances $\boldsymbol{P}_k$ implique la circularité des covariances spectrales $\boldsymbol{R}_k$,

$$\boldsymbol{R}_k = \boldsymbol{R}_l = \boldsymbol{A}\boldsymbol{P}_l\boldsymbol{A}^* + \boldsymbol{R}_\epsilon, \ \forall \boldsymbol{k} \text{ telle que } \|\boldsymbol{k}\| = l.$$

Par conséquent, dans l'expression (V.5), on peut re-partitionner le produit sur tous les modes $\boldsymbol{k}$ en un produit sur les cercles concentriques $\mathcal{D}_l$ (voir figure (V.4)),

$$\mathrm{p}(\boldsymbol{x}_{1..K} \mid \boldsymbol{\theta}) = \prod_{l=\|\boldsymbol{k}\|} |2\,\pi\,\boldsymbol{R}_l|^{-w_l} \exp\left[-\mathrm{Tr}\left(\boldsymbol{R}_l^{-1} \sum_{\boldsymbol{k}\in\mathcal{D}_l} \boldsymbol{x}_k \boldsymbol{x}_k^*\right)\right] \tag{V.6}$$

où $w_l = |\mathcal{D}_l|$ est le nombre de fréquences appartenant au cercle $\mathcal{D}_l$.

En introduisant les covariances empiriques des observations $\hat{\boldsymbol{R}}_l = \sum_{\boldsymbol{k}\in\mathcal{D}_l} \boldsymbol{x}_k \boldsymbol{x}_k^* \ / \ w_l$, le logarithme de la vraisemblance (V.4) s'écrit, à une constante additive près, comme une somme pondérée des divergences de Kullback-Leibler entre les matrices de covariances spectrales théoriques et les covariances empiriques des domaines $\mathcal{D}_l$:

$$\begin{aligned} \log \mathrm{p}(\boldsymbol{x}_{1..K} \mid \boldsymbol{\theta}) &= -\sum_{l=1}^{L} w_l \left(-\ln|\boldsymbol{R}_l^{-1}\hat{\boldsymbol{R}}_l| + \mathrm{Tr}\left(\boldsymbol{R}_l^{-1}\hat{\boldsymbol{R}}_l\right) - N_d\right) + cste \\ &= -\sum_{l=1}^{L} w_l D_{KL}(\boldsymbol{R}_l, \hat{\boldsymbol{R}}_l) + cste. \end{aligned} \tag{V.7}$$

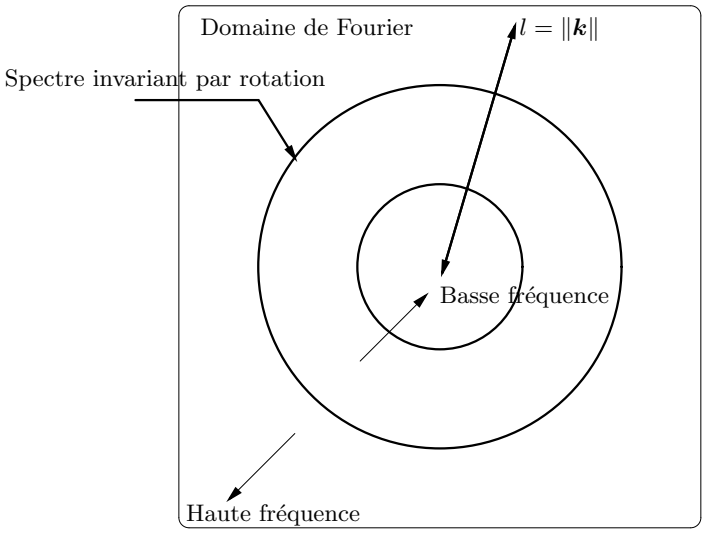


Fig. V.4: Les spectres des sources sont circulaires. Le critère du maximum de vraisemblance est un ajustement des matrices de covariance spectrales sur les cercles concentriques du domaine de Fourier.

Le critère du maximum *a posteriori* peut s'interpréter comme une version régularisée de l'ajustement des matrices de covariances :

$$\log \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..K}) = \underbrace{-\sum_{l=1}^{L} w_l D_{KL}(\boldsymbol{R}_l, \hat{\boldsymbol{R}})}_{\textit{Ajustement de matrices de covariance spectrale}} + \underbrace{\log \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{I})}_{\textit{Régularisation du critère}}$$

La méthode de séparation se base ainsi sur l'exploitation de la non stationnarité spectrale pour identifier la matrice de mélange. Elle est similaire à l'approche adoptée dans [Pham et Cardoso, 2001] où on exploite la non stationnarité dans le domaine fréquentiel 1-D pour la séparation d'un mélange non bruité. Notre méthode peut être considérée comme une extension de [Pham et Cardoso, 2001] au cas 2-D bruité exploitant la structure cachée pour implémenter la solution avec l'algorithme EM.

Remarque 14 *On peut accélérer la méthode de séparation en supposant que les spectres de puissance des sources $\boldsymbol{P}_l$ sont constants par anneaux [Cardoso et al., 2002]. Autrement dit, nous allons élargir les cercles en les transformant en anneaux (voir figure (V.5)). Le critère possède la même forme que l'expression (V.7) en prenant les domaines $\mathcal{D}_l = \{\boldsymbol{k} \mid k_{min}(l) \leq \|\boldsymbol{k}\| \leq k_{max}(l)\}$.*

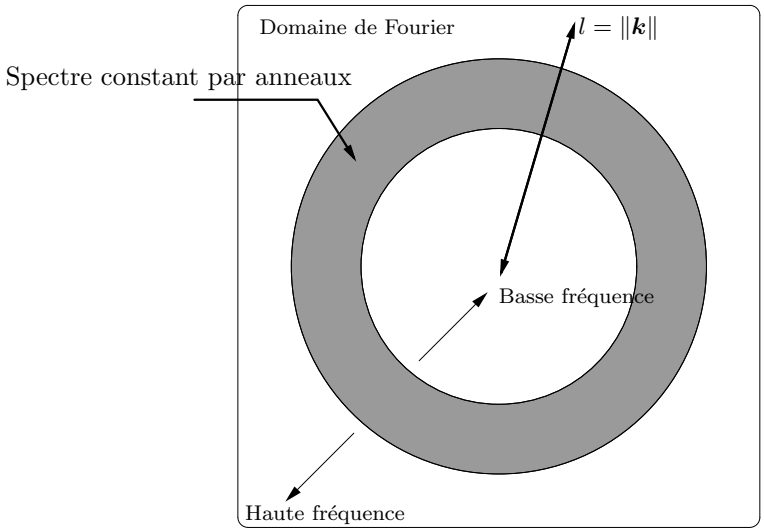


FIG. V.5: Les spectres des sources sont constants par anneaux. Le critère du maximum de vraisemblance est un ajustement des matrices de covariance spectrales sur les anneaux concentriques du domaine de Fourier.

[C] EM SPECTRAL

Malgré sa forme explicite, le critère (V.7) est difficile à optimiser. Cependant, on peut profiter de la structure à variables cachées du problème (les sources étant les variables manquantes) pour implémenter l'algorithme EM [Dempster *et al.*, 1977]. Les détails de la dérivation des équations de ré-estimation de l'algorithme EM pour la séparation des composantes astrophysiques dans le domaine spectral sont dans [Snoussi *et al.*, 2001]. L'étape "*Expectation*" de l'algorithme EM nécessite le calcul des statistiques suivantes :

$$\begin{cases} \boldsymbol{R}_{xx}(l) &= \frac{1}{w_l} \sum_{k \in \mathcal{D}_l} \boldsymbol{x}_k \boldsymbol{x}_k^* \\ \boldsymbol{R}_{xs}(l) &= \frac{1}{w_l} \sum_{k \in \mathcal{D}_l} \boldsymbol{x}_k \mathrm{E}\big[\boldsymbol{s}_k \mid \boldsymbol{x}_k, \boldsymbol{\theta}\big]^* \\ \boldsymbol{R}_{ss}(l) &= \frac{1}{w_l} \sum_{k \in \mathcal{D}_l} \mathrm{E}\big[\boldsymbol{s}_k \boldsymbol{s}_k^* \mid \boldsymbol{x}_k, \boldsymbol{\theta}\big] \end{cases} \tag{V.8}$$

Le calcul des espérances *a posteriori* conditionnelles $\mathrm{E}\,[\boldsymbol{s} \mid \boldsymbol{x}, \boldsymbol{\theta}]$ et $\mathrm{E}\,[\boldsymbol{s}\boldsymbol{s}^* \mid \boldsymbol{x}, \boldsymbol{\theta}]$, dans le cas gaussien, à chaque fréquence $\boldsymbol{k}$, donne :

$$\begin{aligned} \mathrm{E}\big[\boldsymbol{s}_k \mid \boldsymbol{x}, \boldsymbol{\theta}\big] &= \boldsymbol{W}_k \, \boldsymbol{x}_k \\ \mathrm{E}\big[\boldsymbol{s}_k \boldsymbol{s}_k^* \mid \boldsymbol{x}, \boldsymbol{\theta}\big] &= \boldsymbol{W}_k \boldsymbol{x}_k \boldsymbol{x}_k^* \boldsymbol{W}_k^* + \boldsymbol{V}_k \end{aligned}$$

où les matrices $\boldsymbol{W}_k$ (de Wiener) et $\boldsymbol{V}_k$ (covariance *a posteriori*) sont données par,

$$\begin{aligned} \boldsymbol{V}_k &= \left(\boldsymbol{A}^* \boldsymbol{R}_\epsilon^{-1} \boldsymbol{A} + \boldsymbol{P}_k^{-1}\right)^{-1} \\ \boldsymbol{W}_k &= \left(\boldsymbol{A}^* \boldsymbol{R}_\epsilon^{-1} \boldsymbol{A} + \boldsymbol{P}_k^{-1}\right)^{-1} \boldsymbol{A}^* \boldsymbol{R}_\epsilon^{-1} \end{aligned}$$

L'hypothèse des spectres constants par domaines $\mathcal{D}_l$ accélère le calcul des statistiques (V.8). En effet, les matrices $\boldsymbol{V}_k = \boldsymbol{V}_l$ et $\boldsymbol{W}_k = \boldsymbol{W}_l$ sont constantes dans le domaine $\mathcal{D}_l$ et les statistiques s'écrivent,

$$\left\{ \begin{array}{rcl} \boldsymbol{R}_{xx}(l) &=& \frac{1}{w_l} \sum\limits_{k \in \mathcal{D}_l} \boldsymbol{x}_k \boldsymbol{x}_k^* \longrightarrow \text{ calculée hors ligne} \\ \boldsymbol{R}_{xs}(l) &=& \boldsymbol{R}_{xx}(l) \boldsymbol{W}_l^* \\ \boldsymbol{R}_{ss}(l) &=& \boldsymbol{W}_l \boldsymbol{R}_{xx}(l) \boldsymbol{W}_l^* + \boldsymbol{V}_l \end{array} \right. \tag{V.9}$$

Nous commençons par donner le pseudo code de l'algorithme EM dans le cas où on néglige les lobes de convolution des instruments.

EM spectral

```
Initialisation :
   calcul hors ligne des covariances empiriques R_xx(l)
   fixer des valeurs initiales pour A, R_ε et P_l
répéter jusqu'à convergence,
    //--- étape-E ---//
   calculer les statistiques pour l=1 à L,
    V_l = (A R_ε^{-1} A^* + P_l^{-1})^{-1}
    R_xs(l) = R_xx(l) R_ε^{-1} A V_l
    R_ss(l) = V_l A^* R_ε^{-1} R_xx(l) R_ε^{-1} A V_l + V_l
   fin de la boucle sur l,
   R_xs = (1/K) Σ w_l R_xs(l)
   R_ss = (1/K) Σ w_l R_ss(l)
    //--- étape-M----//
   A = R_xs R_ss^{-1}
   R_ε = diag(R_xx − R_xs R_ss^{-1} R_xs^*)
   P_l = diag(R_ss(l)), pour l=1 à L
    Renormaliser A et P_l
fin de répéter
```

(V.10)

En tenant compte de la convolution et en supposant que le noyau $\boldsymbol{B}_l$ est constant sur le domaine $\mathcal{D}_l$, les équations de ré-estimation de la matrice $\boldsymbol{A}$ et de la covariance $\boldsymbol{R}_\epsilon$ sont modifiées et la ré-estimation des spectres demeure inchangée.

EM spectral convolutif

1 : Initialisation :
2 : calcul hors ligne des covariances empiriques $\boldsymbol{R}_{xx}(l)$
3 : fixer des valeurs initiales pour $\boldsymbol{A}$, $\boldsymbol{R}_\epsilon$, $\boldsymbol{P}_l$
4 : répéter jusqu'à convergence,
5 : //--- étape-E ---//
6 : calculer les statistiques pour l=1 à L,
7 : $\boldsymbol{V}_l = \left(\boldsymbol{A}\boldsymbol{R}_\epsilon^{-1}\boldsymbol{A}^* + \boldsymbol{P}_l^{-1}\right)^{-1}$
8 : $\boldsymbol{R}_{xs}(l) = \boldsymbol{R}_{xx}(l)\boldsymbol{R}_\epsilon^{-1}\boldsymbol{A}\boldsymbol{V}_l$
9 : $\boldsymbol{R}_{ss}(l) = \boldsymbol{V}_l\boldsymbol{A}^*\boldsymbol{R}_\epsilon^{-1}\boldsymbol{R}_{xx}(l)\boldsymbol{R}_\epsilon^{-1}\boldsymbol{A}\boldsymbol{V}_l + \boldsymbol{V}_l$
10 : $\boldsymbol{R}_\epsilon(l) = diag(\boldsymbol{R}_{xx}(l) + \boldsymbol{B}_l\boldsymbol{A}\boldsymbol{R}_{ss}(l)\boldsymbol{A}^*\boldsymbol{B}_l^* - \boldsymbol{B}_l\boldsymbol{A}\boldsymbol{R}_{xs}(l) - \boldsymbol{R}_{xs}(l)\boldsymbol{A}^*\boldsymbol{B}_l^*)$ (V.11)
11 : fin de la boucle sur l,
12 : //--- étape-M ---//
13 : $Vec\,(\boldsymbol{A}) = \left[\sum_{l=1}^{L} w_l \boldsymbol{R}_{ss}(l) \otimes \boldsymbol{B}_l^2\right]^{-1} Vec\left(\sum_{l=1}^{L} w_l \boldsymbol{B}_l \boldsymbol{R}_{xs}(l)\right)$
14 : $\boldsymbol{R}_\epsilon = \frac{1}{K}\sum_{l=1}^{L} w_l \boldsymbol{R}_\epsilon(l)$
15 : $\boldsymbol{P}_l = diag(\boldsymbol{R}_{ss}(l))$, pour l=1 à L
16 : Renormaliser $\boldsymbol{A}$ et $\boldsymbol{P}_l$
17 : fin de répéter

Les modifications dues à la convolution sont :

1. l'ajout de la ligne 10,
2. la modification des lignes 14 et 15 de l'EM spectral.

V.3.2 Domaine des harmoniques sphériques

Lorsqu'on traîte des données sur l'ensemble du ciel, la base naturelle est celle des harmoniques sphériques. Un signal $x(\vec{r})$ se décompose sous la forme :

$$x(\vec{r}) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} x(l,m) Y_l^m(\vec{r}).$$

où Y_l^m est la base des harmoniques sphériques. Les coefficients $x(l,m)$ sont définis par :

$$x(l,m) = \int_{4\pi} x(\vec{r}) Y_l^m(\vec{r})^* d\,\Omega. \tag{V.12}$$

La transformation en harmoniques sphériques des observations $\{\boldsymbol{x}(\vec{r})\}_{\vec{r}\in\mathcal{D}}$ ($\mathcal{D}$ est l'ensemble du ciel) conserve la linéarité du mélange (V.4) et la convolution devient une multiplication. La relation entre les observations et les sources s'écrit donc,

$$\boldsymbol{x}(l,m) = \boldsymbol{B}(l)\boldsymbol{A}\boldsymbol{s}(l,m) + \boldsymbol{\epsilon}(l,m)$$

où $\boldsymbol{B}(l) = diag(b_d(l))$ est la transformée du noyau de convolution qui ne dépend que de l à cause de sa symétrie. Pour un noyau gaussien, $b_d(l) \approx \exp\left[-\sigma_d^2 l(l+1)/2\right]$ et $\sigma_d = \Theta_d/\sqrt{8\ln 2}$, où Θ_d est la largeur à mi-hauteur de la gaussienne.

La suite du développement est similaire au cas de la transformée en Fourier. Les sources $\boldsymbol{s}(l,m)$ sont supposées blanches gaussiennes non stationnaires. Les spectres de puissances des sources $\{\sigma_{lm}^2\}$ contenus dans les termes diagonaux des matrices $\boldsymbol{P}_{l,m} = \mathrm{E}\left[\boldsymbol{s}_{lm}\boldsymbol{s}_{lm}^*\right]$ ne dépendent pas de l'indice m à cause de l'hypothèse de l'isotropie. Les termes dans le produit constituant la vraisemblance peuvent être partitionnés de telle manière qu'on tienne compte de la symétrie des spectres de puissance :

$$\mathrm{p}(\{\boldsymbol{x}_{lm}\} \mid \boldsymbol{\theta}) = \prod_{l=1}^{L} |2\,\pi\,\boldsymbol{R}_l|^{-w_l} \exp\left[-\mathrm{Tr}\left(\boldsymbol{R}_l^{-1}\sum_{m=-l}^{l}\boldsymbol{x}_{lm}\boldsymbol{x}_{lm}^*\right)\right] \tag{V.13}$$

où $w_l = |\mathcal{D}_l| = (2l+1)$ est le nombre de cœfficients appartenant à la bande $\mathcal{D}_l = \{m \mid -l \leq m \leq l\}$.

On peut élargir les bandes $\mathcal{D}_l$ en partitionnant L en Q intervalles. Chaque bande $\mathcal{D}_q$ pour $q = 1..Q$ est définie par (voir figure (V.6)) :

$$\mathcal{D}_q = \{l, m \mid l_{min}(q) \leq l \leq l_{max}(q) \; et \; -l \leq m \leq l\}$$

On suppose que les spectres de puissance sont constants à l'intérieur de chaque bande $\mathcal{D}_q$ qui contient $w_q = \sum_{l=l_{min}(q)}^{l_{max}(q)} (2l+1)$ modes.

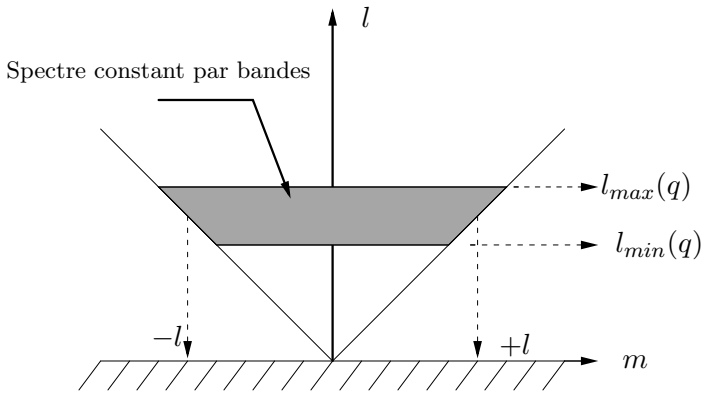


FIG. V.6: Les spectres des sources sont constants par bandes. Le critère du maximum de vraisemblance est un ajustement des matrices de covariance.

Le logarithme de la vraisemblance est donc, à une constante additive près, une somme pondérée des divergences de Kullback-Leibler entre les covariances théoriques $\boldsymbol{R}_q = \boldsymbol{A}\boldsymbol{P}_q\boldsymbol{A}^* + \boldsymbol{R}_\epsilon$ et les covariances empiriques $\hat{\boldsymbol{R}}_q = \sum_{(l,m)\in\mathcal{D}_q} \boldsymbol{x}_{lm}\boldsymbol{x}_{lm}^* \,/\, w_q$,

$$\begin{aligned} \log \mathrm{p}(\{\boldsymbol{x}_{lm}\} \mid \boldsymbol{\theta}) &= -\sum_{q=1}^{Q} w_q \left(-\ln |\boldsymbol{R}_q^{-1}\hat{\boldsymbol{R}}_q| + \mathrm{Tr}\left(\boldsymbol{R}_q^{-1}\hat{\boldsymbol{R}}_q\right) - N_d \right) + cste \\ &= -\sum_{q=1}^{Q} w_q D_{KL}(\boldsymbol{R}_q, \hat{\boldsymbol{R}}_q) + cste. \end{aligned} \tag{V.14}$$

L'algorithme EM possède la même structure que l'algorithme EM spectral (V.10) (ou (V.11) quand on tient compte de la convolution). Au lieu de travailler sur les anneaux concentriques $\mathcal{D}_l$ (figure (V.5)), on travaille sur les bandes $\mathcal{D}_q$ (figure (V.6)).

V.4 Résultats de simulation

Nous avons appliqué la méthode de séparation proposée sur des simulations d'observation de la mission **Planck** sur l'ensemble du ciel.

Simulations d'observations

Les composantes (sources) sont : le CMB, l'émission thermique de la poussière, l'effet synchrotron et les effets SZ thermiques et cinétiques des amas de galaxies (voir figure (V.8)). Elles sont obtenues de la façon suivante.

Les modes s(l,m) du CMB sont générés suivant une statistique gaussienne de variances $C(l)$, prédites par CMBFast [Seljak et Zaldarriaga, 2000], en utilisant des paramètres cosmologiques standards. Les composantes galactiques ont été obtenues à partir des cartes observées par d'autres expériences à des fréquences très différentes. L'émission de la poussière galactique est modélisée en utilisant les cartes à 300 GHz de l'analyse des données DIRBE-IRAS. L'émission synchrotron est simulée à partir des cartes à 408 MHz auxquelles ont été ajoutées des structures aux petites échelles [Stolyarov *et al.*, 2002]. Les effets SZ thermiques et cinétiques ont été entièrement simulés [Eke *et al.*, 1998]. Notons que l'effet SZ cinétique et le CMB ont des lois d'émission proportionnelles. Les simulations ont été réalisées jusqu'à la résolution de 3.5 minutes d'arc.

Ces cinq composantes ainsi que du bruit blanc aux niveaux nominaux des instruments de Planck ont été mélangés suivant le modèle (V.4) à toutes les fréquences des instruments de Planck ($30, 44, 70, 100$ GHz pour l'instrument de basse fréquence et $100, 143, 217, 353, 545, 857$ GHz pour l'instrument haute fréquence). La figure (V.9) montre les 10 cartes d'observation. Les résolutions angulaires des observations sont par ordre croissant de fréquence : $33, 23, 14, 10, 10.6, 7.4, 4.9, 4.5, 4.5, 4.5$ minutes d'arc.

Résultats

Les éléments de la matrice de mélange (les spectres électromagnétiques) sont estimés avec une bonne précision. Le tableau (V.7) donne le rapport entre les éléments de la matrice de mélange estimés et les paramètres vrais relatifs au CMB. Les éléments de la matrice de mélange relatifs à l'effet SZ thermique sont estimés avec une bonne précision. La loi d'émission de l'effet synchrotron est très bien contrainte aux basses fréquences, ainsi que la loi d'émission de la poussière aux plus hautes fréquences. Par ailleurs, notre méthode permet de contraindre les composantes galactiques à des fréquences très éloignées de leur maximum d'émission.

Fréquence	30	44	70	100 (LFI)
CMB	0.999984	1.000254	0.999780	1.000081

Fréquence	100 (HFI)	143	217	353	545
CMB	1	0.999993	0.999836	0.998972	0.990155

Fig. V.7: Rapport entre les valeurs estimées et les vraies valeurs du spectre électromagnétique.

La figure (V.10) montre l'estimation du spectre de puissance du CMB. On note que la méthode proposée permet d'estimer précisément le spectre de puissance jusqu'au multipôle $l = 2500$. Aux plus petites échelles, la dispersion commence à être significative. Ce résultat n'est pas surprenant puisque le bruit et l'effet du lobe sont importants à ces échelles angulaires pour tous les détecteurs. Par ailleurs, les spectres de puissance estimés ne semblent pas être contaminés par l'effet SZ cinétique. Afin d'illustrer les performances de l'estimation du spectre en aveugle (ne connaissant pas la matrice de mélange $\boldsymbol{A}$), nous avons tracé sur la figure (V.11) les erreurs relatives $|\tilde{C}(q) - C(q)| \;/\; C(q)$ ($C(q)$ est le vrai spectre) commises sur l'estimation du spectre en aveugle et en semi-aveugle (en fixant les éléments de $\boldsymbol{A}$ à leurs vraies valeurs). On note l'équivalence des deux cas et donc que l'estimation conjointe de la matrice de mélange n'affecte pas la précision de l'estimation du spectre du CMB.

La figure (V.12) illustre les cartes reconstruites par filtrage de Wiener après la convergence de l'algorithme de séparation. Nous avons choisi de déterminer quatre composantes. Une des composantes dans les simulations, l'effet SZ cinétique, est négligeable à toutes les fréquences. En plus, elle ne peut pas être séparé du CMB par notre approche puisque ces deux composantes ont des lois d'émission proportionnelles (CMB et SZ cinétique forment une seule composante). On note la bonne qualité de la reconstruction des cartes en les comparant aux vraies cartes de la figure (V.8).

Bibliographie

[Baccigalupi *et al.*, 2000] C. Baccigalupi, L. Bedini, C. Burigana, G. De Zotti, A. Farusi, D. Maino, M. Maris, F. Perrotta, E. Salerno, L. Toffolatti et A. Tonazzini. *Monthly Notices of the Royal Astronomical Society*, 318 : 769–780, novembre 2000.

[Bennet et *al*, 2003] C. Bennet et *al.* First year Wilkinson Microwave Anisotropy Probe (WMAP) observations : Preliminary maps and basic results. *submitted to ApJ*, 2003.

[Benoît et *al*, 2003] A. Benoît et *al.* The cosmic microwave background anisotropy power spectrum measured by Archeops. *Astronomy and Astrophysics*, 399 : L19–L23, mars 2003.

[Bouchet et Gispert, 1999] F. R. Bouchet et R. Gispert. *New Astronomy*, 4 : 443–479, novembre 1999.

[Cardoso *et al.*, 2002] J. Cardoso, H. Snoussi, J. Delabrouille et G. Patanchon. Blind separation of noisy gaussian stationary sources. application to cosmic microwave background imaging. In *Eusipco*, Toulouse, septembre 2002.

[De Bernardis et *al*, 2000] P. De Bernardis et *al.* A flat Universe from high-resolution maps of the cosmic microwave background radiation. *Nature*, 404 : 955–959, 2000.

[De Zotti *et al.*, 1999] G. De Zotti, L. Toffolatti, F. Argüeso, R. D. Davies, P. Mazzotta, R. B. Partridge, G. F. Smoot et N. Vittorio. In *AIP Conf. Proc. 476 : 3K cosmology*, page 204, 1999.

[Delabrouille *et al.*, 2001] J. Delabrouille, G. Patanchon et E. Audit. *Monthly Notices of the Royal Astronomical Society*, 2001.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird et D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39 : 1–38, 1977.

[Eke *et al.*, 1998] V. R. Eke, J. F. Navarro et C. S. Frenk. The evolution of X-Ray Clusters in a Low Density Universe. *APJ*, 503 : 569, novembre 1998.

[Grainge et *al*, 2003] K. Grainge et *al.* The CMB power spectrum out to l=1400 measured by the VSA. *MNRAS in press*, 2003.

[Hanany et *al*, 2000] S. Hanany et *al.* MAXIMA-1 : A Measurement of the cosmic microwave background anisotropy on angular scales of 10 arcminutes to 5 degrees. *ApJ Letters*, 545 : L5–L9, décembre 2000.

[Hobson *et al.*, 1998] M. Hobson, A. W. Jones, A. N. Lasenby et F. R. Bouchet. *Monthly Notices of the Royal Astronomical Society*, 300 : 1–29, octobre 1998.

[Hu et Sugiyama, 1996] W. Hu et N. Sugiyama. *APJ*, 471 : 542, novembre 1996.

[Jungman *et al.*, 1996] G. Jungman, M. Kamionkowski, A. Kosowsky et D. N. Spergel. *Physical Review Letters*, 76 : 1007–1010, février 1996.

[Kuo et *al*, 2002] C. Kuo et *al.* High resolution observations of the CMB power spectrum with ACBAR. *ApJ, available at astro-ph/0212289*, 2002.

[Pearson et *al*, 2003] T. Pearson et *al.* The anisotropy of the microwave background to l = 3500 : Mosaic observations with the Cosmic Background Imager. *Accepted by The Astrophysical Journal*, 2003.

[Pham et Cardoso, 2001] D.-T. Pham et J. Cardoso. Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. Signal Processing*, 49, 9 (11) : 1837–1848, 2001.

[Seljak et Zaldarriaga, 2000] U. Seljak et M. Zaldarriaga. *ApJ. Suppl. ser.*, 129 : 431, 2000.

[Snoussi *et al.*, 2001] H. Snoussi, G. Patanchon, J. Macías-Pérez, A. Mohammad-Djafari et J. Delabrouille. Bayesian blind component separation for cosmic microwave background observations. In R. L. Fry, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 125–140. MaxEnt Workshops, Amer. Inst. Physics, août 2001.

[Stolyarov *et al.*, 2002] V. Stolyarov, M. P. Hobson, M. A. J. Ashdown et A. N. Lasenby. *Monthly Notices of the Royal Astronomical Society*, 336 : 99–111, 2002.

[Tegmark et Estathiou, 1996] M. Tegmark et G. Estathiou. A method for subtracting foregrounds from multifrequency CMB sky maps. *Monthly Notices of the Royal Astronomical Society*, 281 : 1297, 1996.

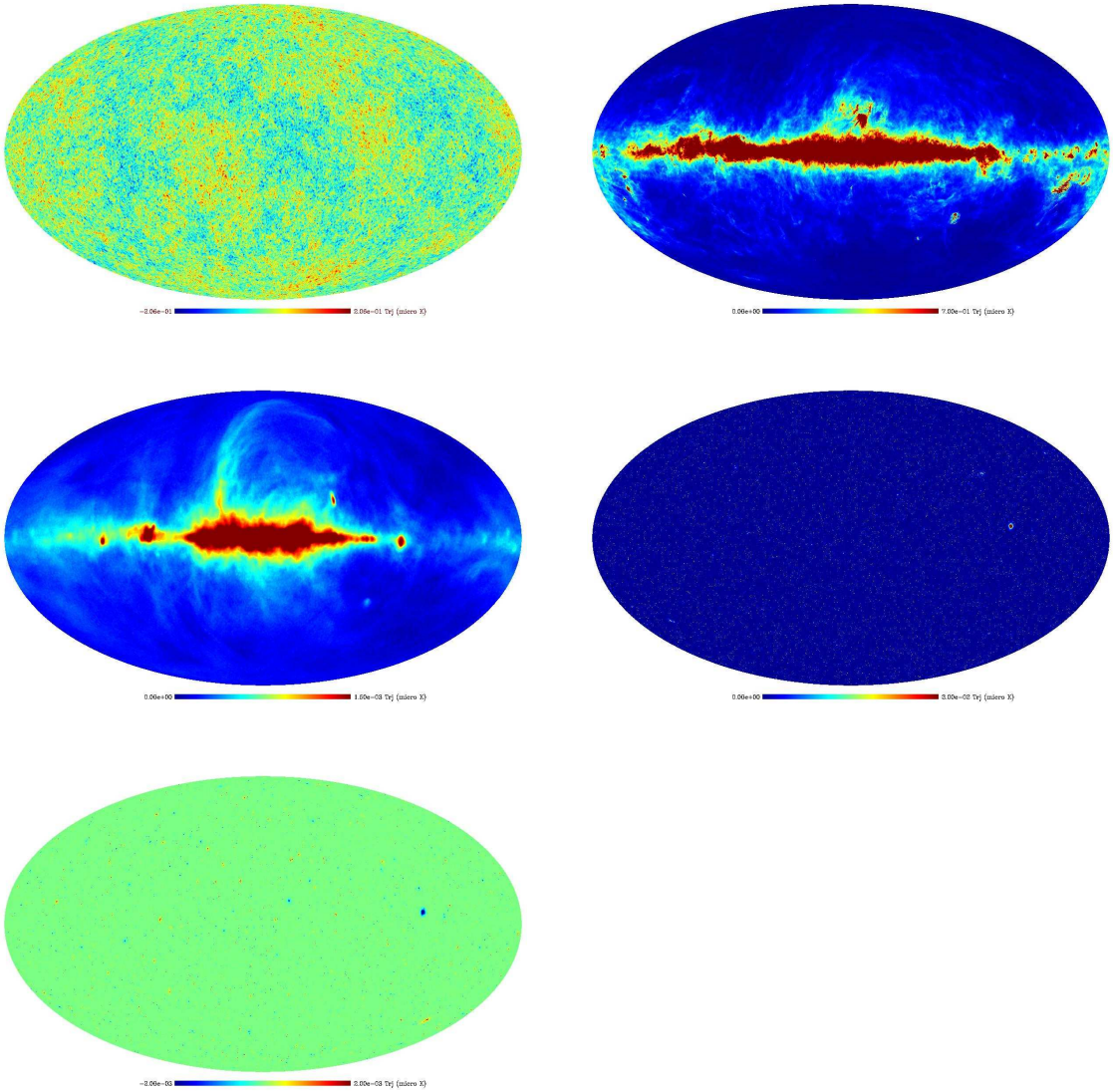


FIG. V.8: Les cartes sur l'ensemble du ciel représentant les composantes sources utilisées pour tester la méthode de séparation

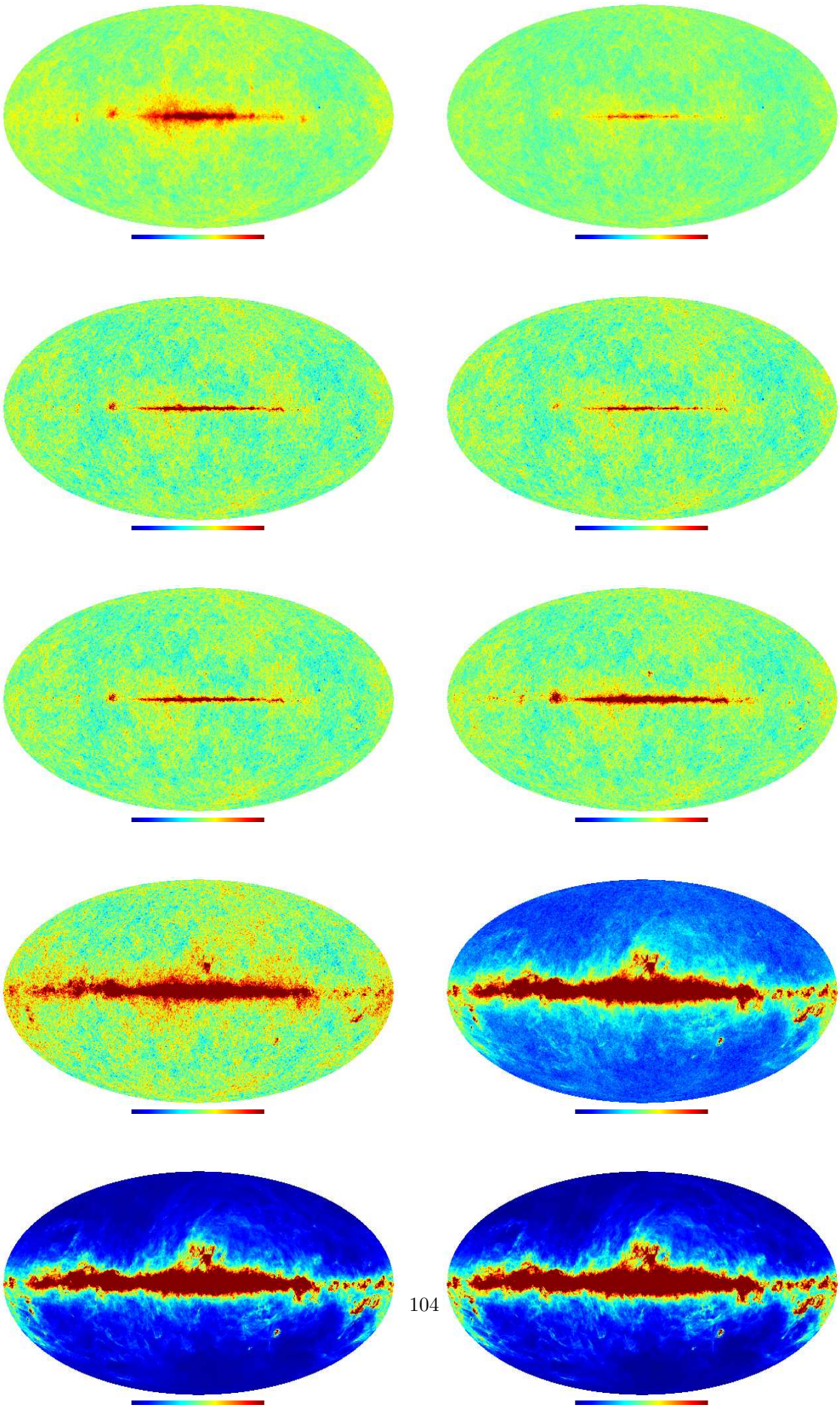


FIG. V.9: Les simulations des observations de la mission Planck.

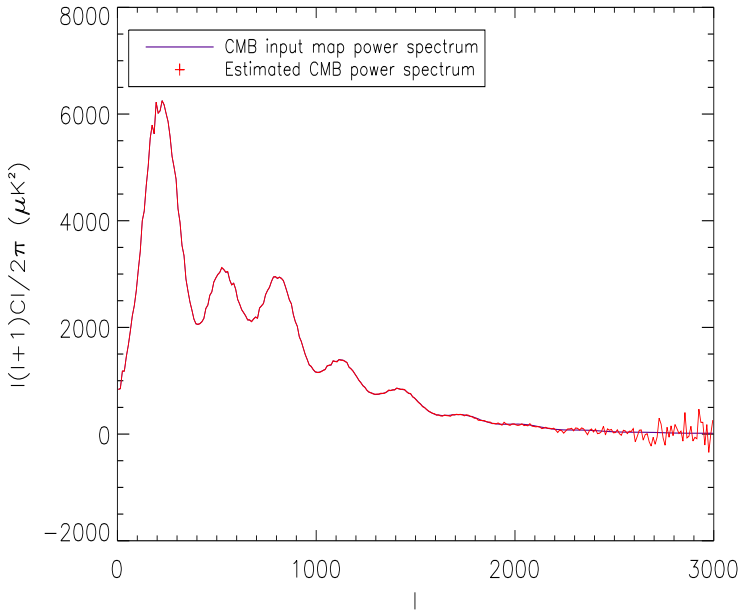


FIG. V.10: Estimation du spectre de puissance du CMB.

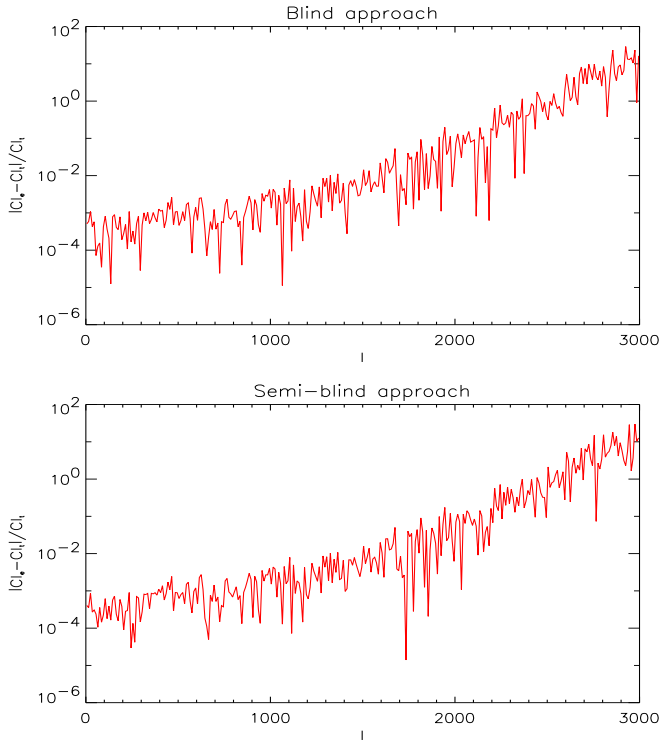


FIG. V.11: Erreurs relatives de l'estimation du spectre du CMB dans le cas aveugle et semi-aveugle.

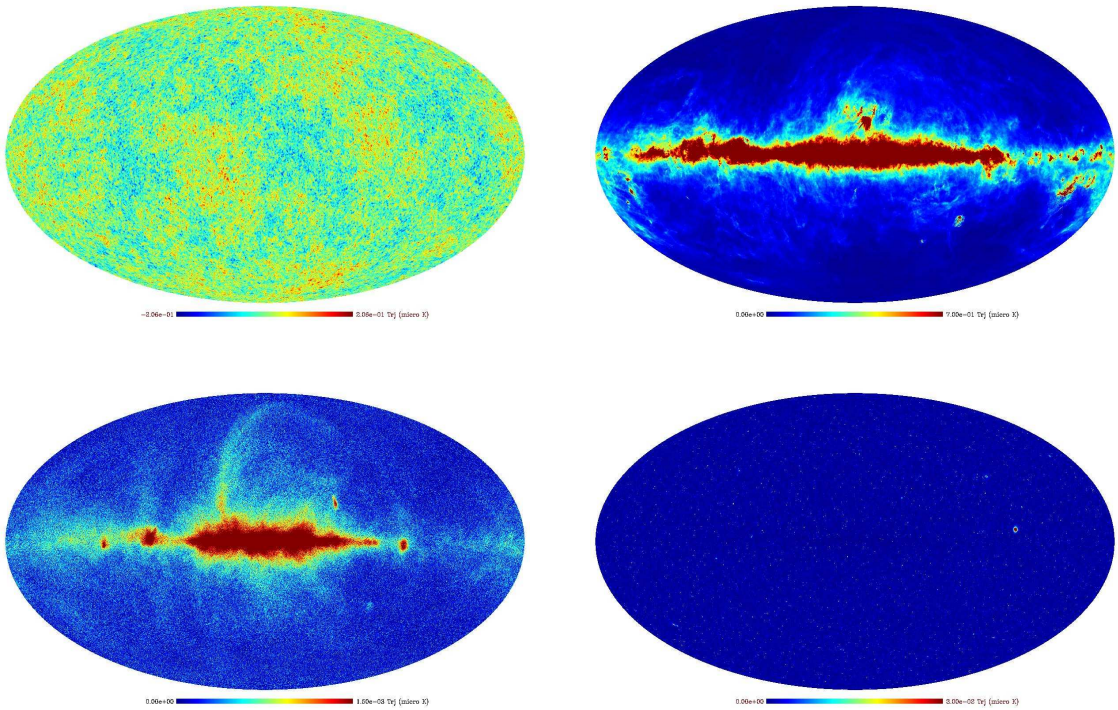
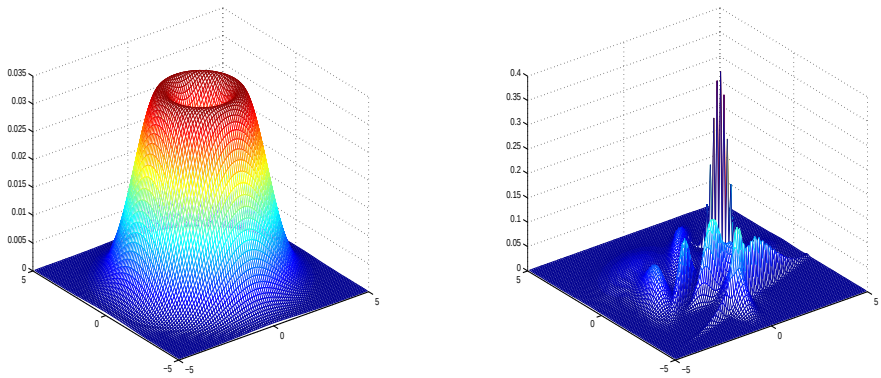


FIG. V.12: Les cartes des composantes reconstruites en aveugle.

Chapitre VI

DÉGÉNÉRESCENCE DU MAXIMUM DE VRAISEMBLANCE



Dans ce chapitre, nous décrivons le problème de la dégénérescence du maximum de vraisemblance dans le cas d'un mélange de gaussiennes multivariables. Nous montrons que la vraisemblance n'est pas bornée et nous caractérisons l'ensemble des points de singularité. La pénalisation de la vraisemblance par un *a priori* **inverse wishart** sur les matrices de covariance élimine cette dégénérescence sans compliquer les équations de ré-estimation de l'algorithme EM. Nous apportons également d'autres modifications à l'algorithme EM afin de tenir compte de certaines contraintes sur la structure des matrices de covariance. Nous montrons que le risque de dégénérescence existe aussi dans le cas de la séparation de sources et que la pénalisation par un *a priori* conjugué inverse wishart élimine cette dégénérescence.

VI.1 Introduction

On considère un processus doublement stochastique :

1. une première couche de variables discrètes $(z_t)_{t=1..T}$ où chaque variable z_t prend ses valeurs dans un ensemble discret $\mathcal{Z} = \{1..K\}$,
2. et une deuxième couche de variables continues $(\boldsymbol{s}_t)_{t=1..T}$ où chaque vecteur $\boldsymbol{s}_t$ prend ses valeurs dans $\mathbb{R}^n$.

Conditionnellement à la première couche $z_{1..T}$, les variables $(\boldsymbol{s}_t)_{t=1..T}$ sont temporellement blanches : $\mathrm{p}(\boldsymbol{s}_{1..T} \mid z_{1..T}) = \prod_{t=1}^{T} \mathrm{p}(\boldsymbol{s}_t \mid z_t)$. On suppose que les lois $\mathrm{p}(\boldsymbol{s} \mid z)$ sont paramétriques ayant la même forme $f(\boldsymbol{s} \mid \zeta_z)$ mais se distinguent par la valeur du paramètre ζ_z qui dépend de la variable $z \in \{1..K\}$. Dans la suite, on suppose que cette loi est gaussienne et donc que le vecteur des paramètres ζ_z contient la moyenne $\boldsymbol{\mu}_z$ et la covariance $\boldsymbol{R}_z$ relatives à z.

La première couche $z_{1..T}$ peut être considérée comme un processus de classification. Chaque observation $\boldsymbol{s}$ appartient à un groupe z modélisé statistiquement par une gaussienne $\mathcal{N}(. \mid \boldsymbol{\mu}_z, \boldsymbol{R}_z)$.

Dans ce travail, on suppose que les étiquettes $z_{1..T}$ suivent une loi générale paramétrique $\mathrm{p}(z_{1..T} \mid \boldsymbol{\pi})$. La forme de cette loi n'intervient pas d'une manière significative dans les développements qui vont suivre. On rappelle toutefois quelques cas particuliers souvent traîtés dans la littérature relative à ce sujet.

– Les étiquettes $z_{1..T}$ sont i.i.d. : le vecteur des paramètres $\boldsymbol{\pi}$ est alors formé par les K probabilités discrètes : $\{\pi_k = \mathrm{p}(z = k)\}_{k=1..K}$. Les sources sont marginalement blanches :

$$\mathrm{p}(\boldsymbol{s}_{1..T}) = \prod_{t=1}^{T} \mathrm{p}(\boldsymbol{s}_t) = \prod_{t=1}^{T} \sum_{k=1}^{K} \mathrm{p}(z_t = k)\, \mathrm{p}(\boldsymbol{s}_t \mid z_t = k) \tag{VI.1}$$

Ce cas est le plus connu dans la littérature sous le nom de "modèle de mélange de gaussiennes" (du fait de la somme dans l'expression (VI.1)).

– Les étiquettes $z_{1..T}$ forment une chaîne de Markov : la propriété de Markov modélise la dépendance temporelle des étiquettes et par conséquent la dépendance temporelle des sources. Le vecteur des paramètres $\boldsymbol{\pi}$ est formé dans ce cas par le vecteur des probabilités initiales $\boldsymbol{\pi}^0$ et par la matrice de transition $\mathbb{P}$. La probabilité d'une chaîne d'étiquettes $z_{1..T}$ s'écrit :

$$\mathrm{p}(z_{1..T} \mid \boldsymbol{\pi}) = \boldsymbol{\pi}^0(z_0)\, \mathbb{P}_{z_0 z_1} \dots \mathbb{P}_{z_{t-1} z_t} \dots \mathbb{P}_{z_{T-1} z_T}.$$

L'avantage de cette modélisation est de tenir compte de la dépendance des observations $\boldsymbol{s}_{1..T}$ via la couche cachée $\boldsymbol{z}_{1..T}$. Ce modèle est désigné dans la littérature par le modèle de Markov caché (HMM en anglais).

– Les étiquettes sont définies sur une image ($z_{1..T} = \boldsymbol{Z}$) et forment un champ de Markov. En définissant un système de voisinage ∂ (voir chapitre (IV) pour plus de détails), la propriété de Markov s'exprime par :

$$Pr(Z_r \mid Z_{\mathcal{S}\setminus r}) = Pr(Z_r \mid Z_{\partial(r)}). \tag{VI.2}$$

Le vecteur $\boldsymbol{\pi}$ contient ainsi les probabilités conditionnelles[1] (VI.2). On désigne ce modèle par un modèle de champ de Markov caché (HMF en anglais).

Ce modèle suscite un grand intérêt dans la communauté du traitement du signal et de l'image. Parmi ses avantages, on peut mentionner les points suivants.

– Le modèle de mélange représente une alternative intéressante à la modélisation non paramétrique. En augmentant le nombre d'étiquettes K, on peut atteindre toute loi de probabilité (regarder [Roeder et Wasserman, 1997] pour l'utilisation des mélanges de gaussiennes pour l'estimation des densités).

[1] Dans le cas 2-D, on utilise plutôt la forme de Gibbs équivalente à un champ de Markov. Le vecteur $\boldsymbol{\pi}$ contient donc les paramètres des potentiels U_c de la distribution de Gibbs.

– Les signaux réels s'apprêtent bien à cette modélisation. A titre d'exemple, le traitement des signaux de parole est un bon domaine d'application pour les chaînes de Markov cachées [Rabiner, 1989]. Dans [Snoussi et Mohammad-Djafari, 2002a] et [Snoussi et Mohammad-Djafari, 2002b], ce modèle a été utilisé dans les problèmes de séparation de sources.
– Cette modélisation représente un outil statistique efficace pour les problèmes de classification [McLachlan et Basford, 1987] et de ségmentation 2-D [Descombes, 1993].
– L'identification des paramètres du mélange repose sur l'algorithme EM [Dempster *et al.*, 1977] qui peut être implémenté d'une manière efficace et rapide.

Maximum de vraisemblance

Quelque soit le modèle pris pour les étiquettes, la loi marginale des sources s'écrit :

$$\mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{\theta}) = \sum_{z_{1..T}} \mathrm{p}(z_{1..T} \mid \boldsymbol{\pi}) \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{s}_t \,;\, \boldsymbol{\mu}_{z_t}, \boldsymbol{R}_{z_t}) \tag{VI.3}$$

où $\boldsymbol{\theta}$ représente les paramètres $(\boldsymbol{\pi}, \boldsymbol{\mu}_z, \boldsymbol{R}_z)$.

Ayant observé les données $\boldsymbol{s}_{1..T}$, notre objectif est l'identification de $\boldsymbol{\theta}$. Parmi les différentes approches possibles (rappelées dans [McLachlan et Peel, 2000]), le maximum de vraisemblance est la méthode la plus utilisée. Ceci est dû aux propriétés asymptotiques de consistance et d'efficacité de l'estimateur du maximum de vraisemblance (sous certaines conditions de régularité) et la possibilité d'implémenter l'algorithme EM [Dempster *et al.*, 1977] (*Expectation-Maximization*) qui représente un outil efficace dans les situations où on a des problèmes à variables cachées.

On note $\boldsymbol{\Theta}$ l'ensemble des paramètres :

$$\boldsymbol{\Theta} = \left\{ \boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}_k, \boldsymbol{R}_k) \mid \sum_{z_{1..T}} \mathrm{p}(z_{1..T} \mid \boldsymbol{\pi}) = 1,\ \boldsymbol{\mu}_k \in \mathbb{R}^n,\ \boldsymbol{R}_k \in \mathcal{C},\ k = 1..K \right\}$$

où $\mathcal{C}$ est un sous espace **fermé** de l'ensemble des matrices symétriques positives. On donnera des exemples de tels sous espaces dans la section (VI.4) (voir aussi [Burg, 1982]).

Remarque 15 *On n'impose pas aux matrices de covariance d'appartenir strictement à l'ensemble des matrices symétriques définies positives (régulières). En effet, cet ensemble topologique est ouvert. Sa frontière est formée par les matrices symétriques positives singulières. On préfère travailler plutôt avec son adhérence (ensemble fermé) qui coïncide avec l'ensemble de toutes les matrices symétriques positives. Nous donnerons plus loin les raisons de ce choix.*

L'estimateur du maximum de vraisemblance est défini par :

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{\theta}). \tag{VI.4}$$

Placement du travail

Les points suivants résument les contributions de ce travail.

Caractérisation de la dégénérescence : la dégénérescence du maximum de vraisemblance due au fait que la vraisemblance (VI.3) n'est pas bornée est bien connue dans la littérature [Kiefer et Wolfowitz, 1956; Day, 1969; McLachlan et Peel, 2000; Ridolfi et Idier, 1999]. Dans ce travail, nous donnons une caractérisation mathématique rigoureuse de l'ensemble des points de singularité. Cette caractérisation généralise celle étudiée dans le cas scalaire [Ridolfi et Idier, 1999] au cas plus général des données vectorielles. Cette caractérisation explique la croissance du risque de dégénérescence avec la dimension n des données.

Elimination de la dégénérescence : la pénalisation de la vraisemblance avec un *a priori* conjugué (inverse wishart) élimine cette dégénérescence [Ormoneit et Tresp, 1998]. Cette solution est beaucoup plus efficace que l'optimisation sous contrainte [Hathaway, 1986]. Notre contribution consiste à construire, d'une manière générale, en se basant sur la caractérisation des points de singularité, une classe d'*a priori* garantissant l'élimination de la dégénérescence. L'*a priori* **inverse wishart** fait partie de cette classe et présente l'avantage de garder explicites les équations de ré-estimation de l'algorithme EM.

Existence de la solution : nous montrons, en se basant sur les propriétés topologiques de l'espace des paramètres $\boldsymbol{\Theta}$, l'existence d'au moins un maximiseur de la vraisemblance pénalisée.

Estimation des matrices structurées : Grâce à la pénalisation, nous avons pu tenir compte de la contrainte de régularité des matrices de covariance. La difficulté de cette contrainte (régularité des matrices) est que l'espace correspondant est ouvert. Par contre, dans le cas où on contraint les matrices à varier dans des espaces fermés, nous montrons comment modifier d'une manière simple l'algorithme EM afin de tenir compte de cette contrainte.

Dégénérescence en séparation de sources : nous montrons aussi, dans le cas où les données $\boldsymbol{s}_{1..T}$ ne sont pas directement observées mais plutôt mélangées et bruitées, que le risque de dégénérescence existe. Nous donnons une caractérisation des points de singularité (qui est plus compliquée que dans le cas précédent) et nous montrons que la pénalisation par un inverse wishart sur la matrice de covariance du bruit élimine ce risque de dégénérescence.

VI.2 Dégénérescence du maximum de vraisemblance

Proposition 2 (La vraisemblance n'est pas bornée) $\forall\ \boldsymbol{s}_{1..T} \in (\mathbb{R}^n)^T$, *il existe des points de singularité* $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ *tels que :* $\lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}^*} p(\boldsymbol{s}_{1..T}\,|\,\boldsymbol{\theta}) = \infty$. *Ces points de singularité* $\boldsymbol{\theta}^* = (\boldsymbol{\pi}^*, \boldsymbol{\mu}_z^*, \boldsymbol{R}_z^*)_{\boldsymbol{z}\in\mathcal{Z}}$ *sont tels que au moins une des matrices* $\boldsymbol{R}_z^*$ *est positive singulière et la moyenne* $\boldsymbol{\mu}_z$ *correspondante appartient à l'intersection de* $[n - rang(\boldsymbol{R}_z)]$ *hyperplans de* $\mathbb{R}^n$.

Preuve : Dans la suite, toutes les matrices sont symétriques. On rappelle que toutes les matrices de covariance sont supposées appartenir à $\mathcal{C}$ sous espace fermé de l'espace $\mathcal{S}_+$ des matrices positives. La raison pour laquelle on préfère travailler avec des espaces fermés est que toute suite convergente à valeurs dans un espace fermé $\mathcal{F}$ converge vers un point appartenant à $\mathcal{F}$. $\mathcal{S}_+$ est lui même fermé et connexe. L'ensemble des matrices positives définies $\mathcal{R}_+$ est un sous espace ouvert de $\mathcal{S}_+$. Sa frontière avec l'ensemble de toutes les matrices est formée par les matrices positives singulières. L'adhérence de $\mathcal{R}_+$ est égale à $\mathcal{S}_+$.

On considère un point $\boldsymbol{\theta}^*$ de $\boldsymbol{\Theta}$ vérifiant la propriété suivante :

Propriété 1 $\exists k \in \mathcal{Z}$ *tel que* $\boldsymbol{R}_k^*$ *est une matrice positive singulière de rang* $p_k < n$.

On partitionne l'ensemble $\mathcal{Z} = \mathcal{Z}_s \cup \mathcal{Z}_r$ de telle façon que $\forall k \in \mathcal{Z}_s,\ \boldsymbol{R}_k^*$ est singulière et $\forall k \in \mathcal{Z}_r,\ \boldsymbol{R}_k^*$ est régulière ($\mathcal{Z}_s \neq \emptyset$). $\forall k \in \mathcal{Z}_s$ la matrice $\boldsymbol{R}_k^*$ est diagonalisable dans le groupe orthogonal : il existe une matrice unitaire $\boldsymbol{U}_k$ telle que $\boldsymbol{U}_k\,\boldsymbol{U}_k^T = \boldsymbol{I}$ et

$$\boldsymbol{R}_k^* = \boldsymbol{U}_k^T \boldsymbol{\Lambda}_k \boldsymbol{U}_k\ , \qquad \boldsymbol{\Lambda}_k = \begin{bmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & \lambda^*_{n-p_k+1} & & \\ & & & & \ddots & \\ & & & & & \lambda^*_n \end{bmatrix}$$

où les valeurs propres $(\lambda_i^*)_{i=n-p_k+1}^{i=n}$ sont toutes strictement positives.

Afin de montrer la proposition (2), il suffit de construire une suite de points $(\boldsymbol{\theta}^{(q)})_{q\in\mathbb{N}}$ telle que :

$$\lim_{q\to\infty}\boldsymbol{\theta}^{(q)}=\boldsymbol{\theta}^{*}\implies\lim_{q\to\infty}\mathrm{p}(\boldsymbol{s}_{1..T}\mid\boldsymbol{\theta}^{(q)})=\infty.$$

Tout d'abord, fixons $k_0\in\mathcal{Z}_s$ (il existe puisque $\mathcal{Z}_s\neq\emptyset$) et supposons que $\boldsymbol{\theta}^*$ vérifie en plus la propriété suivante :

Propriété 2 *La moyenne $\boldsymbol{\mu}^*_{k_0}$ est telle que $\left[\boldsymbol{U}_{k_0}(\boldsymbol{s}_i-\boldsymbol{\mu}^*_{k_0})\right]_j=0,\ j=1..(n-p_{k_0})$ pour un instant i donné.*

Autrement dit, la moyenne $\boldsymbol{\mu}^*_{k_0}$ appartient à l'intersection de $(n-p_{k_0})$ hyperplans de $\mathbb{R}^n$. Lorsque $p_{k_0}=0$, c'est-à-dire que la matrice $\boldsymbol{R}^*_{k_0}$ est nulle, la moyenne $\boldsymbol{\mu}^*_{k_0}$ coincide avec l'observation $\boldsymbol{s}_i$.

Considérons la suite $\boldsymbol{\theta}^{(q)}$ suivante :

$$\begin{cases} k\in\mathcal{Z}_r & \begin{cases}\boldsymbol{\mu}_k^{(q)}=\boldsymbol{\mu}_k^*\\ \boldsymbol{R}_k^{(q)}=\boldsymbol{R}_k^*\end{cases}\\[2ex] k\in\mathcal{Z}_s & \begin{cases}\boldsymbol{\mu}_k^{(q)}=\boldsymbol{\mu}_k^*\\ \boldsymbol{R}_k^{(q)}=\boldsymbol{U}_k^T\boldsymbol{\Lambda}_k^{(q)}\boldsymbol{U}_k,\end{cases}\end{cases} \tag{VI.5}$$

avec les matrices diagonales $\boldsymbol{\Lambda}_k^{(q)}$ définies par :

$$\boldsymbol{\Lambda}_k^{(q)}=\begin{bmatrix}\lambda_1^{(q)} & & & & & & \\ & \ddots & & & & & \\ & & \lambda_{n-p_k}^{(q)} & & & & \\ & & & \lambda^*_{n-p_k+1} & & & \\ & & & & \ddots & & \\ & & & & & \lambda^*_n\end{bmatrix}$$

où toutes les suites numériques $\lambda_i^{(q)}$ sont strictement positives (assurant que les $\boldsymbol{R}_k^{(q)}$ restent dans $\mathcal{R}_+$) et tendent vers 0. Ceci implique que les matrices $\boldsymbol{\Lambda}_k^{(q)}$ tendent vers les matrices $\boldsymbol{\Lambda}_k^*$ et donc que la suite $\boldsymbol{\theta}^{(q)}$ tend vers le point singulier $\boldsymbol{\theta}^*$.

Remarque 16 *Concernant le paramètre $\boldsymbol{\pi}^*$, on suppose qu'il garantit la positivité de $\Pr(z_{1..T}\mid\boldsymbol{\pi}^*)\ \forall z_{1..T}\in\mathcal{Z}^T$ et donc que $\min\limits_{z_{1..T}}\Pr(z_{1..T}\mid\boldsymbol{\pi}^*)>c>0$. Cette hypothèse n'est pas nécessaire mais facilite la démonstration de la proposition. On reviendra sur ce point plus loin.*

Dans la suite, on va montrer qu'en **contrôlant les vitesses de convergence** des suites $\lambda_j^{(q)}$, de chaque composante $k\in\mathcal{Z}_s$, la vraisemblance $\mathrm{p}(\boldsymbol{s}_{1..T}\mid\boldsymbol{\theta}^{(q)})$ tend vers l'infini.

La vraisemblance s'écrit,

$$\mathrm{p}(\boldsymbol{s}_{1..T}\mid\boldsymbol{\theta}^{(q)})=\sum_{z_{1..T}}\mathrm{p}(z_{1..T}\mid\boldsymbol{\pi}^{(q)})\prod_{t=1}^{T}\mathcal{N}(\boldsymbol{s}_t\,;\,\boldsymbol{\mu}_{z_t}^{(q)},\boldsymbol{R}_{z_t}^{(q)}) \tag{VI.6}$$

Parmi les termes de la somme dans (VI.6), on choisit la configuration $z_{1..T}$ telle que, à l'instant i, $z_i=k_0$ et $\forall t\neq i,\ z_t=l\neq k_0$. La vraisemblance (VI.6) vérifie l'inégalité suivante :

$$\begin{aligned}\mathrm{p}(\boldsymbol{s}_{1..T}\mid\boldsymbol{\theta}^{(q)})\geq c\,|2\,\pi\,\boldsymbol{R}_{k_0}^{(q)}|^{-\frac{1}{2}}\exp\left[-\tfrac{1}{2}\mathrm{Tr}\left({\boldsymbol{R}_{k_0}^{(q)}}^{-1}(\boldsymbol{s}_i-\boldsymbol{\mu}_{k_0}^{(q)})(\boldsymbol{s}_i-\boldsymbol{\mu}_{k_0}^{(q)})^T\right)\right]\\ |2\,\pi\,\boldsymbol{R}_l^{(q)}|^{-\frac{T-1}{2}}\exp\left[-\tfrac{1}{2}\mathrm{Tr}\left({\boldsymbol{R}_l^{(q)}}^{-1}\sum_{t\neq i}(\boldsymbol{s}_t-\boldsymbol{\mu}_l^{(q)})(\boldsymbol{s}_t-\boldsymbol{\mu}_l^{(q)})^T\right)\right]\end{aligned} \tag{VI.7}$$

En revenant aux formes décomposées de $\boldsymbol{R}_{k_0}^{(q)}$ et $\boldsymbol{R}_{l}^{(q)}$ dans leurs bases orthogonales :

$$\begin{cases} \boldsymbol{R}_{k_0}^{(q)} = \boldsymbol{U}_{k_0}^T \, diag\left[\lambda_1^{(q)}, \ldots, \lambda_{n-p_{k_0}}^{(q)}, \lambda_{n-p_{k_0}+1}^{*}, \ldots, \lambda_n^{*}\right] \boldsymbol{U}_{k_0} \\ \boldsymbol{R}_{l}^{(q)} = \boldsymbol{U}_{l}^T \, diag\left[\lambda_1^{(q)}, \ldots, \lambda_{n-p_{l}}^{(q)}, \lambda_{n-p_{l}+1}^{*}, \ldots, \lambda_n^{*}\right] \boldsymbol{U}_{l} \end{cases}$$

et en tenant compte de la propriété (2) de la moyenne $\boldsymbol{\mu}_{k_0}$, l'inégalité (VI.7) se transforme en :

$$\begin{aligned} \mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{\theta}^{(q)}) \geq c \, |2\,\pi\,\boldsymbol{R}_{k_0}^{(q)}|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2} \sum_{j=n-p_{k_0}+1}^{n} \frac{\left[\boldsymbol{U}_{k_0}(\boldsymbol{s}_i - \boldsymbol{\mu}_{k_0}^{(q)})\right]_j^{\,2}}{\lambda_j^{*}}\right] \\ |2\,\pi\,\boldsymbol{R}_{l}^{(q)}|^{-\frac{T-1}{2}} \exp\left[-\tfrac{1}{2} \sum_{j=1}^{n} \frac{\sum_{t\neq i} \left[\boldsymbol{U}_{l}(\boldsymbol{s}_t - \boldsymbol{\mu}_{l}^{(q)})\right]_j^{\,2}}{\gamma_j^{(q)}}\right] \end{aligned} \tag{VI.8}$$

Le point clé de la démonstration est le fait que les suites des valeurs propres $[\lambda_j^{(q)}]_{j=1}^{n-p_{k_0}}$ qui tendent vers 0 ont disparu de l'argument de la fonction exponentielle à cause de la propriété vérifiée par la moyenne $\boldsymbol{\mu}_{k_0}$ mais restent dans le dénominateur du minorant dans (VI.8) à travers le déterminant $|\boldsymbol{R}_{k_0}^{(q)}|^{-\frac{1}{2}}$. Pour conclure à la divergence du minorant, on doit envisager deux cas :

1. $l \in \mathcal{Z}_r$: dans ce cas toutes les valeurs propres $[\gamma_j^{(q)}]_{j=1}^{n}$ tendent vers des valeurs strictement positives. Par conséquent, lorsque les suites $[\lambda_j^{(q)}]_{j=1}^{n-p_{k_0}}$ tendent vers 0 ($\boldsymbol{\theta}^{(q)} \rightarrow \boldsymbol{\theta}^{*}$), le minorant diverge vers l'infini. Ceci quelque soit la vitesse de convergence des $\lambda_j^{(q)}$.

2. $l \in \mathcal{Z}_s$: les valeurs propres $[\gamma_j^{(q)}]_{j=1}^{n-p_l}$ tendent aussi vers 0 et sont présentes dans l'argument de l'exponentielle. Dans ce cas, on doit contrôler les vitesses de convergence relatives des suites λ_j et γ_j. Par exemple, si $\lambda_j^{(q)} = e^{-q}$ et $\gamma_j^{(q)} = 1/\log q$, le minorant dans (VI.8) diverge vers l'infini quand $q \rightarrow \infty$.

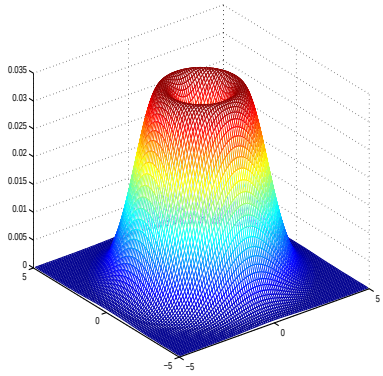
Nous avons ainsi prouvé que chaque point $\boldsymbol{\theta}^{*}$ vérifiant les deux propriétés (1) et (2) est un point de singularité. Autrement dit, on peut trouver une suite de points $\boldsymbol{\theta}^{(q)}$ convergeant vers le point $\boldsymbol{\theta}^{*}$ telle que la vraisemblance de $\boldsymbol{\theta}^{(q)}$ diverge vers l'infini.

Remarque 17 *On note que, dans la démonstration, l'hypothèse de positivité de* $\Pr(z_{1..T} \mid \boldsymbol{\pi}^{*})$ *n'est pas nécessaire. En effet, dans le cas où* $\Pr(z_{1..T} \mid \boldsymbol{\pi}^{(q)})$ *tend vers* 0*, on peut toujours contrôler la convergence de telle façon que le minorant de l'expression (VI.8) diverge.*

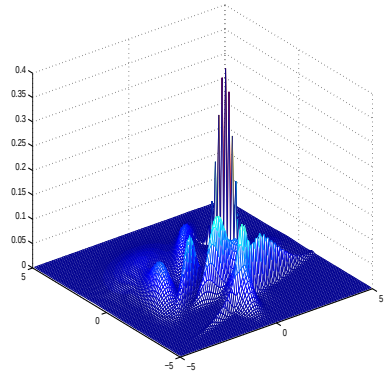
Remarque 18 *En fixant* $n = 1$*, on retrouve le cas scalaire étudié dans [Ridolfi et Idier, 1999]. Dans ce cas, la propriété (1) implique que l'une des variances* σ_k *tend vers* 0 *et la propriété (2) implique que la moyenne* μ_k *de la même composante* k *coïncide avec une observation* s_i*.*

Remarque 19 *On aurait pu caractériser un ensemble de singularité plus restreint qui est similaire au cas scalaire. Ceci en considérant les* $\boldsymbol{\theta}^{*}$ *tel que l'une des matrices* $\boldsymbol{R}_k$ *est nulle et la moyenne* $\boldsymbol{\mu}_k$ *coincide avec un vecteur d'observation* $\boldsymbol{s}_i$*. Nous avons voulu caractériser d'une manière générale l'ensemble des singularités (propriétés (1) et (2)) afin de montrer que le risque de dégénérescence augmente avec la dimension* n *(le nombre de points de singularité est infini dans le cas* $n > 1$*).*

La figure (VI.1) illustre cette dégénérescence. Dans cet exemple de simulation, on a pris une distribution originale (graphe à gauche de la figure (VI.1)) d'un vecteur aléatoire 2-D qui consiste en un mélange de 10 gaussiennes. Les gaussiennes ont la même covariance et des moyennes situées sur un cercle. Le graphe à droite de la figure (VI.1) montre la distribution estimée avec le maximum de vraisemblance. On note la dégénérescence du maximum de vraisemblance qui diverge vers des gaussiennes très piquées.



Distribution originale



Distribution estimée avec maximum de vraisemblance à partir de 100 échantillons

FIG. VI.1: Echec de l'estimation des paramètres d'une distribution mélange de 10 gaussiennes avec la méthode du maximum de vraisemblance.

VI.3 Solution bayésienne

La dégénérescence a été mentionnée par plusieurs auteurs [Kiefer et Wolfowitz, 1956; Day, 1969]. On trouve dans [Hathaway, 1986] une solution à ce problème de dégénérescence dans le cas scalaire. Cette solution consiste à implémenter l'algorithme EM en imposant la contrainte que les variances soient strictement positives $\sigma_z > c > 0,\ z = 1..K$. Le choix du paramètre c et la contrainte de positivité rendent cette solution complexe. Le cas vectoriel va nécessairement accroître cette complexité puisqu'on doit imposer, dans ce cas, la régularité des matrices de covariance. Dans le cas scalaire, une solution bayésienne a été considérée dans [Ridolfi et Idier, 1999]. Elle consiste à pénaliser la vraisemblance avec un *a priori* **inverse gamma**. Dans le cas vectoriel [Ormoneit et Tresp, 1998], l'utilisation d'un *a priori* conjugué (inverse wishart pour les matrices de covariance $\boldsymbol{R}_z$) élimine également cette dégénérescence. Nous allons formuler une solution générale au problème de dégénérescence en se basant sur la caractérisation des points de singularité étudiée dans la section précédente.

Intuitivement, la dégénérescence de la vraisemblance se produit quand l'une des matrices de covariance $\boldsymbol{R}_z$ se rapproche de la frontière de singularité $\mathcal{F}r(\mathcal{S}_+)$. L'approche bayésienne consiste à multiplier la vraisemblance[2] par la distribution *a priori* $\mathrm{p}(\boldsymbol{\theta}) = \prod_{z=1}^{K} \mathrm{p}(\boldsymbol{R}_z)$. On peut donc essayer, tout en restant dans l'espace $\boldsymbol{\Theta}$, de jouer sur la forme de $\mathrm{p}(\boldsymbol{R}_z)$ afin d'éliminer la dégénérescence de la distribution *a posteriori* $\mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{\theta}) \prod_{z=1}^{K} \mathrm{p}(\boldsymbol{R}_z)$.

En étudiant le terme dans l'expression (VI.8) responsable de la dégénérescence, la loi *a priori* $\mathrm{p}(\boldsymbol{R}_z)$ doit vérifier les deux conditions suivantes :

($\boldsymbol{C}$.1) $\lim_{\boldsymbol{R}_z \to \mathcal{F}r(\mathcal{S}_+)} |\boldsymbol{R}_z|^{-N} \mathrm{p}(\boldsymbol{R}_z) = 0$, quelque soit la manière avec laquelle la matrice $\boldsymbol{R}_z$ tend vers la frontière de singularité.

($\boldsymbol{C}$.2) La loi *a priori* $\mathrm{p}(\boldsymbol{R}_z)$ soit bornée.

La première condition ($\boldsymbol{C}$.1) assure que la vraisemblance pénalisée tend vers 0 quand on se rapproche de la frontière de singularité. La deuxième condition ($\boldsymbol{C}$.2) assure que la loi *a priori* ne cause pas à son tour des dégénérescences et que la vraisemblance pénalisée reste bornée sur tout l'espace $\boldsymbol{\Theta}$.

[2] En tant que fonction des moyennnes $\boldsymbol{\mu}_z$, la vraisemblance est bornée. C'est pourquoi , il suffit de considérer un *a priori* sur les matrices de covariance.

Parmi les lois *a priori* vérifiant les deux conditions ($\boldsymbol{C}$.1) et ($\boldsymbol{C}$.2), on propose la distribution **inverse wishart** (wishart pour les matrices $\boldsymbol{R}_z^{-1}$) :

$$\boldsymbol{R}_z \sim \mathcal{IW}_n(\nu_z, \boldsymbol{\Sigma}_z) \propto |\boldsymbol{R}_z^{-1}|^{\frac{\nu_z+(n+1)}{2}} \exp\left[-\frac{1}{2}\nu_z \mathrm{Tr}\left(\boldsymbol{R}_z^{-1}\boldsymbol{\Sigma}_z^{-1}\right)\right]$$

où ν_z est le degré de liberté de la distribution wishart et $\boldsymbol{\Sigma}_z$ est une matrice définie positive.

Proposition 3 $\forall \boldsymbol{s}_{1..T} \in (\mathbb{R}^n)^T$, *la vraisemblance* $\mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{\theta})$ *pénalisée par l'a priori* **inverse wishart** *:*

$$\mathrm{p}(\boldsymbol{\theta}) = \prod_{z=1}^{K} \mathcal{IW}_n(\boldsymbol{R}_z\,;\, \nu_z, \boldsymbol{\Sigma}_z)$$

est bornée sur $\boldsymbol{\Theta}$*. En plus, elle tend vers* 0 *quand l'une des matrices de covariance se rapproche de la frontière de singularité.*

Remarque 20 *Le fait que la distribution a posteriori tend vers* 0 *garantit que les estimateurs* MAP *des matrices de covariance n'appartiennent pas à la frontière de singularité.*

Preuve : la vraisemblance pénalisée s'écrit :

$$\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{s}_{1..T}) \propto \prod_{z=1}^{K} \mathrm{p}(\boldsymbol{R}_z) \sum_{z_{1..T}} \mathrm{p}(z_{1..T} \mid \boldsymbol{\pi}) \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{s}_t\,;\, \boldsymbol{\mu}_{z_t}, \boldsymbol{R}_{z_t}) \tag{VI.9}$$

Pour chaque configuration $z_{1..T}$ (terme de la somme finie dans (VI.9)), on a les inégalités suivantes :

$$\begin{cases} \mathrm{p}(z_{1..T} \mid \boldsymbol{\pi}) \leq 1 \\ \\ \mathcal{N}(\boldsymbol{s}_t; \boldsymbol{\mu}_{z_t}, \boldsymbol{R}_{z_t}) \leq |2\,\pi\,\boldsymbol{R}_{z_t}|^{-\frac{1}{2}} \end{cases}$$

on a donc,

$$\begin{aligned} \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{s}_{1..T}) &\leq \left(\sum_{z=1}^{K} |\boldsymbol{R}_{z_t}|^{-\frac{1}{2}}\right)^T \prod_{z=1}^{K} \mathrm{p}(\boldsymbol{R}_z) \\ &\leq |\boldsymbol{R}_{k_{min}}|^{-\frac{T}{2}} \prod_{z=1}^{K} \mathrm{p}(\boldsymbol{R}_z) \qquad \text{(VI.10)} \\ &\leq |\boldsymbol{R}_{k_{min}}|^{-\frac{T}{2}} \prod_{z=1}^{K} |\boldsymbol{R}_z^{-1}|^{\frac{\nu_z+(n+1)}{2}} \exp\left[-\frac{1}{2}\nu_z \mathrm{Tr}\left(\boldsymbol{R}_z^{-1}\boldsymbol{\Sigma}_z^{-1}\right)\right] \end{aligned}$$

où la classe k_{min} est telle que $|\boldsymbol{R}_{k_{min}}| \leq |\boldsymbol{R}_k|$ pour tout $k \in \{1..K\}$.

Finalement, en utilisant l'inégalité suivante, valable pour toute matrice $\boldsymbol{M}$ symétrique positive ([Gostiaux, 1993b] page 261) :

$$(det\,\boldsymbol{A})^{1/n} \leq \frac{1}{n}\mathrm{Tr}\,(\boldsymbol{A})\,,$$

on arrive à l'inégalité qui va nous permettre de conclure la démonstration :

$$\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{s}_{1..T}) \leq |\boldsymbol{R}_{k_{min}}|^{-\frac{T}{2}} \prod_{z=1}^{K} |\boldsymbol{R}_z^{-1}|^{\frac{\nu_z+(n+1)}{2}} \exp\left[-\frac{n\,\nu_z|\boldsymbol{\Sigma}_z^{-1}|^{1/n}}{2\,|\boldsymbol{R}_z|^{1/n}}\right] \tag{VI.11}$$

On voit clairement que si l'une des matrices $\boldsymbol{R}_z$ approche la frontière $\mathcal{F}r(\mathcal{S}_+)$, alors le déterminant $|\boldsymbol{R}_{k_{min}}|$ tend vers 0 et par conséquent, le majorant dans (VI.11) de la distribution *a posteriori* tend vers 0, ce qui achève la démonstration.

Remarque 21 *En plus du fait qu'elle remplit les conditions* ($\boldsymbol{C}$.1) *et* ($\boldsymbol{C}$.2)*, la distribution a priori inverse wishart présente l'avantage de garder la même structure des équations de ré-estimation de l'algorithme* EM*. Le seul changement concerne la ré-estimation des covariances* $\boldsymbol{R}_z$*. Ainsi, à l'itération* m *de l'*EM *:*

$$\textit{Sans pénalisation} \quad \longrightarrow \quad \boldsymbol{R}_z^{(m)} = \frac{\sum_t (\boldsymbol{s}_t - \boldsymbol{\mu}_z^{(m)})(\boldsymbol{s}_t - \boldsymbol{\mu}_z^{(m)})^* \, \mathrm{p}(z \mid \boldsymbol{s}_t, \boldsymbol{\theta}^{(m-1)})}{\sum_t \mathrm{p}(z \mid \boldsymbol{s}_t, \boldsymbol{\theta}^{(m-1)})}$$

$$\textit{Avec pénalisation} \quad \longrightarrow \quad \boldsymbol{R}_z^{(m)} = \frac{\sum_t (\boldsymbol{s}_t - \boldsymbol{\mu}_z^{(m)})(\boldsymbol{s}_t - \boldsymbol{\mu}_z^{(m)})^* \, \mathrm{p}(z \mid \boldsymbol{s}_t, \boldsymbol{\theta}^{(m-1)}) + \nu_z \boldsymbol{\Sigma}_z^{-1}}{\sum_t \mathrm{p}(z \mid \boldsymbol{s}_t, \boldsymbol{\theta}^{(m-1)}) + (\nu_z + n + 1)}$$

où on remarque que la seule modification consiste à rajouter les termes $\nu_z \boldsymbol{\Sigma}_z^{-1}$ *et* $(\nu_z + n + 1)$ *respectivement dans le numérateur et dans le dénominateur.*

VI.3.1 Existence de la solution

Afin de simplifier les notations, nous allons noter $f(\boldsymbol{\theta})$ la vraisemblance pénalisée. La proposition (2) garantit que la fonction $f(\boldsymbol{\theta})$ est continue (par prolongement) sur l'espace $\boldsymbol{\Theta}$ et qu'elle tend vers 0 quand l'une des matrices $\boldsymbol{R}_z$ approche la frontière de singularité. Supposons que l'ensemble $\mathcal{C}$ des matrices de covariance contient au moins une matrice $\boldsymbol{R}_0$ définie positive. La valeur $f(\boldsymbol{\theta}_0)$ est strictement positive. En supposant que l'espace $\mathcal{C}$ est connexe (donc bien enchaîné), ceci assure qu'une recherche continue du maximum de la fonction f ne va pas traverser la frontière de singularité et que le maximum (s'il existe) n'appartient pas à cette frontière.

Afin de prouver l'existence d'au moins un maximum, le fait que f soit bornée sur la frontière n'est pas suffisant. Nous allons d'abord munir l'espace des matrices de covariance $(n \times n)$ d'une métrique en le considérant comme un espace vectoriel normé de dimension n^2. Le produit scalaire entre deux matrices $\boldsymbol{M}$ et $\boldsymbol{N}$ est défini par :

$$< \boldsymbol{M}, \boldsymbol{N} >= \mathrm{Tr}\left(\boldsymbol{M}^T \boldsymbol{N}\right),$$

et donc la norme est définie par :

$$||\boldsymbol{M}||^2 = \sum_{i,j} M_{ij}^2.$$

L'espace $\mathcal{C}$ est un sous-espace **fermé** de $\mathcal{S}_+$ mais il n'est pas nécessairement borné et donc il n'est pas **compact**. On ne peut pas alors appliquer le théorème qui précise qu'une fonction continue sur un compact est bornée et atteint ses bornes. Cependant, nous allons montrer dans la suite que la fonction f est bornée et atteint son maximum sur $\boldsymbol{\Theta}$.

Supposons qu'il existe au moins K matrices $\boldsymbol{R}_z^0 \in \mathcal{C}$ définies positives, $f(\boldsymbol{\theta}_0)$ a une valeur finie strictement positive. Soit b un réel strictement positif tel que :

$$\forall z \in \mathcal{Z}, ||\boldsymbol{R}_z^0||^2 \leq b.$$

On définit l'ensemble $\mathcal{B}_b$ des matrices de norme inférieure ou égale à b. $\mathcal{B}_b$ est fermé borné. Les matrices $\boldsymbol{R}_z^0$ appartiennent alors à $\mathcal{C} \cap \mathcal{B}_b$. L'espace $\mathcal{C} \cap \mathcal{B}_b$ est aussi fermé borné donc compact. La fonction f atteint sur ce compact son maximum. Nous allons montrer maintenant que lorsque b tend vers l'infini, la fonction $f(\boldsymbol{\theta})$ tend vers 0 pour les $\boldsymbol{\theta}$ tels que $||\boldsymbol{R}_z||^2 > b$ ($\boldsymbol{R}_z$ à l'extérieur de la boule $\mathcal{B}_b$).

En reprenant l'inégalité (VI.10), on montre que :

$$\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{s}_{1..T}) \leq \sum_{z_{1..T}} \mathcal{Q}_{z_{1..T}} \tag{VI.12}$$

où le terme $\mathcal{Q}_{z_{1..T}}$ est défini pour chaque configuration $z_{1..T}$ par :

$$\mathcal{Q}_{z_{1..T}} = \prod_{z=1}^{K} \frac{1}{|\boldsymbol{R}_z|^{\frac{\nu_z+(n+1)+|\mathcal{T}_z|}{2}}} \exp\left[-\frac{1}{2}\nu_z \mathrm{Tr}\left(\boldsymbol{R}_z^{-1}\boldsymbol{\Sigma}_z^{-1}\right)\right] \tag{VI.13}$$

$\mathcal{T}_z = \{t \mid z_t = z\}$ étant la répartition des indices temporels selon la configuration $z_{1..T}$.

En utilisant le fait que la norme d'une matrice positive est égale à la somme des carrés de ses valeurs propres (voir [Gostiaux, 1993b] page 261) :

$$||\boldsymbol{M}||^2 = \sum_{j=1}^{n} \lambda_j^2,$$

on aboutit aux inégalités suivantes :

$$b \,<\, ||\boldsymbol{R}_z||^2 = \sum_{j=1}^{n} \lambda_{zj}^2 \,\leq\, n\,\lambda_{z,\,max}^2$$

Ainsi, lorsque b tend vers l'infini, les valeurs propres maximales des matrices $\boldsymbol{R}_z$ tendent vers l'infini. Par conséquent, en utilisant la majoration (VI.12) et l'expression (VI.13), il est facile de prouver que :

$$\lim_{b\to\infty} f(\boldsymbol{\theta}) = 0, \textit{ pour } \boldsymbol{\theta} \text{ tel que } ||\boldsymbol{R}_z||^2 > b.$$

Cette limite veut dire que $\forall\, \epsilon > 0$, il existe un rayon b_ϵ tel que si l'une des matrices $\boldsymbol{R}_z$ se trouve à l'extérieur de la boule $\mathcal{B}_{b_\epsilon}$ alors la vraisemblance pénalisée est inférieure à ϵ. En prenant $0 < \epsilon < f(\boldsymbol{\theta}_0)$, on aura $f(\boldsymbol{\theta}) \leq f(\boldsymbol{\theta}_0)$ pour tous les points $\boldsymbol{\theta}$ à l'extérieur de $\mathcal{C} \cap \mathcal{B}_{b_\epsilon}$. L'ensemble $\mathcal{C} \cap \mathcal{B}_{b_\epsilon}$ est compact, donc f atteint son maximum au point $\hat{\boldsymbol{\theta}} \in \mathcal{C} \cap \mathcal{B}_{b_\epsilon}$. En particulier $f(\hat{\boldsymbol{\theta}}) \geq f(\boldsymbol{\theta}_0)$, et donc supérieure à tous les points à l'extérieur de $\mathcal{C} \cap \mathcal{B}_{b_\epsilon}$. On a ainsi prouvé l'existence d'un maximum global pour la vraisemblance pénalisée. La figure (VI.2) illustre cette démonstration.

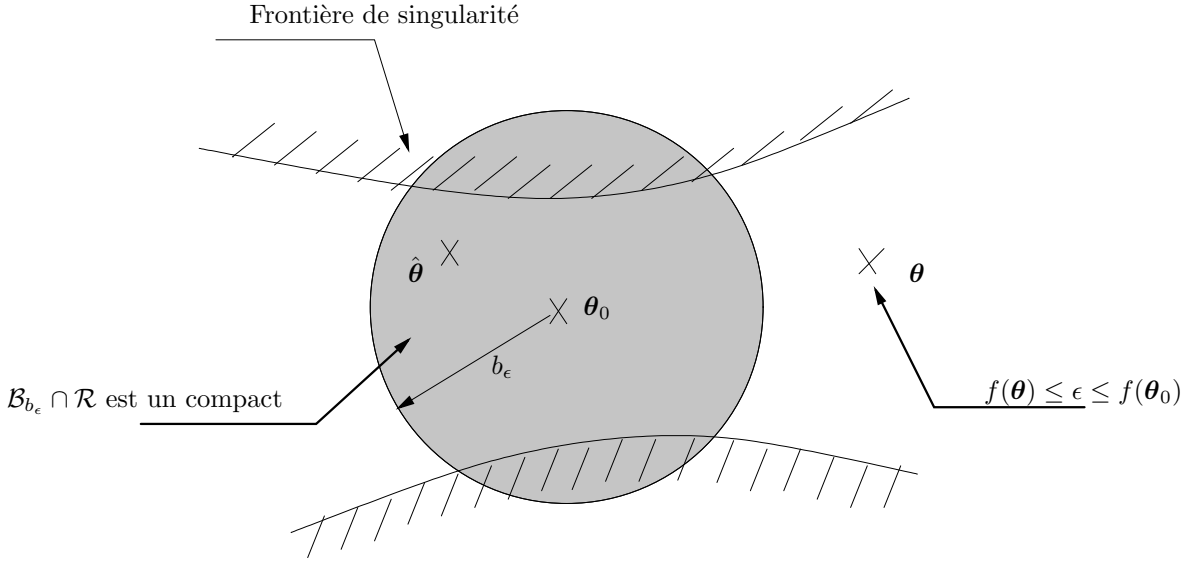
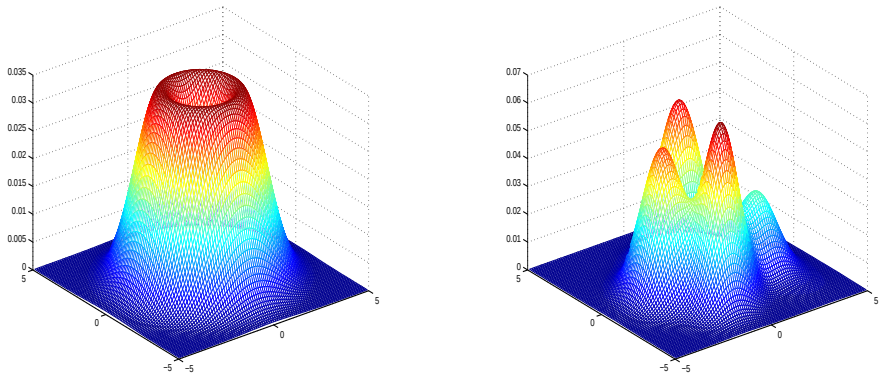


FIG. VI.2: Illustration de la preuve d'existence d'un maximum globale pour la vraisembance pénalisée.

La figure (VI.3) illustre l'effet de la régularisation apporté par la pénalisation de la vraisemblance. Nous avons utilisé les mêmes conditions de simulation que dans l'exemple de la figure (VI.1). Avec la pénalisation, le risque de dégénérescence est nul. Ce que nous avons noté dans les simulations.



Distribution originale.

Distribution estimée avec l'algorithme EM pénalisé.

FIG. VI.3: Effet de la régularisation apporté par la pénalisation de la vraisemblance.

VI.4 Estimation des matrices de covariance structurées

Dans ce paragraphe, nous allons voir comment on peut tenir compte de certaines contraintes sur la structure des matrices de covariance $\boldsymbol{R}_z$. Autrement dit, l'espace $\mathcal{C}$ auquel doit appartenir les matrices $\boldsymbol{R}_z$ n'est pas tout l'espace $\mathcal{S}_+$ des matrices positives mais il y est strictement inclu : $\mathcal{C} = \mathcal{S}_+ \cup \mathcal{V} \subsetneq \mathcal{S}_+$, où $\mathcal{V}$ caractérise la structure imposée aux matrices de covariance. Dans le cas où la structure imposée vérifie la propriété suivante :

Propriété 3 *Si* $\boldsymbol{R} \in \mathcal{C}$ *alors* $\delta\boldsymbol{R} \in \mathcal{V}$

Autrement dit, la variation de la matrice $\boldsymbol{R}$ définie par :

$$\delta\boldsymbol{R} = \begin{pmatrix} \delta R(1,1) & \delta R(1,2) & \ddots & \delta R(1,n-1) & \delta R(1,n) \\ \delta R(2,1) & \delta R(2,2) & \ddots & \delta R(2,n-1) & \delta R(2,n) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \delta R(n,1) & \delta R(n,2) & \ddots & \delta R(n,n-1) & \delta R(n,n) \end{pmatrix}$$

garde la même structure que la matrice $\boldsymbol{R} \in \mathcal{C}$. On donne deux exemples de telles matrices :

Exemple 5 *Si on suppose que le vecteur* $\boldsymbol{s}_t$ *représente une série stationnaire alors les matrices de covariances sont* **Tœplitz** *et vérifient la propriété (3).*

Exemple 6 *Dans certaines applications, on connait les structures des matrices* $\boldsymbol{R}_z$ *mais à des cœfficients multiplicatifs près :*

$$\boldsymbol{R}_z = \alpha_z \boldsymbol{\Sigma}_z, \ z = 1..K,$$

où les matrices $\boldsymbol{\Sigma}_z$ *(par exemple les spectres de processus stationnaires) sont connues mais les cœfficients* α_z *(les puissances des spectres) ne sont pas connus. Ces matrices* $\boldsymbol{R}_z$ *vérifient la propriété (3).*

Exemple 7 *Les variétés linéaires :*

$$\mathcal{C} = \{\boldsymbol{R} = \sum_{l=1}^{L} x_l \boldsymbol{Q}_l \mid \boldsymbol{x} \in \mathbb{R}^L\}$$

où $\{\boldsymbol{Q}_l\}_{l=1}^L$ *est la base de* $\mathcal{C}$*, est un espace fermé vérifiant la propriété (3). D'ailleurs, les deux exemples précédents en sont des cas particuliers.*

Remarque 22 *On note que ce type de contrainte de structure n'est pas de même nature que la contrainte de régularité étudiée dans la section précédente. Contraindre algorithmiquement les matrices à être régulières n'est pas facile. On a vu que la pénalisation est une solution efficace qui nous permet d'éviter cette difficulté.*

Dans la suite, nous allons généraliser le travail de [Burg, 1982] [3] (où les auteurs estiment une matrice de covariance structurée d'un processus gaussien) au cas de mélange de gaussiennes multivariables en proposant un algorithme EM "renversé".

Commençons par rappeler l'algorithme EM standard pour l'estimation des paramètres d'un mélange de gaussiennes. C'est un algorithme itératif qui, partant d'un point initial $\boldsymbol{\theta}^{(0)}$, transforme $\boldsymbol{\theta}^{(k)}$ en $\boldsymbol{\theta}^{(k+1)}$ selon les deux étapes suivantes :

Etape-E : En considérant les observations $\boldsymbol{s}_{1..T}$ comme des données incomplètes et les étiquettes $z_{1..T}$ comme les données manquantes, on calcule la fonctionnelle $\mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$:

$$\mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) = \mathrm{E}\big[\log p(\boldsymbol{s}_{1..T}, z_{1..T} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \mid \boldsymbol{s}_{1..T}, \boldsymbol{\theta}^{(k)}\big].$$

Etape-M : Remettre à jour $\boldsymbol{\theta}^{(k+1)}$ en maximisant la fonctionnelle $\mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$:

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$$

Les expressions des $\boldsymbol{\mu}_z^{(k+1)}$ et de $\boldsymbol{\pi}^{(k+1)}$ sont simples à dériver. On s'intéresse plus particulièrement à la maximisation de la fonctionnelle $\mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ par rapport aux matrices de covariance $\boldsymbol{R}_z$ sous la contrainte qu'elles appartiennent à l'ensemble $\mathcal{C}$. Nous avons alors le problème suivant :

$$\textit{maximisation de } \mathcal{Q}_r(. \mid \boldsymbol{\theta}^{(k)}) = \sum_{z=1}^{K} g(\boldsymbol{R}_z, \boldsymbol{S}_z) \;\; \textit{s.c.} \;\; \{\boldsymbol{R}_z \in \mathcal{C},\, z = 1..K\}.$$

L'optimisation par rapport aux matrices $\boldsymbol{R}_z$ est alors découplée. Les fonctions $g(\boldsymbol{R}_z, \boldsymbol{S}_z)$ sont définies par :

$$g(\boldsymbol{R}_z, \boldsymbol{S}_z) = -\left(1 + \frac{\nu_z + n + 1}{N_z}\right) \log|\boldsymbol{R}_z| - \mathrm{Tr}\left(\boldsymbol{R}_z^{-1}\left(\boldsymbol{S}_z + \frac{\nu_z}{N_z}\boldsymbol{\Sigma}_z^{-1}\right)\right),$$

où $\boldsymbol{S}_z$ est la matrice de covariance empirique moyennée et N_z est le nombre moyen des étiquettes de classe z :

$$\begin{cases} \boldsymbol{S}_z = \dfrac{\displaystyle\sum_{t=1}^{T} (\boldsymbol{s}_t - \boldsymbol{\mu}_z^{(k+1)})\,(\boldsymbol{s}_t - \boldsymbol{\mu}_z^{(k+1)})^T\, p(z \mid \boldsymbol{s}_t,\, \boldsymbol{\theta}^{(k)})}{\displaystyle\sum_{t=1}^{T} p(z \mid \boldsymbol{s}_t,\, \boldsymbol{\theta}^{(k)})} \\ \\ N_z = \displaystyle\sum_{t=1}^{T} p(z \mid \boldsymbol{s}_t,\, \boldsymbol{\theta}^{(k)}) \end{cases}$$

Les matrices $\boldsymbol{R}_z$ doivent vérifier les conditions de gradient suivantes :

$$\begin{cases} \delta g(\boldsymbol{R}_z, \boldsymbol{S}_z) = \mathrm{Tr}\left(\left\{\boldsymbol{R}_z^{-1}(\boldsymbol{S}_z + \frac{\nu_z}{N_z}\boldsymbol{\Sigma}_z^{-1})\boldsymbol{R}_z^{-1} - (1 + \frac{\nu_z+n+1}{N_z})\boldsymbol{R}_z^{-1}\right\}\delta\boldsymbol{R}_z\right) = 0, \\ \\ z = 1..K. \end{cases} \tag{VI.14}$$

Les contraintes de structure sont exprimées à travers le terme $\delta\boldsymbol{R}_z$. Tous les déplacements ne sont pas autorisés et doivent être conformes aux contraintes de structure.

[3] Dans [Burg, 1982], on trouve d'autres exemples de contraintes de structure.

VI.4.1 Cas sans contraintes de structure

Les variations des matrices $\boldsymbol{R}_z$ sont quelconques et donc les gradients $\frac{\partial g(\boldsymbol{R}_z,\boldsymbol{S}_z)}{\partial \boldsymbol{R}_z}$ sont identiquement nuls :

$$\frac{\partial g(\boldsymbol{R}_z,\boldsymbol{S}_z)}{\partial \boldsymbol{R}_z}=0 \Longrightarrow \boldsymbol{R}_z^{(k+1)}=\frac{\boldsymbol{S}_z+\frac{\nu_z}{N_z}\boldsymbol{\Sigma}_z^{-1}}{1+\frac{\nu_z+n+1}{N_z}} \tag{VI.15}$$

On retrouve ainsi les équations de ré-estimation standards de l'algorithme EM. On note que la pénalisation de la vraisemblance par l'*a priori* inverse wishart garantit que l'estimée de $\boldsymbol{R}_z$ est définie positive grâce à la présence du terme $\frac{\nu_z}{N_z}\boldsymbol{\Sigma}_z^{-1}$ dans le numérateur de l'expression de $\boldsymbol{R}_z^{(k+1)}$ (VI.15).

VI.4.2 Cas avec contraintes de structure

La résolution des équations de gradient (VI.14) en imposant des contraintes de structure est compliquée à cause de la présence des termes en $\boldsymbol{R}_z^{-1}$. Nous allons proposer un algorithme EM renversé en généralisant l'algorithme "Inverse Iteration Algorithm" de [Burg, 1982] au cas de mélange gaussiennes. L'idée principale consiste à résoudre les équations du gradient $\delta g(\boldsymbol{R}_z,\boldsymbol{S}_z-\boldsymbol{D}_z)$ en $\boldsymbol{D}_z$ et non plus en $\boldsymbol{R}_z$. En effet si on regarde l'expression (VI.14) de $\delta g(\boldsymbol{R}_z,\boldsymbol{S}_z)$, on s'aperçoit qu'elle est non linéaire en fonction de $\boldsymbol{R}_z$ mais qu'elle est linéaire en fonction de $\boldsymbol{S}_z$. Il est alors plus facile d'imposer la contrainte de structure sur la matrice $\boldsymbol{S}_z$. A chaque itération k de l'algorithme EM, les matrices de covariance sont calculées de la manière suivante :

1. `Trouver` $\boldsymbol{D}_z \in \mathcal{C}$ `telle que` $g(\boldsymbol{R}_z^{(k)},\boldsymbol{S}_z-\boldsymbol{D}_z)$ `satisfait les conditions du gradient :` $\delta g(\boldsymbol{R}_z^{(k)},\boldsymbol{S}_z-\boldsymbol{D}_z)=0$.
2. $\boldsymbol{R}_z^{(k)} \longleftarrow \boldsymbol{R}_z^{(k)}+\boldsymbol{D}_z$

Remarque 23 *Nous avons attribué l'adjectif "renversé" en suivant les arguments de [Burg, 1982]. En effet, la recherche de l'incrément $\boldsymbol{D}_z$, de telle façon que la fonction $g(\boldsymbol{R}_z,\boldsymbol{S}_z-\boldsymbol{D}_z)$ satisfait les conditions du gradient, peut s'interpréter comme une modification virtuelle de la matrice empirique $\boldsymbol{S}_z$ en fixant la matrice $\boldsymbol{R}_z$. Ensuite, la matrice $\boldsymbol{R}_z$ subit le déplacement inverse, d'où l'adjectif "renversé".*

Nous allons montrer dans la suite que cette modification de l'algorithme EM n'affecte pas sa monotonie et que le calcul de $\boldsymbol{D}_z$ consiste à résoudre un système linéaire.

[A] Monotonie de l'EM renversé

On veut prouver que $\boldsymbol{D}_z$ est une direction qui garantit la croissance de la vraisemblance pénalisée. Pour ceci, il suffit de montrer que le produit scalaire du gradient de la fonctionnelle $\mathcal{Q}(\boldsymbol{\theta}\mid\boldsymbol{\theta}^{(k)})$ et de la direction $\boldsymbol{D}_z$ (solution de $\delta g(\boldsymbol{R}_z^{(k)},\boldsymbol{S}_z-\boldsymbol{D}_z)=0$) est positif :

$$\mathrm{Tr}\left(\left\{\boldsymbol{R}_z^{-1}(\boldsymbol{S}_z+\frac{\nu_z}{N_z}\boldsymbol{\Sigma}_z^{-1})\boldsymbol{R}_z^{-1}-(1+\frac{\nu_z+n+1}{N_z})\boldsymbol{R}_z^{-1}\right\}\boldsymbol{D}_z\right)>0. \tag{VI.16}$$

Le terme à gauche peut se décomposer en :

$$\mathrm{Tr}\left(\left\{\boldsymbol{R}_z^{-1}(\boldsymbol{S}_z-\boldsymbol{D}_z+\frac{\nu_z}{N_z}\boldsymbol{\Sigma}_z^{-1})\boldsymbol{R}_z^{-1}-(1+\frac{\nu_z+n+1}{N_z})\boldsymbol{R}_z^{-1}\right\}\boldsymbol{D}_z\right)+\mathrm{Tr}\left(\boldsymbol{R}_z^{-1}\boldsymbol{D}_z\boldsymbol{R}_z^{-1}\boldsymbol{D}_z\right),$$

où le premier terme est nul par construction de $\boldsymbol{D}_z$. Le deuxième terme s'écrit, en considérant la décomposition de Cholesky de la matrice $\boldsymbol{R}_z^{-1}=\boldsymbol{G}\boldsymbol{G}^T$,

$$\begin{aligned}\mathrm{Tr}\left(\boldsymbol{R}_z^{-1}\boldsymbol{D}_z\boldsymbol{R}_z^{-1}\boldsymbol{D}_z\right) &= \mathrm{Tr}\left(\boldsymbol{G}\boldsymbol{G}^T\boldsymbol{D}_z\boldsymbol{G}\boldsymbol{G}^T\boldsymbol{D}_z\right)\\ &= \mathrm{Tr}\left(\boldsymbol{G}^T\boldsymbol{D}_z\boldsymbol{G}\boldsymbol{G}^T\boldsymbol{D}_z\boldsymbol{G}\right)\\ &= ||\boldsymbol{G}^T\boldsymbol{D}_z\boldsymbol{G}||^2.\end{aligned}$$

Si la matrice $\boldsymbol{D}_z$ est non nulle, ce terme est strictement positive, ce qui achève la démonstration de l'inégalité (VI.16).

D'après les propriétés de l'algorithme EM (rappelées dans le chapitre (II)), la croissance de la fonctionnelle $\mathcal{Q}(. \mid \boldsymbol{\theta}^{(k)})$ implique la croissance de la vraisemblance pénalisée incomplète $\mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta})$:

$$\mathcal{Q}(\boldsymbol{\theta}^{(k+1)} \mid \boldsymbol{\theta}^{(k)}) \geq \mathcal{Q}(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}^{(k)}) \Longrightarrow \mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{\theta}^{(k+1)})\mathrm{p}(\boldsymbol{\theta}^{(k+1)}) \geq \mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{\theta}^{(k)})\mathrm{p}(\boldsymbol{\theta}^{(k)})$$

La matrice $\boldsymbol{D}_z$ est alors une direction assurant la croissance de la vraisemblance pénalisée.

[B] Calcul de $\boldsymbol{D}_z$

A l'itération k de l'EM renversé, l'incrément $\boldsymbol{D}_z$ doit vérifier l'équation de gradient suivante :

$$\begin{cases} \mathrm{Tr}\left(\left\{ {\boldsymbol{R}_z^{(k)}}^{-1}(\boldsymbol{S}_z - \boldsymbol{D}_z + \frac{\nu_z}{N_z}\boldsymbol{\Sigma}_z^{-1}){\boldsymbol{R}_z^{(k)}}^{-1} - (1 + \frac{\nu_z+n+1}{N_z}){\boldsymbol{R}_z^{(k)}}^{-1} \right\} \boldsymbol{Q}\right) = 0, \\ \forall \boldsymbol{Q} \in \mathcal{C}. \end{cases} \tag{VI.17}$$

Au lieu de chercher $\boldsymbol{D}_z$, on peut chercher directement la matrice $\boldsymbol{R}_z^{'} = \boldsymbol{R}_z^{(k)} + \boldsymbol{D}_z$. L'équation précédente devient :

$$\begin{cases} \mathrm{Tr}\left(\left\{ {\boldsymbol{R}_z^{(k)}}^{-1}(\boldsymbol{S}_z + \frac{\nu_z}{N_z}\boldsymbol{\Sigma}_z^{-1}){\boldsymbol{R}_z^{(k)}}^{-1} - (1 + \frac{\nu_z+n+1}{N_z}){\boldsymbol{R}_z^{(k)}}^{-1}\boldsymbol{R}_z^{'}{\boldsymbol{R}_z^{(k)}}^{-1} \right\} \boldsymbol{Q}\right) = 0, \\ \forall \boldsymbol{Q} \in \mathcal{C}. \end{cases} \tag{VI.18}$$

En prenant une base $\{\boldsymbol{Q}_l\}_{l=1}^{L}$ de l'espace $\mathcal{C}$, ceci revient à trouver le vecteur $\boldsymbol{x}$ tel que la matrice $\boldsymbol{R}_z^{'} = \sum x_l \boldsymbol{Q}_l$ vérifie l'équation (VI.18) pour les L matrices $\boldsymbol{Q}_l$. Nous avons alors un système linéaire à L équations et L inconnues :

$$\begin{cases} (1 + \frac{\nu_z+n+1}{N_z}) \sum_{l=1}^{L} x_l \mathrm{Tr}\left({\boldsymbol{R}_z^{(k)}}^{-1}\boldsymbol{Q}_l{\boldsymbol{R}_z^{(k)}}^{-1}\boldsymbol{Q}_j\right) = \mathrm{Tr}\left({\boldsymbol{R}_z^{(k)}}^{-1}(\boldsymbol{S}_z + \frac{\nu_z}{N_z}\boldsymbol{\Sigma}_z^{-1}){\boldsymbol{R}_z^{(k)}}^{-1}\boldsymbol{Q}_j\right) \\ j = 1..L. \end{cases}$$

On peut mettre ce système sous forme matricielle :

$$\boldsymbol{M}\boldsymbol{x} = \boldsymbol{b}, \tag{VI.19}$$

en définissant la matrice $\boldsymbol{M}$ par :

$$M_{jl} = \mathrm{Tr}\left({\boldsymbol{R}_z^{(k)}}^{-1}\boldsymbol{Q}_l{\boldsymbol{R}_z^{(k)}}^{-1}\boldsymbol{Q}_j\right), \tag{VI.20}$$

et le vecteur $\boldsymbol{b}$ par :

$$b_j = \frac{N_z}{N_z + \nu_z + n + 1}\mathrm{Tr}\left({\boldsymbol{R}_z^{(k)}}^{-1}(\boldsymbol{S}_z + \frac{\nu_z}{N_z}\boldsymbol{\Sigma}_z^{-1}){\boldsymbol{R}_z^{(k)}}^{-1}\boldsymbol{Q}_j\right).$$

Maintenant, il faut vérifier que la matrice $\boldsymbol{M}$ est définie positive[4] pour que l'équation linéaire (VI.19) ait une solution. Autrement dit, il faut vérifier que :

$$\forall\ \boldsymbol{v} \neq 0, \sum_{j,l} v_j M_{jl} v_l > 0.$$

[4]En fait, il faut vérifier que la matrice $\boldsymbol{M}$ est régulière mais comme elle est symétrique, il suffit de vérifier qu'elle est définie.

En utilisant les expressions (VI.20) des éléments de la matrice $\boldsymbol{M}$,

$$\begin{aligned}
\sum_{j,l} v_j M_{jl} v_l &= \sum_{j,l} \mathrm{Tr}\left(\boldsymbol{R}_z^{(k)^{-1}} v_l \boldsymbol{Q}_l \boldsymbol{R}_z^{(k)^{-1}} v_j \boldsymbol{Q}_j\right) \\
&= \mathrm{Tr}\left(\boldsymbol{R}_z^{(k)^{-1}} \boldsymbol{B} \boldsymbol{R}_z^{(k)^{-1}} \boldsymbol{B}\right), \;\; \boldsymbol{B} = \sum_j v_j \boldsymbol{Q}_j, \\
&= \mathrm{Tr}\left(\boldsymbol{G}^T \boldsymbol{B} \boldsymbol{G} \boldsymbol{G}^T \boldsymbol{B} \boldsymbol{G}\right), \;\; (\boldsymbol{R}_z^{(k)^{-1}} = \boldsymbol{G}\boldsymbol{G}^T) \\
&= ||\boldsymbol{G}^T \boldsymbol{B} \boldsymbol{G}||^2.
\end{aligned}$$

Puisque la matrice $\boldsymbol{B}$ est non nulle ($\boldsymbol{v} \neq 0$), la norme au carré $||\boldsymbol{G}^T\boldsymbol{B}\boldsymbol{G}||^2$ est strictement positive. Ce qui prouve que la matrice $\boldsymbol{M}$ est définie positive et donc que l'équation (VI.17) admet une solution unique $\hat{\boldsymbol{D}}_z = \sum_l \hat{x}_l \boldsymbol{Q}_l$ avec $\hat{\boldsymbol{x}} = \boldsymbol{M}^{-1}\boldsymbol{b}$.

VI.5 Sources mélangées

Dans cette section, on considère le cas où les sources $\boldsymbol{s}_{1..T}$ ne sont pas directement observées. On suppose qu'elles ont subi une transformation linéaire (avec une matrice inconnue $\boldsymbol{A}$) en plus d'un bruit blanc additif gaussien de covariance inconnue $\boldsymbol{R}_\epsilon$. Les observations effectives sont représentées par les signaux vectoriels $\boldsymbol{x}_{1..T}$. A chaque instant t, cette transformation est modélisée par la relation matricielle suivante :

$$\boldsymbol{x}_t = \boldsymbol{A}\,\boldsymbol{s}_t + \boldsymbol{\epsilon}_t, \;\; t = 1..T,$$

où $\boldsymbol{x}_t$ est le vecteur $(m \times 1)$ des observations, $\boldsymbol{s}_t$ le vecteur $(n \times 1)$ des sources, $\boldsymbol{A}$ est la matrice de mélange $(m \times n)$ et $\boldsymbol{\epsilon}_t$ est le bruit blanc de covariance $\boldsymbol{R}_\epsilon$.

Si les sources sont modélisées par des mélanges de gaussiennes, le vecteur total $\boldsymbol{\gamma}$ des paramètres inconnus à identifier contient désormais la matrice de mélange, la covariance du bruit et les moyennes et covariances *a priori* des sources. L'ensemble Γ des paramètres devient :

$$\Gamma = \{\boldsymbol{A} \in \mathcal{M}_{m,n}, \boldsymbol{R}_\epsilon \in \mathcal{S}_+, \boldsymbol{\mu}_z \in \mathbb{R}^n, \boldsymbol{R}_z \in \mathcal{S}_+, \; z = 1..K\}.$$

Nous allons montrer que l'identification du paramètre $\boldsymbol{\gamma}$ à partir des observations $\boldsymbol{x}_{1..T}$ par la méthode du maximum de vraisemblance souffre des mêmes problèmes de dégénérescence étudiés dans les sections précédentes. En effet, la vraisemblance de $\boldsymbol{\gamma}$ s'écrit :

$$\begin{aligned}
\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\gamma}) &= \int_{\boldsymbol{s}_{1..T}} \mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{s}_{1..T}, \boldsymbol{A}, \boldsymbol{R}_\epsilon)\, \mathrm{p}(\boldsymbol{s}_{1..T} \mid \{\boldsymbol{\mu}_z, \boldsymbol{R}_z\}_{z=1}^{z=K}) d\,\boldsymbol{s}_{1..T} \\
&= \sum_{z_{1..T}} \mathrm{p}(z_{1..T}) \left[\prod_{t=1}^{T} \int_{\boldsymbol{s}_t} \mathrm{p}(\boldsymbol{x}_t \mid \boldsymbol{s}_t, \boldsymbol{A}, \boldsymbol{R}_\epsilon)\, \mathrm{p}(\boldsymbol{s}_t \mid \boldsymbol{\mu}_z, \boldsymbol{R}_z) d\,\boldsymbol{s}_t\right] \qquad \text{(VI.21)} \\
&= \sum_{z_{1..T}} \mathrm{p}(z_{1..T}) \left[\prod_{t=1}^{T} \mathcal{N}(\boldsymbol{x}_t \,;\, \boldsymbol{A}\boldsymbol{\mu}_z, \boldsymbol{A}\boldsymbol{R}_z\boldsymbol{A}^T + \boldsymbol{R}_\epsilon)\right]
\end{aligned}$$

On note, sous cette forme, que la distribution des observations $\boldsymbol{x}_{1..T}$ est un mélange de gaussiennes multivariables. La loi des étiquettes $z_{1..T}$ de ce mélange est la même que celle des étiquettes des sources $\boldsymbol{s}_{1..T}$. Les observations appartenant à la classe z suivent une gaussienne de moyenne $\boldsymbol{m}_z = \boldsymbol{A}\boldsymbol{\mu}_z$ et de covariance $\boldsymbol{P}_z = \boldsymbol{A}\boldsymbol{R}_z\boldsymbol{A}^T + \boldsymbol{R}_\epsilon$. Les moyennes et les covariances $(\boldsymbol{m}_z, \boldsymbol{P}_z)$ possèdent ainsi une structure particulière et ne varient pas d'une façon indépendante comme c'était le cas pour les moyennes et les covariances $(\boldsymbol{\mu}_z, \boldsymbol{R}_z)$ des

sources. La multiplication par la matrice $\boldsymbol{A}$ et l'ajout du bruit sont à l'origine de cette dépendance algébrique. Cependant, malgré cette dépendance, nous allons prouver que les mêmes risques de dégénérescence existent dans ce cas. Intuitivement, en suivant les arguments avancés dans le cas sans mélange, ceci revient à justifier que bien que les matrices $\boldsymbol{P}_z$ soient liées, on peut soit les partager en matrices régulières et matrices singulières soit contrôler les vitesses relatives de leurs convergences vers la frontière de singularité. Ces situations ne se produisent qu'à cause de la diversité des matrices $\boldsymbol{R}_z$. En effet, si la singularité provient uniquement de la covariance $\boldsymbol{R}_\epsilon$ ou de la matrice $\boldsymbol{A}$, elle aura un effet commun sur toutes les matrices $\boldsymbol{P}_z$ et la dégénérescence ne se produit pas.

On rappelle que les matrices de covariance $\boldsymbol{R}_z$ et $\boldsymbol{R}_\epsilon$ appartiennent à l'espace fermé $\mathcal{S}_+$ des matrices symétriques positives et les contraindre à être définies complique la mise en œuvre de l'identification. Cette difficulté est due au fait que l'ensemble des matrices définies positives est un ouvert. Pour la même raison, on n'impose pas que la matrice $\boldsymbol{A}$ soit de rang plein.

On commence par s'assurer que parmi les matrices $\boldsymbol{P}_z = \boldsymbol{A}\boldsymbol{R}_z\boldsymbol{A}^T + \boldsymbol{R}_\epsilon$, il peut y exister au moins une matrice singulière et au moins une matrice régulière.

Proposition 4 *$\forall$ $\boldsymbol{A}$ non nulle, il existe des matrices $\{\boldsymbol{P}_z = \boldsymbol{A}\boldsymbol{R}_z\boldsymbol{A}^T + \boldsymbol{R}_\epsilon, z = 1..K\}$ telles que $\{z \mid \boldsymbol{P}_z$ est singulière $\} \neq \emptyset$ et $\{z \mid \boldsymbol{P}_z$ est régulière $\} \neq \emptyset$.*
$\boldsymbol{R}_\epsilon$ est nécessairement une matrice positive singulière et au moins une des matrices $\boldsymbol{R}_z$ est singulière.

Preuve : Sans affecter la généralité de la démonstration, on va construire les matrices $\boldsymbol{P}_z$ de telle façon que la matrice $\boldsymbol{P}_1$ soit singulière et que les autres matrices $\{\boldsymbol{P}_z\}_{z=2}^K$ soient régulières.

On note $\mathcal{K}er(\boldsymbol{M}) = \{\boldsymbol{v} \mid \boldsymbol{M}\boldsymbol{v} = 0\}$ le noyau d'une matrice $\boldsymbol{M}$ et $\mathcal{C}(\boldsymbol{M}) = \{\boldsymbol{v} \mid \boldsymbol{v}^T\boldsymbol{M}\boldsymbol{v} = 0\}$ son cône d'isotropie. Nous allons utiliser une propriété importante des matrices réelles symétriques ([Gostiaux, 1993a], pp 418) :

Propriété 4 *Si $\boldsymbol{M}$ est symétrique positive alors son noyau est égal à son cône d'isotropie.*

En utilisant cette propriété, les noyaux des matrices $\boldsymbol{P}_z$ vérifient la relation suivante :

$$\mathcal{K}er(\boldsymbol{P}_z) = \mathcal{K}er(\boldsymbol{A}\boldsymbol{R}_z\boldsymbol{A}^T) \cap \mathcal{K}er(\boldsymbol{R}_\epsilon)$$

Pour que la matrice $\boldsymbol{P}_1$ soit singulière et que les autres $\boldsymbol{P}_z$ soient régulières, les paramètres $\boldsymbol{\gamma}$ doivent vérifier les relations suivantes :

$$\begin{cases} \mathcal{K}er(\boldsymbol{A}\boldsymbol{R}_1\boldsymbol{A}^T) \cap \mathcal{K}er(\boldsymbol{R}_\epsilon) & \neq & \{0\} \\ \mathcal{K}er(\boldsymbol{A}\boldsymbol{R}_z\boldsymbol{A}^T) \cap \mathcal{K}er(\boldsymbol{R}_\epsilon) & = & \{0\}\,,\ z = 2..K \end{cases} \tag{VI.22}$$

Avant de construire les matrices $\boldsymbol{R}_z$ et $\boldsymbol{R}_\epsilon$ vérifiant la condition (VI.22), on note que :

1. Si la matrice $\boldsymbol{R}_\epsilon$ est régulière alors toutes les matrices $\boldsymbol{P}_z$ sont régulières. En effet, selon le principe **mini-max** de la caractérisation des valeurs propres de la somme de deux matrices hermitiennes, les valeurs propres des matrices $\boldsymbol{P}_z$ sont supérieures à celles de $\boldsymbol{R}_\epsilon$ et par conséquent strictement positives.
2. Si toutes les matrices $\boldsymbol{R}_z$ sont régulières, alors les matrices $\boldsymbol{P}_z$ sont soit toutes régulières soit toutes singulières. En effet, en général, nous avons :

$$\mathcal{K}er(\boldsymbol{A}^T) \subseteq \mathcal{K}er(\boldsymbol{A}\boldsymbol{R}_z\boldsymbol{A}^T),\ z = 1..K. \tag{VI.23}$$

Dans le cas particulier où toutes les matrices $\boldsymbol{R}_z$ sont régulières, cette inclusion devient une égalité. Toutes les matrices $\boldsymbol{P}_z$ ont donc le même noyau $\mathcal{K}er(\boldsymbol{A}^T) \cap \mathcal{K}er(\boldsymbol{R}_\epsilon)$ et sont ainsi soit toutes régulières soit toutes singulières.

Supposons maintenant que toutes les matrices $\boldsymbol{R}_z$ sont régulières sauf la première matrice $\boldsymbol{R}_1$. Nous allons essayer de construire les matrices $\boldsymbol{R}_1$ et $\boldsymbol{R}_\epsilon$ vérifiant la condition (VI.22). Soit $\boldsymbol{x}_s$ un vecteur non nul appartenant à $[\mathcal{K}er(\boldsymbol{A}^T)]^\perp$. Soit $(\boldsymbol{x}_j)_{j\in J}$ une famille de vecteurs appartenant à $\mathcal{K}er(\boldsymbol{A}^T)$ telle que $\{\boldsymbol{x}_s\} \cup (\boldsymbol{x}_j)_{j\in J}$ soit une base orthonormale (ceci est assuré par le principe de la base incomplète). Les matrices $\boldsymbol{R}_1 = \sum_{j\in J} \alpha_j \, \boldsymbol{x}_j \boldsymbol{x}_j^T \ (\alpha_j \geq 0)$ et $\boldsymbol{R}_\epsilon = \sum_{j\in J} \beta_j \, \boldsymbol{x}_j \boldsymbol{x}_j^T \ (\beta_j \geq 0)$ vérifie les relations suivantes :

$$\begin{cases} \boldsymbol{x}_s \in \mathcal{K}er(\boldsymbol{A}\boldsymbol{R}_1\boldsymbol{A}^T) \cap \mathcal{K}er(\boldsymbol{R}_\epsilon) \\ \\ \mathcal{K}er(\boldsymbol{A}\boldsymbol{R}_z\boldsymbol{A}^T) \cap \mathcal{K}er(\boldsymbol{R}_\epsilon) = \{0\} \end{cases}$$

Par conséquent les matrices $\boldsymbol{R}_1$ et $\boldsymbol{R}_\epsilon$ vérifient la condition (VI.22) ce qui achève la preuve de la proposition (4).

Remarque 24 *On note dans la proposition (4) que l'ensemble des classes z telles que les matrices $\boldsymbol{R}_z$ soient singulières contient celui des classes z telles que les matrices $\boldsymbol{P}_z$ soient singulières. En effet, la matrice $\boldsymbol{R}_z$ peut bien être singulière sans que la matrice $\boldsymbol{P}_z$ le soit.*

L'existence de certains points de singularité[5] pour la vraisemblance $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\gamma})$ est alors simple à prouver.

Proposition 5 $\forall\, \boldsymbol{x}_{1..T} \in (\mathbb{R}^m)^T$, *il existe des points de singularité $\boldsymbol{\gamma}^* \in \Gamma$ où la vraisemblance, définie par (VI.21), diverge :*

$$\lim_{\boldsymbol{\gamma}\to\boldsymbol{\gamma}^*} \mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\gamma}) = \infty.$$

Nécessairement, la covariance du bruit $\boldsymbol{R}_\epsilon$ est singulière et l'une au moins des covariances $\boldsymbol{R}_z$ est singulière.

La preuve de cette proposition repose sur les mêmes arguments que ceux du cas sans mélange étudié dans la section (VI.2), grâce à la proposition (4).

VI.6 Elimination de la dégénérescence dans le cas du mélange

En s'inspirant du cas sans mélange traîté dans les sections précédentes, on va étudier la solution bayésienne qui consiste à choisir un *a priori* inverse wishart pour les matrices de covariance. En fait, nous savons que la dégénérescence se produit, même dans le cas du mélange, quand les matrices de covariance $\boldsymbol{R}_z$ et $\boldsymbol{R}_\epsilon$ approchent la frontière de singularité $\mathcal{F}r(\mathcal{S}_+)$ des matrices positives. En choisissant un *a priori* de telle manière que la distribution *a posteriori* tend vers 0 quand l'une des matrices approchent la frontière de singularité, on élimine le risque de dégénérescence et on garantit en plus que les covariances estimées sont définies positives.

Nous allons montrer qu'il suffit de mettre un *a priori* inverse wishart sur la covariance du bruit $\boldsymbol{R}_\epsilon$ pour éliminer la dégénérescence.

Proposition 6 $\forall\, \boldsymbol{x}_{1..T} \in (\mathbb{R}^m)^T$, *la distribution a posteriori :*

$$\mathrm{p}(\boldsymbol{\gamma} \mid \boldsymbol{x}_{1..T}) \propto \mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\gamma})\, \mathcal{IW}_m(\boldsymbol{R}_\epsilon \,;\, \nu_\epsilon, \boldsymbol{\Sigma}_\epsilon)$$

où la vraisemblance $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\gamma})$ est définie par (VI.21), est bornée et tend vers 0 quand $\boldsymbol{\gamma}$ se rapproche de l'un des points de singularité $\boldsymbol{\gamma}^*$.

Preuve : L'expression (VI.21) de la vraisemblance $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\gamma})$ a la même forme que la vraisemblance dans le cas sans mélange sauf que l'*a priori* est porté sur la matrice $\boldsymbol{R}_\epsilon$ et non sur les matrices $\boldsymbol{P}_z$. On peut

[5] A la différence du cas sans mélange, vu la complexité de l'ensemble des paramètres, on ne cherche pas à caractériser tout l'ensemble des points de singularité.

donc se baser sur les mêmes calculs menés dans la preuve de la proposition (3) :

$$\begin{aligned} \mathrm{p}(\boldsymbol{\gamma} \mid \boldsymbol{x}_{1..T}) &\leq \mathrm{p}(\boldsymbol{R}_\epsilon) \left(\sum_{z=1}^{K} |\boldsymbol{P}_z|^{-\frac{1}{2}} \right)^T \\ &\leq |\boldsymbol{R}_\epsilon^{-1}|^{\frac{\nu_\epsilon+(m+1)}{2}} \exp\left[-\tfrac{1}{2}\nu_\epsilon \mathrm{Tr}\left(\boldsymbol{R}_\epsilon^{-1} {\boldsymbol{\Sigma}_\epsilon}^{-1} \right) \right] |\boldsymbol{P}_{k_{min}}|^{-\frac{T}{2}} \end{aligned} \tag{VI.24}$$

où la classe k_{min} est telle que $|\boldsymbol{P}_{k_{min}}| \leq |\boldsymbol{P}_k|$ pour tout $k \in \{1..K\}$.

Comme on a les deux inégalités suivantes :

$$\begin{cases} |\boldsymbol{R}_\epsilon^{-1} {\boldsymbol{\Sigma}_\epsilon}^{-1}|^{1/m} \leq \frac{1}{m} \mathrm{Tr}\left(\boldsymbol{R}_\epsilon^{-1} {\boldsymbol{\Sigma}_\epsilon}^{-1} \right) \\ |\boldsymbol{R}_\epsilon| \leq |\boldsymbol{P}_{k_{min}}| \end{cases}$$

on obtient

$$\mathrm{p}(\boldsymbol{\gamma} \mid \boldsymbol{x}_{1..T}) \leq \frac{1}{|\boldsymbol{R}_\epsilon|^{\frac{\nu_\epsilon+(m+1)+T}{2}}} \exp\left[-\frac{m\nu_\epsilon}{2} \frac{|{\boldsymbol{\Sigma}_\epsilon}^{-1}|^{1/m}}{|\boldsymbol{R}_\epsilon|^{1/m}} \right]$$

où le terme à droite tend vers 0 quand $\boldsymbol{R}_\epsilon$ approche la singularité. Ce qui prouve que la densité *a posteriori* $\mathrm{p}(\boldsymbol{\gamma} \mid \boldsymbol{x}_{1..T})$ tend vers 0 quand l'une des conditions nécessaires de la dégénérescence ($\boldsymbol{R}_\epsilon$ est singulière) est vérifiée.

VI.7 Conclusion

Nous avons montré que le critère du maximum de vraisemblance, dans le cas de l'estimation des paramètres d'un mélange de gaussiennes multivariables, souffre d'un problème de **dégénérescence**. La vraisemblance diverge vers l'infini quand le paramètre se rapproche de l'un des points de singularité. Ces points de singularité sont caractérisés par la propriété suivante :

Propriété 5 *Au moins l'une des matrices de covariance $\boldsymbol{R}_z$ est positive singulière et la moyenne correspondante $\boldsymbol{\mu}_z$ appartient à l'intersection de $n - rang(\boldsymbol{R}_z)$ hyperplans de $\mathbb{R}^n$.*

Cette caractérisation montre que le risque de dégénérescence est plus important dans le cas vectoriel que dans le cas scalaire où les points de singularité sont tels qu'une des variances σ_z^2 est nulle et la moyenne correspondante μ_z coïncide avec une observation.

Une des conditions nécessaires de cette dégénérescence est que l'une des matrices de covariance approche la frontière de singularité. Contraindre les matrices $\boldsymbol{R}_z$ à être régulières complique le problème d'identification. Pour se rendre compte de cette complexité, il suffit de considérer le cas scalaire où la contrainte ($\sigma_z/\sigma_{z'} > c > 0,\ z, z' \in \mathcal{Z}$) (solution envisagée par [Hathaway, 1986]) complique la mise en œuvre de l'algorithme EM.

Le contexte bayésien offre une solution efficace pour éliminer le risque de dégénérescence sans contraindre algorithmiquement les matrices à être régulières. Cette solution consiste à choisir une densité *a priori* de telle façon que la densité *a posteriori* soit bornée et qu'elle tend vers 0 quand l'une des matrices de covariance approche la frontière de singularité. En se basant sur les propriétés topologiques de l'espace des paramètres, nous avons montré l'existence d'un maximiseur de la densité *a posteriori* et que les covariances correspondantes sont nécessairement régulières. L'*a priori* inverse wishart remplit cette condition et présente en plus l'avantage que l'algorithme EM garde une forme simple et efficace à implémenter.

Nous avons traîté en plus d'autres types de contraintes qui consistent à imposer une structure particulière aux matrices de covariance. Dans le cas où l'espace de contrainte est stable par variations continues des éléments de la matrice de covariance, nous avons généralisé la procédure de [Burg, 1982] pour proposer un algorithme **EM renversé** capable d'identifier les paramètres du mélange en préservant la structure imposée aux covariances.

Finalement, nous avons étudié ces risques de dégénérescence dans le cas où les sources ne sont plus directement accessibles mais plutôt mélangées et bruitées. La singularité de la covariance du bruit et d'au moins une des covariances de l'*a priori* des sources est une condition nécessaire de cette dégénérescence. En s'inspirant du cas sans mélange et en choisissant un *a priori* inverse wishart sur la covariance du bruit, nous avons montré que la distribution *a posteriori* du paramètre global (qui est constitué de la matrice de mélange, de la covariance du bruit et des moyennes et variances de la distribution *a priori* des sources) est bornée et tend vers 0 quand on se rapproche d'un point singulier provoquant la dégénérescence de la vraisemblance.

Références

[Burg, 1982] J. P. Burg. Estimation of structured covariance matrices. *Proceeding of IEEE*, 70 (9) : 963–974, septembre 1982.

[Day, 1969] N. Day. Estimating the components of a mixture of normal distributions. *Biometrika*, 56 : 463–474, 1969.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird et D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39 : 1–38, 1977.

[Descombes, 1993] X. Descombes. *Champs Markoviens en analyse d'image*. thèse de doctorat, École Nationale Supérieure des Télécommunications, Paris, décembre 1993.

[Gostiaux, 1993a] B. Gostiaux. *Cours de mathématiques spéciales. Algèbre*. Presses Universitaires de France, Paris, 1993.

[Gostiaux, 1993b] B. Gostiaux. *Cours de mathématiques spéciales. Analyse fonctionnelle et calcul différentiel*. Presses Universitaires de France, Paris, 1993.

[Hathaway, 1986] R. J. Hathaway. A constrained EM algorithm for univariate normal mixtures. *J. Statist. Comput. Simul.*, 23 : 211–230, 1986.

[Kiefer et Wolfowitz, 1956] J. Kiefer et J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.*, 27 : 887–906, 1956.

[McLachlan et Basford, 1987] G. J. McLachlan et K. E. Basford. *Mixture Models, inference and applications to clustering*, volume 84 de *statistics*. Dekker, 1987.

[McLachlan et Peel, 2000] G. J. McLachlan et D. Peel. *Finite Mixture Models*. Wiley series in probability and statistics. Wiley, 2000.

[Ormoneit et Tresp, 1998] D. Ormoneit et V. Tresp. Averaging, maximum penalized likelihood and Bayesian estimation for improving Gaussian mixture probability density estimates. *IEEE Transactions on Neural Networks*, 9 (4) : 639–649, juillet 1998.

[Rabiner, 1989] R. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77 (2) : 257–286, février 1989.

[Ridolfi et Idier, 1999] A. Ridolfi et J. Idier. Penalized maximum likelihood estimation for univariate normal mixture distributions. In *Actes 17e coll. GRETSI*, pages 259–262, Vannes, septembre 1999.

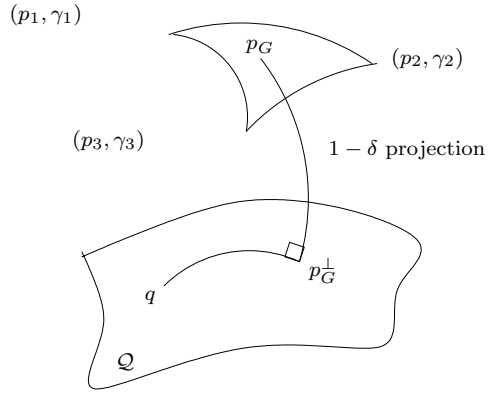
[Roeder et Wasserman, 1997] K. Roeder et L. Wasserman. Practical bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.*, 92 : 894–902, 1997.

[Snoussi et Mohammad-Djafari, 2002a] H. Snoussi et A. Mohammad-Djafari. Bayesian unsupervised learning for source separation with mixture of gaussians prior. *To appear in Int. Journal of VLSI Signal Processing Systems*, 2002.

[Snoussi et Mohammad-Djafari, 2002b] H. Snoussi et A. Mohammad-Djafari. MCMC Joint Separation and Segmentation of Hidden Markov Fields. In *Neural Networks for Signal Processing XII*, pages 485–494. IEEE workshop, septembre 2002.

Chapitre VII

SÉLECTION D'*a priori* ET GÉOMÉTRIE DE L'INFORMATION



Ce chapitre est consacré au problème de la sélection de la loi *a priori* dans un contexte bayésien. Nous présentons une approche originale [Snoussi et Mohammad-Djafari, 2002a] basée sur la théorie de la prédiction bayésienne [Zhu et Rohwer, 1995a] en utilisant les outils de la géométrie de l'information [Amari et Nagaoka, 2000]. On montre l'importance du choix de la géométrie dans l'espace des distributions de probabilité. La règle de Bayes permet de définir la masse par la loi *a posteriori*. Une fois la géométrie et la masse fixées, on construit un critère variationnel dont la minimisation donne la loi *a priori* qu'on a notée δ-*a priori*. Avec les outils de la géométrie différentielle, on introduit la notion d'*a priori* projeté pour les familles paramétriques. Ce travail est appliqué au mélange de familles δ-plates comme le mélange de familles exponentielles (0-plates) et en séparation de sources.

Information geometry and prior selection

H. Snoussi, A. Mohammad-Djafari

∗ Laboratoire des Signaux et Systèmes (L2S), Supélec, Plateau de Moulon, 91192 Gif-sur-Yvette Cedex, France

email = snoussi@lss.supelec.fr, djafari@lss.supelec.fr

Abstract

In this contribution, we study the problem of prior selection arising in Bayesian inference. There is an extensive literature on the construction of non informative priors and the subject seems far from a definite solution [Kass et Wasserman, 1994]. Here we revisit this subject with differential geometry tools and propose to construct the prior in a Bayesian decision theoretic framework. We show how the construction of a prior by projection is the best way to take into account the restriction to a particular family of parametric models. For instance, we apply this procedure to the curved parametric families where the ignorance is directly expressed by the relative geometry of the restricted model in the wider model containing it.

VII.1 Introduction

Experimental science can be modeled as a learning machine mapping the inputs $\boldsymbol{x}$ to the outputs $\boldsymbol{y}$ (see figure (VII.1)). The complexity of the physical mechanism underlying the mapping inputs/outputs or the lack of information make the prediction of the outputs given the inputs (forward model) or the estimation of the inputs given the outputs (inverse problem) a difficult task. When a parametric forward model $p(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{\theta})$ is assumed to be available from the knowledge of the system, one can use the classical ML or when a prior model $p(\boldsymbol{x},\boldsymbol{\theta}) = p(\boldsymbol{x}\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta})$ is assumed to be available too, the classical Bayesian methods can be used to obtain the joint *a posteriori* $p(\boldsymbol{x},\boldsymbol{\theta}\,|\,\boldsymbol{y})$ and then both $p(\boldsymbol{x}\,|\,\boldsymbol{y})$ and $p(\boldsymbol{\theta}\,|\,\boldsymbol{y})$ from which we can make any inference about $\boldsymbol{x}$ and $\boldsymbol{\theta}$. But in many practical situations the question of modeling $p(\boldsymbol{y}\,|\,\boldsymbol{x})$ and $p(\boldsymbol{x})$ is still open and to validate a model, one uses what is called the training data $\boldsymbol{z} = (\boldsymbol{x}_i,\boldsymbol{y}_i)_{i=1..N}$. Then the role of statistical learning become trying to find a joint distribution $p(\boldsymbol{z})$ belonging in general to the whole set of probability distributions and to exploit the maximum of relevant information to provide some desired predictions. In this paper, we suppose that we are given some training data $\boldsymbol{x}_{1..N}$ and $\boldsymbol{y}_{1..N}$ and some information about the mapping which consists in a model $\mathcal{Q} = \{P(\boldsymbol{z})\}$ of probability distributions, parametric ($\mathcal{Q} = \{P(\boldsymbol{z}\,|\,\boldsymbol{\theta})\}$) or non parametric. Our objective is to construct a learning rule τ mapping the set $\mathcal{Z}$ of training data $\boldsymbol{z} = (\boldsymbol{x}_{1..N},\boldsymbol{y}_{1..N})$ to a probability distribution $p \in \mathcal{Q}$ or to a probability distribution in the whole set of probabilities $p \in \mathcal{P}$:

$$\begin{array}{rccc} \tau: & \mathcal{Z} & \longrightarrow & \mathcal{Q} \\ & \boldsymbol{z} & \mapsto & q = \tau(\boldsymbol{z}) \end{array}$$

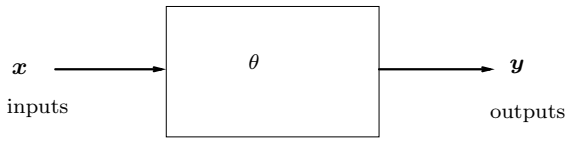


FIG. VII.1: Learning machine model of experimental science

The Bayesian statistical learning leads to a solution depending on the prior distribution of the unknown distribution p. In the parametric case, this is equivalent to the prior $\Pi(\boldsymbol{\theta})$ on the parameter $\boldsymbol{\theta}$. Finding a general expression for $\Pi(\boldsymbol{\theta})$ and how this expression reflects the relationship between a restricted model and the closer set of ignorance containing it are the main objectives of this paper. We show the prior expression depends on the chosen geometry (subjective choice) of the set of probability measures. We show that the entropic prior [Rodríguez, 1991] and the conjugate prior of exponential families are special cases related to special geometries.

In section I, we review briefly some concepts of Bayesian geometrical statistical learning and the role of differential geometry. In section II, we develop the basics of prior selection in a Bayesian decision perspective and we discuss the effect of model restriction both from non parametric to parametric modelization and from parametric family to a curved family. In section III, we study the particular case of δ-flat families where previous results have explicit formula. In section IV, we come across the case of δ-flat families mixture. In section V, we apply these results to a couple of learning examples, the mixture of multivariate Gaussian classification and blind source separation. We end with a conclusion and indicate some future scopes.

VII.2 Statistical geometric learning

VII.2.1 Mass and Geometry

The statistical learning consists in constructing a learning rule τ which maps the training measured data $\boldsymbol{z}$ to a probability distribution $q = \tau(\boldsymbol{z}) \in \mathcal{Q} \subset \mathcal{P} = \{p \mid \int p = 1\}$ (the predictive distribution). The subset $\mathcal{Q}$ is in general a parametric model and it is called the computational model. Therefore, our target space is the space of distributions and it is fundamental to provide this space with, at least in this work, two attributes which are the mass (a scalar field) and a geometry. The mass is defined by an *a priori* distribution $\Pi(p)$ on the space $\mathcal{P}$ before collecting the data $\boldsymbol{z}$ and modified according to Bayesian rule after observing the data to give the *a posteriori* distribution (see figure (VII.2)) :

$$P(p \mid \boldsymbol{z}) \propto P(\boldsymbol{z} \mid p)\, \Pi(p)$$

where $P(\boldsymbol{z} \mid p)$ is $p(\boldsymbol{z})$ the likelihood of the probability p to generate the data $\boldsymbol{z}$.

The geometry can be defined by the δ-divergence D_δ :

$$D_\delta(p, q) = \frac{\int p}{1-\delta} + \frac{\int q}{\delta} - \frac{\int p^\delta q^{1-\delta}}{\delta(1-\delta)}$$

which is an invariant measure under reparametrization of the restricted parametric model $\mathcal{Q}$. It is shown [Amari 1985, Amari85] that, in the parametric manifold $\mathcal{Q}$, the δ-divergence induces a dualistic structure

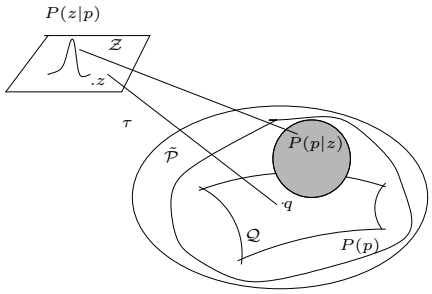


FIG. VII.2: *a posteriori* mass proportional to the product of the *a priori* mass and the likelihood function

$(g, \nabla^\delta, \nabla^{1-\delta})$, where g is the Fisher metric, ∇^δ the δ connection with Christoffel symbols $\Gamma^\delta_{ij,k}$ and $\nabla^* = \nabla^{1-\delta}$ its dual connection :

$$\begin{cases} g_{ij} & = & E_{\boldsymbol{\theta}}\left[\partial_i l(\boldsymbol{\theta})\, \partial_j l(\boldsymbol{\theta})\right] \\ \Gamma^\delta_{ij,k} & = & E_{\boldsymbol{\theta}}\left[(\partial_i \partial_j l(\boldsymbol{\theta}) + \delta \partial_i l(\boldsymbol{\theta}) \partial_j l(\boldsymbol{\theta}))\, \partial_k l(\boldsymbol{\theta})\right] \end{cases}$$

The parametric manifold $\mathcal{Q}$ is δ-flat if and only if there exists a parameterization $[\theta_i]$ such that the Christoffel symbols vanish : $\Gamma^\delta_{ij,k}(\boldsymbol{\theta}) = 0$. The coordinates $[\theta_i]$ are called the affine coordinates. If for a different coordinate system $[\theta_i^{'}]$, the connection coefficients are null then the two coordinate systems $[\theta_i]$ and $[\theta_i^{'}]$ are related by an affine transformation, i.e there exists a $(n \times n)$ matrix $\boldsymbol{A}$ and a vector $\boldsymbol{b}$ such that $\boldsymbol{\theta}^{'} = \boldsymbol{A\theta} + \boldsymbol{b}$.

All the above definitions can be extended to non parametric families by replacing the partial derivatives with the Fréchet derivatives. Embedding the model $\mathcal{Q}$ in the whole space of finite measures $\tilde{\mathcal{P}}$ [Zhu et Rohwer, 1995a,b] not only the space of probability distributions $\mathcal{P}$, many results can be proven easily for the main reason that $\tilde{\mathcal{P}}$ is δ-flat and δ-convex $\forall\ \delta$ in $[0\,,\,1]$. However, $\mathcal{P}$ is δ-flat for only $\delta = \{0,\, 1\}$ and δ-convex for $\delta = 1$. For notation convenience, we use the δ-coordinates $\overset{\delta}{l}$ of a point $p \in \tilde{\mathcal{P}}$ defined as :

$$\overset{\delta}{l}(p) = p^\delta/\delta$$

A curve linking 2 points a and b is a function $\gamma :\ [0\,,\,1] \longrightarrow \tilde{\mathcal{P}}$, such that $\gamma(0) = a$ and $\gamma(1) = b$. A curve is a δ-geodesic in the δ-geometry if it is a straight line in the δ-coordinates.

VII.2.2 Bayesian learning

The loss quantity of a decision rule τ with a fixed δ-geometry can be measured by the δ-divergence $D_\delta(p, \tau(\boldsymbol{z}))$ between the true probability p and the decision $\tau(\boldsymbol{z})$. This divergence is first averaged with respect to all possible measured data $\boldsymbol{z}$ and then with respect to the unknown true probability p which gives the generalization error $E(\tau)$:

$$E_\delta(\tau) = \int_p P(p) \int_{\boldsymbol{z}} P(\boldsymbol{z}\,|\,p) D_\delta(p, \tau(\boldsymbol{z}))$$

Therefore, the optimal rule τ_δ is the minimizer of the generalization error :

$$\tau_\delta = \arg\min_\tau \{E_\delta(\tau)\}$$

The coherence of Bayesian learning is shown in [Zhu et Rohwer, 1995a,b] and means that the optimal estimator τ_δ can be computed pointwise as a function of $\boldsymbol{z}$ and we don't need a general expression of the optimal estimator τ_δ :

$$\hat{p}(\boldsymbol{z}) = \tau_\delta(\boldsymbol{z}) = \arg\min_q \int_p P(p\,|\,\boldsymbol{z}) D_\delta(p, q) \tag{VII.1}$$

By variational calculation, the solution of (VII.1) is straightforward and gives :

$$\hat{p}^{\delta} = \int p^{\delta} P(p \,|\, z)$$

The above solution is exactly the gravity center of the set $\tilde{\mathcal{P}}$ with mass $P(p \,|\, \boldsymbol{z})$, the *a posteriori* distribution of p and the δ-geometry induced by the δ-divergence D_δ. Here we have the analogy with the static mechanics and the importance of the geometry defined on the space of distributions. The whole space of finite measures $\tilde{\mathcal{P}}$ is δ-convex and thus, independently on the *a posteriori* distribution $P(p \,|\, \boldsymbol{z})$ the solution $\hat{p}$ belongs to $\tilde{\mathcal{P}}$ $\forall\ \delta \in [0\,,\,1]$.

VII.2.3 Restricted Model

In practical situations, we restrict the space of decisions to a subset $\mathcal{Q} \in \tilde{\mathcal{P}}$. $\mathcal{Q}$ is in general a parametric manifold that we suppose to be a differentiable manifold. Thus $\mathcal{Q}$ is parametrized with a coordinate system $[\theta_i]_{i=1}^n$ where n is the dimension of the manifold. $\mathcal{Q}$ is also called the computational model and we prefer this appellation because the main reason of the restriction is to design and manipulate the points p with their coordinates which belong to an open subset of $\mathbb{R}^n$. However, the computational model $\mathcal{Q}$ is not disconnected from non parametric manipulations and we will show that both *a priori* and final decisions can be located outside the model $\mathcal{Q}$.

Let's compare now the non parametric learning with the parametric learning when we are constrained to a parametric model $\mathcal{Q}$:

1. **Non parametric modeling :** The optimal estimate is the minimizer of the generalization error where the true unknown point p is allowed to belong to the whole space $\tilde{\mathcal{P}}$ and the minimizer q is constrained to $\mathcal{Q}$:

$$\hat{q}(\boldsymbol{z}) = \tau_\delta(\boldsymbol{z}) = \arg\min_{q \in \mathcal{Q}} \int_{p \in \tilde{\mathcal{P}}} P(p \,|\, \boldsymbol{z}) D_\delta(p, q) \tag{VII.2}$$

   Thus the solution is the δ-projection of the barycentre $\hat{p}$ of ($\tilde{\mathcal{P}}$, $P(p \,|\, \boldsymbol{z})$, D_δ) onto the model $\mathcal{Q}$ (see figure (VII.3)).

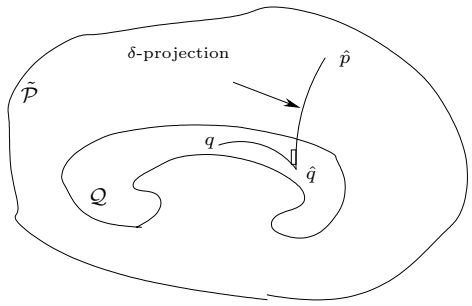


Fig. VII.3: Projection of the non parametric solution onto the computational model

2. **Parametric modeling :** The optimal estimate is the minimizer of the same cost function as in the non parametric case but the true unknown point p is also constrained to be in $\mathcal{Q}$:

$$\hat{q}(\boldsymbol{z}) = \tau_\delta(\boldsymbol{z}) = \arg\min_{q \in \mathcal{Q}} \int_{p \in \mathcal{Q}} P(p \,|\, \boldsymbol{z}) D_\delta(p, q) = \arg\min_{q \in \mathcal{Q}} \int_{\boldsymbol{\theta}} P(\boldsymbol{\theta} \,|\, \boldsymbol{z}) D_\delta(p_\theta, q) d\,\boldsymbol{\theta} \tag{VII.3}$$

   The solution is the δ-projection of the barycentre $\hat{p}$ of ($\mathcal{Q}$, $P(\boldsymbol{\theta} \,|\, \boldsymbol{z})$, D_δ) onto the model $\mathcal{Q}$ (see figure (VII.4)).

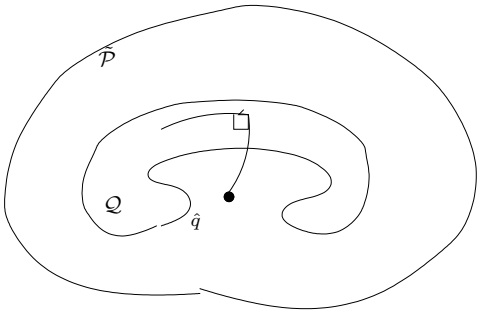


FIG. VII.4: Projection of the barycentre solution onto the parametric model

The interpretation of the parametric modeling as a non parametric one and the effect of such restriction can be done in two ways :

1. The cost function to be minimized in equation (VII.3) is the same as the cost function in (VII.2) when p is allowed to belong to the whole set $\tilde{\mathcal{P}}$ and the *a posteriori* $P(p \mid \boldsymbol{z})$ is zero outside the model $\mathcal{Q}$. This is the case when the prior $P(p)$ has $\mathcal{Q}$ as its support. However this interpretation implies that the best solution $\hat{p}$ which is the barycentre of $\mathcal{Q}$ can be located outside the model $\mathcal{Q}$ and thus has *a priori* a zero probability !
2. The second interpretation is to say that the cost function to be minimized in equation (VII.3) is the same as the cost function in (VII.2) when the *a posteriori* $P(\boldsymbol{\theta} \mid \boldsymbol{z})$ is the projected mass of the *a posteriori* $P(p \mid \boldsymbol{z})$ onto the model $\mathcal{Q}$. We note here the role of the geometry defined on the space $\mathcal{P}$ and the relative geometric shape of the manifold. For instance, the ignorance is directly related to the geometry of the model $\mathcal{Q}$. The projected *a posteriori* or *a priori* can be computed by :

$$f^{\perp}(q) \propto \int_{p \in \mathcal{S}_q} f(p)$$

where $f(p)$ designs the *a priori* or the *a posteriori* distribution and $\mathcal{S}_q = \{p \in \tilde{\mathcal{P}} \mid p^{\perp} = q\}$ the set of points p whose the δ-projection is the q in $\mathcal{Q}$.

The manipulation of these concepts in the general case is very abstract. However, in section IV, we present the explicit computations in the case of restricted autoparallel parametric submanifold $\mathcal{Q}_1 \in \mathcal{Q}$ of δ-flat families.

VII.3 Prior selection

The present section is the main contribution of this paper. We address here the problem of prior selection in a Bayesian decision framework. By prior selection, we mean how to construct a prior $P(p)$ respecting the following rule : Exploit the prior knowledge without adding irrelevant information. We note that this represents a trade off between some desirable behaviour and uniformity of the prior. We want to insist here, that the prior selection must be performed before collecting the data $\boldsymbol{z}$, otherwise the coherence of the Bayesian rule is broken down.

In a decision framework, the desirable behaviour can be stated as follows : Before collecting the training data, provide a reference distribution p_0 as a decision. The reference distribution can be provided by an expert or by our previous experience. Now, we have the inverse problem of the statistical learning. Before, the *a posteriori* distribution (mass) is fixed and we have to find the optimal decision (barycentre). Now, the optimal decision p_0 (barycentre) is fixed and we have to find the optimal repartition $\Pi(p)$ according to the uniformity constraint. In order to have the usual notions of integration and derivation, we assume that our objective is to find the prior on the parametric model $\mathcal{Q} = \{q_\theta \mid \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^n\}$.

The cost function can be constructed as a weighted sum of the generalization error of the reference prior and the divergence of the prior from the Jeffreys prior (The square root of the determinant of the Fisher information [Box et Tiao, 1972]) representing the uniformity. It is worth noting that we are considering two

different spaces : the space $\tilde{\mathcal{P}}$ of finite measures and the space of prior distributions on the finite measures. Since we have two distinct spaces, we can choose two different geometries on each space. For example, if we consider the δ-geometry on the space $\tilde{\mathcal{P}}$ and the 1-geometry on the space of priors, we have the following cost function :

$$J(\Pi) = \gamma_e \int \Pi(\boldsymbol{\theta}) D_\delta(p_\theta, p_0) d\boldsymbol{\theta} + \gamma_u \int \Pi(\boldsymbol{\theta}) \log \Pi(\boldsymbol{\theta}) / \sqrt{g(\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{VII.4}$$

where γ_e is the confidence degree in the reference distribution p_0 and γ_u the uniformity degree. Considered independently, these two coefficients are not significant. However, their ratio is relevant in the following. The cost (VII.4) can be rewritten as :

$$\begin{cases} J(\Pi) = \gamma_e E(\tau_0) + \gamma_u \int \Pi(\boldsymbol{\theta}) \log \Pi(\boldsymbol{\theta}) / \sqrt{g(\boldsymbol{\theta})} d\boldsymbol{\theta} \\ \\ \frac{\partial \tau_0}{\partial \boldsymbol{z}} = 0 \end{cases}$$

where $E(\tau_0)$ is the generalisation error of a fixed learning rule τ_0. By variational calculation, we obtain the solution of the minimization of the function (VII.4) :

$$\Pi(\boldsymbol{\theta}) \propto e^{-\frac{\gamma_e}{\gamma_u} D_\delta(p_\theta, p_0)} \sqrt{g(\boldsymbol{\theta})} \tag{VII.5}$$

We note that if $\delta = 1$ then the cost function (VII.4) is the kullback-Leibler divergence between the joint distributions of data and parameters as considered in [Rodríguez, 1991] and if $\delta = 0$ we obtain the conjugate prior for exponential families (see examples in section VI). When the value of the ratio γ_e/γ_u goes to 0, we obtain the Jeffreys prior and when this ratio goes to ∞ we obtain the Dirac concentrated on p_0.

The model restriction to the parametric manifold $\mathcal{Q}$ is essentially for computational reasons. However, the reference distribution is a prior decision and does not depend on a post processing after collecting the data. Therefore, the reference distribution p_0 can be located in the whole space of probability measures. We can also have either a discrete set of N reference distributions $(p_0^i)_{i=1}^N$ weighted by $(\gamma_e^i)_{i=1}^N$ or a continuous set of reference distributions (a region or the whole set of probability distributions) with a probability measure $P(p_0)$ corresponding to the weights $(\gamma_e^i)_{i=1}^N$ in the discrete case. We show in the following that the prior solution Π has the same form as (VII.5).

1. $p_0 \notin \mathcal{Q}$: When the reference distribution p_0 is located outside the model $\mathcal{Q}$, the δ-divergence $D_\delta(p_\theta\ ,\ p_0)$ in the expression (VII.4) can be decomposed according to the generalized Pythagore relation [Amari et Nagaoka, 2000] :
$$D_\delta(p_\theta\ ,\ p_0) = D_\delta(p_\theta\ ,\ p_0^\perp) + D_\delta(p_0^\perp\ ,\ p_0)$$
where $p_0^\perp$ is the $1-\delta$-projection of p_0 onto $\mathcal{Q}$ (see figure (VII.5)).

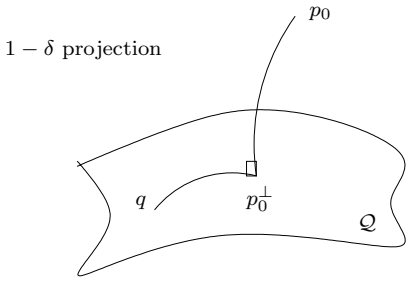


FIG. VII.5: The equivalent of the non parametric reference distribution is its $1-\delta$ projection onto the parametric model $\mathcal{Q}$.

Giving the prior solution :

$$\Pi(\boldsymbol{\theta}) \propto e^{-\frac{\gamma_e}{\gamma_u} D_\delta(p_\theta, p_0^\perp)} \sqrt{g(\boldsymbol{\theta})}$$

2. When we have N reference distributions $\{(p_1, \gamma_1), ..., (p_N, \gamma_N)\}$, the cost function (VII.4) becomes :

$$J_N(\Pi) = \sum_{i=1}^{N} \gamma_i \int \Pi(\boldsymbol{\theta}) D_\delta(p_\theta, p_i) d\,\boldsymbol{\theta} + \gamma_u \int \Pi(\boldsymbol{\theta}) \log \Pi(\boldsymbol{\theta}) / \sqrt{g(\boldsymbol{\theta})} d\,\boldsymbol{\theta} \tag{VII.6}$$

If we define the $1-\delta$-barycentre p_G of the system $\{(p_1, \gamma_1), ..., (p_N, \gamma_N)\}$ as

$$\overset{1-\delta}{l}(p_G) = \sum_{i=1}^{N} \gamma_i \overset{1-\delta}{l}(p_i) / \sum_{i=1}^{N} \gamma_i$$

and the $p_G^\perp$ the $1-\delta$ projection of p_G onto $\mathcal{Q}$, the solution Π of the minimization of (VII.6) is :

$$\Pi(\boldsymbol{\theta}) \propto e^{-\frac{\sum \gamma_i}{\gamma_u} D_\delta(p_\theta, p_G^\perp)} \sqrt{g(\boldsymbol{\theta})}$$

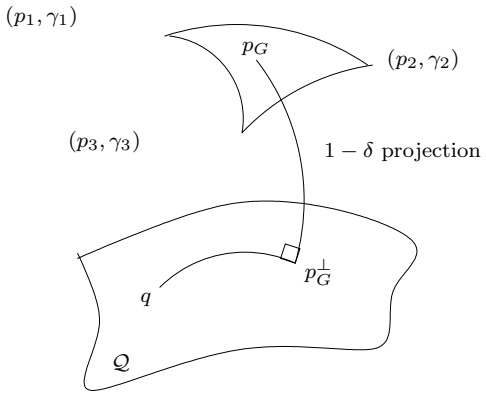


FIG. VII.6: The equivalent reference distribution is the $1-\delta$ projection of the $1-\delta$ barycentre of the N references distributions.

3. When we have a continuous set $\mathcal{P}_r \subseteq \tilde{\mathcal{P}}$ of reference distributions with a mass distribution $P_r(p_0)$, the cost function is transformed to :

$$J_c(\Pi) = \int_{p_0 \in \mathcal{P}_r} P_r(p_0) \int \Pi(\boldsymbol{\theta}) D_\delta(p_\theta, p_0) d\,\boldsymbol{\theta} + \gamma_u \int \Pi(\boldsymbol{\theta}) \log \Pi(\boldsymbol{\theta}) / \sqrt{g(\boldsymbol{\theta})} d\,\boldsymbol{\theta} \tag{VII.7}$$

In the same way, we define the $1-\delta$ barycentre p_G of $(\mathcal{P}_r,\, P_r)$ as :

$$\overset{1-\delta}{l}(p_G) = \int_{\mathcal{P}_r} P_r(p_0) \overset{1-\delta}{l}(p_0) / \int_{\mathcal{P}_r} P_r(p_0)$$

and the $p_G^\perp$ the $1-\delta$ projection of p_G onto $\mathcal{Q}$, the solution Π of the minimization of (VII.7) is :

$$\Pi(\boldsymbol{\theta}) \propto e^{-\frac{\int P(p_0)}{\gamma_u} D_\delta(p_\theta, p_G^\perp)} \sqrt{g(\boldsymbol{\theta})}$$

The above results show that whatever the choice of the reference distribution is, the resulting prior has the same form with a certain (non arbitrary) reference prior belonging to the model $\mathcal{Q}$. The existence of many reference distributions (or even a continuous set) indicates implicitly the existence of hyperparameter and the resulting solution shows that this hyperparameter is integrated and at the same time optimized if the *a priori* average (the barycentre) is considered as an optimization operation.

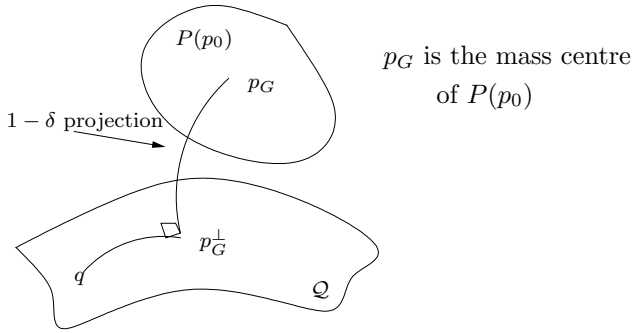


FIG. VII.7: The equivalent reference distribution of a continum reference region is the $1-\delta$ projection of the $1-\delta$ expectation reference.

VII.4 δ-flat families

In this section we study the particular case of δ-flat families. $\mathcal{Q}$ is a δ flat manifold if and only if there exists a coordinate system $[\theta_i]$ such that the connection coefficients $\Gamma_\delta(\boldsymbol{\theta})$ are null. We call $[\theta_i]$ an affine coordinate system. It is known that δ-flatness is equivalent to $1-\delta$ flatness. Therefore, there exist dual affine coordinates $[\eta_i]$ such that $\Gamma_{1-\delta}(\boldsymbol{\eta}) = 0$. One of the many properties of δ-flat families is that we can express, in a simple way, the δ-divergence D_δ as a function of the coordinates $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ and thus any decision can be computed while manipulating the real coordinates. It is shown in [Amari, 1985] that the dual affine coordinates $[\theta_i]$ and $[\eta_i]$ are related by Legendre transformations and the canonical divergence is :

$$D_\delta(p,q) = \psi(p) + \phi(q) - \theta_i(p)\eta_i(q)$$

where ψ and ϕ are the dual potentials such that :

$$\begin{cases} \frac{\partial \eta_j}{\partial \theta_i} = g_{ij} \quad \frac{\partial \theta_i}{\partial \eta_j} = g_{ij}^{-1} \\ \partial_i \psi = \eta_i \quad \partial_i \phi = \theta_i \end{cases}$$

For example, the exponential families are 1-flat with the canonical parameters as 1-affine coordinates, the mixture family is 0-flat with the mixture coefficients as 0-affine coordinates, $\tilde{\mathcal{P}} = \{p, \int p < \infty\}$ is δ flat for all $\delta \in [0\,,\,1]$.

[A] δ optimal estimates in δ flat families

As indicated in section II, the δ optimal estimate is the δ projection of $\int_\theta p^\delta P(\boldsymbol{\theta} \,|\, \boldsymbol{z})$ which is the minimizer of the functional $\int_\theta P(\boldsymbol{\theta} \,|\, \boldsymbol{z}) D_\delta(p_\theta, q)$. We see that, in general, the divergence as a function of the parameters $[\theta_i]$ has not a simple expression. However, with δ-flat manifolds, we obtain an explicit solution. Noting that :

$$\partial_i D_\delta(p_\theta, q) = D_\delta(p_\theta, (\partial_i)_q) = \theta_i(q) - \theta_i(p)$$

the solution is :

$$\hat{q} = q(\hat{\boldsymbol{\theta}}), \;\; \hat{\boldsymbol{\theta}} = \int \boldsymbol{\theta} P(\boldsymbol{\theta} \,|\, \boldsymbol{z}) d\boldsymbol{\theta} = E_{\theta \,|\, \boldsymbol{z}}[\boldsymbol{\theta}]$$

This means that the δ optimal estimate is the *a posteriori* expectation of the δ affine coordinates. Since the only degree of freedom of the affine coordinates is the affine transformation, this estimate is invariant under affine reparameterization.

Noting also that :

$$\partial_i D_{1-\delta}(p,q) = D_{1-\delta}(p, (\partial_i)_q) = \eta_i(q) - \eta_i(p)$$

Then the *a posteriori* expectation of the $1-\delta$ affine coordinates is the $1-\delta$ optimal estimate.

[B] Prior selection with δ flat families

The δ prior Π has the following general expression :

$$\Pi(\boldsymbol{\theta}) \propto e^{-\frac{\gamma_e}{\gamma_u} D_\delta(p_\theta, p_0)} \sqrt{g(\boldsymbol{\theta})}$$

where $p_0 \in \mathcal{Q}$ is the equivalent reference distribution in the manifold $\mathcal{Q}$. When we assume that $\mathcal{Q}$ is δ flat with affine coordinates $[\theta_i]$ and dual affine coordinates $[\eta_i]$, the expression of the prior becomes :

$$\Pi(\boldsymbol{\theta}) \propto e^{-\frac{\gamma_e}{\gamma_u}\left(\psi(\boldsymbol{\theta}) - \theta_i\, \eta_i^0\right)} \sqrt{g(\theta)}$$

where $[\theta_i^0]$ and $[\eta_i^0]$ are the affine coordinates of p_0.

Therefore, we have an explicit analytic expression of the prior.

In the Euclidean case, that is when the connection ∇ is equal to its dual connection ∇^*, which is equivalent to equality of the affine coordinates $[\theta_i] = [\eta_i]$, the δ prior distribution is Gaussian with mean $\boldsymbol{\theta}_0$ and precision $2\,\frac{\gamma_e}{\gamma_u}$:

$$\Pi(\theta) \propto e^{-\frac{\gamma_e}{\gamma_u}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2}$$

We detail here the notion of prior projection in the particular case of ∇^*-autoparallel submanifolds $\mathcal{Q}_a \subset \mathcal{Q}$. $\mathcal{Q}_a$ is $(1-\delta)$-autoparallel in $\mathcal{Q}$ if and only if, at every point $p \in \mathcal{Q}_a$, the covariant derivative $\nabla^*_{\partial_a} \partial_b$ remains in the tangent space $\mathcal{T}_p$ of the submanifold $\mathcal{Q}_a$ at the point p. A simple characterization in flat manifolds is that the $(1-\delta)$-affine coordinates $[u_i]$ of $\mathcal{Q}_a$ form an affine subspace of the coordinates $[\eta_i]$. We can show that by a suitable affine reparametrization of $\mathcal{Q}$, the submanifold $\mathcal{Q}_a$ is defined as :

$$\begin{cases} \mathcal{Q}_a = \left\{ p_\eta \in \mathcal{Q} \,|\, \boldsymbol{\eta}_I = \boldsymbol{\eta}_I^0 \text{ is fixed } \right\} \\ \\ I \subset \{1..n\} \end{cases}$$

where $n - |I|$ is the dimension of $\mathcal{Q}_a$. If we consider the space $\mathcal{Q}_a^c$ such the complementary dual affine coordinates $\boldsymbol{\theta}_{II} = \boldsymbol{\theta}_{II}^0$ are fixed ($II = \{1..n\} - I$), then the tangent spaces $\mathcal{T}_p$ and $\mathcal{T}_p^c$ at the point $p(\boldsymbol{\eta}_I^0,\, \boldsymbol{\theta}_{II}^0)$ are orthogonal. Consequently, the projected prior from $\mathcal{Q}$ onto $\mathcal{Q}_a$ is simply :

$$\Pi^\perp(p) = \int_{q \in \mathcal{Q}_a^c} \Pi(q) = \int_{\theta_I} \Pi(\boldsymbol{\theta}_I,\, \boldsymbol{\theta}_{II}) d\,\boldsymbol{\theta}_I$$

Hence, we see that the projected prior onto a ∇^*-autoparallel manifold is the marginalization in the δ affine coordinates and not in with respect to the $\boldsymbol{\eta}_I$ coordinates as it seems intuitive at a first look. This is essential due to the dual affine structure of the space $\tilde{\mathcal{P}}$.

VII.5 Mixture of δ-flat families and singularities

The mixture of distributions has attracted a great attention in that it gives a wider exploration of the probability distributions space based on a simple parametric manifold. For instance, by the mixture of Gaussians (which belongs to a 0-flat family) we can approach any probability distribution in total variation norm. In this section, we study the general case of the mixture of δ flat families. The space can be defined as :

$$\begin{cases} \mathcal{Q} = \{ p_\theta \,|\, p_\theta = \sum_{j=1}^k w_j p_j(.\,; \boldsymbol{\theta}^j) \} \\ \\ p_j \in \mathcal{Q}_j, \quad \mathcal{Q}_j \text{ is } \delta \text{ flat} \end{cases}$$

where the manifolds $\mathcal{Q}_j$ are either distinct or not.

The mixture distribution can be viewed as an incomplete model where the weighted sum is considered as a marginalization over the hidden variable z representing the label of the mixture. Thus $p_\theta = \sum_z p(z) p(x \,|\, z, \boldsymbol{\theta}_z)$ and the weights $p(z)$ are the parameters of a mixture family. We consider now the statistical learning problem within the mixture family. A mixture of δ flat families is not, in general, δ flat. Therefore the δ optimal estimates have no more a simple expression. However, with data augmentation procedure we can construct

iterative algorithms computing the solution. Here, we focus on the computation of the δ prior of the mixture density.

The δ prior has the following expression :

$$\Pi(\boldsymbol{\theta}) \propto e^{-\frac{\gamma_e}{\gamma_u} D_\delta(p_\theta, p_0)} \sqrt{g(\boldsymbol{\theta})} \tag{VII.8}$$

The mixture (marginalization) form of the distribution p_θ leads to a complex expression of the δ divergence and the determinant of the Fisher information. However, the computation of these expressions in the complete data distribution space [Rodríguez, 2001] is feasible and gives explicit formula. By complete data $\boldsymbol{y}$, we mean the union of the observed data $\boldsymbol{x}$ and the hidden data $\boldsymbol{z}$. Therefore, the divergence will be considered between complete data distributions :

$$D_\delta(p^c, p_0^c) = \frac{\int p^c}{1-\delta} + \frac{\int p_0^c}{\delta} - \frac{\int (p^c)^\delta (p_0^c)^{1-\delta}}{\delta(1-\delta)}$$

where p^c is the complete likelihood $p(x,\, z \,|\, \boldsymbol{\theta})$ and $\boldsymbol{\theta}$ includes the parameters of the conditionals $p(x \,|\, z, \theta_z)$ and the discrete probabilities $p(z)$.

The additivity property of the δ-divergence is not conserved unless δ is equal to 0 or 1 [Amari, 1985] :

$$D_\delta(p_1 p_2, q_1 q_2) = D_\delta(p_1, q_1) + D_\delta(p_2, q_2) -$$

$$\delta\,(1-\delta) D_\delta(p_1, q_1) D_\delta(p_2, q_2)$$

Consequently, in the special case of $\delta \in \{0,\, 1\}$, we have the following simple formula :

$$\begin{cases} D_0(p, p_0) = \sum_{j=1}^k w_j^0 \left[D_0(p_j, p_j^0) + \log \frac{w_j^0}{w_j} \right] \\ \\ D_1(p, p_0) = \sum_{j=1}^k w_j \left[D_1(p_j, p_j^0) + \log \frac{w_j}{w_j^0} \right] \end{cases}$$

[A] Singularities with mixture families

It is known that in learning the parameters of Gaussian mixture densities [Snoussi 2001] the maximum likelihood fails because of the degeneracy of the likelihood function to infinity when certain variances go to zero or certain covariance matrices approach the boundary of singularity. In [Snoussi et Mohammad-Djafari, 2001], there is an analysis of the occurrence of this situation in the multivariate Gaussian mixture case. In this section, we give a general condition leading to this problem of degeneracy occurring in the learning within the mixture of δ flat families.

Let $\mathcal{Q}$ a δ flat manifold and $[\theta_i]$ the natural affine coordinates and $[\eta_i]$ the dual affine coordinates. The two coordinate systems are related by Legendre transformation [Amari, 1985] :

$$\begin{cases} \frac{\partial \eta_j}{\partial \theta_i} = g_{ij} \quad \frac{\partial \theta_i}{\partial \eta_j} = g_{ij}^{-1} \\ \\ \partial_i \psi = \eta_i \quad \partial_i \phi = \theta_i \end{cases}$$

where $(g_{ij})_{i=1..n}^{j=1..n}$ is the Fisher matrix and ψ and ϕ are the dual potentials.

It is clear from the expression of the variable transformation between the two affine coordinates that a singularity of the Fisher information matrix g leads to non differentiability in the transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. A singularity of g means that the determinant of this matrix is zero. Therefore, it is interesting to study the behaviour of the dual divergence at the boundary of singularity and we will show in an example that the dual divergences may have different behaviour as the distribution p approaches the boundary of singularity.

To illustrate such behaviour, we take a Gaussian family $\{\mathcal{N}(\mu,\, \sigma^2) \,|\, \mu \in \mathbb{R},\, \sigma \in \mathbb{R}_+\}$ which is a 2-dimensional statistical manifold 0-flat. The 0-affine coordinates are $\boldsymbol{\theta}$ and the 1-affine coordinates are $\boldsymbol{\eta}$

given by the following expressions :

$$\begin{cases} \theta_1 = \frac{\mu}{\sigma^2}, & \theta_2 = \frac{-1}{2\,\sigma^2} \\ \eta_1 = \mu, & \eta_2 = \mu^2 + \sigma^2 \end{cases} \tag{VII.9}$$

The corresponding Fisher information are :

$$|g(\theta)| \propto \sigma^6, \quad |g(\eta)| \propto 1/\sigma^6 \tag{VII.10}$$

The canonical divergence has the following expression :

$$D_\delta(p_1, p_2) = D_{1-\delta}(p_2, p_1) = \psi(p_1) + \phi(p_2) - \theta_i(p_1)\,\eta_i(p_2) \tag{VII.11}$$

where ψ and ϕ are the potentials given by :

$$\psi = \frac{\mu^2}{2\,\sigma^2} + \log\sqrt{2\pi}\sigma, \quad \phi = \frac{-1}{2} - \log\sqrt{2\pi}\sigma \tag{VII.12}$$

We see that the degeneracy occurs when the variance σ goes to zero. A detailed study of how this degeneracy occurs in the Gaussian mixture case is in [Snoussi et Mohammad-Djafari, 2001] and is reviewed in the example of the next section. Here we focus on the difference of behaviour of the two canonical divergences D_0 and D_1.

The expression of the δ prior is :

$$\Pi_\delta \propto e^{-D_\delta(p_\theta, p_0)}\sqrt{g(\theta)}$$

Following the complete data procedure :

$$\begin{cases} \Pi_0 \propto e^{-\frac{\gamma_e}{\gamma_u}\sum w_{i0}\{D_0(p_\theta^i, p_0^i) + \log\frac{w_{i0}}{w_i}\}}\sqrt{g(\boldsymbol{\theta}, \boldsymbol{w})} \\ \Pi_1 \propto e^{-\frac{\gamma_e}{\gamma_u}\sum w_i\{D_1(p_\theta^i, p_0^i) + \frac{w_i}{w_{i0}}\}}\sqrt{g(\boldsymbol{\theta}, \boldsymbol{w})} \end{cases}$$

The resulting prior is factorized and separated into independent priors on the components of the Gaussian mixture. Combining expressions of (VII.9), (VII.10), (VII.11) and (VII.12) we note the following comparison of the 0 and 1 priors through their dependences on the variance σ_j :

$\delta = 0$	$\delta = 1$
$\downarrow$	$\downarrow$
$p \longrightarrow \partial\mathcal{Q}$	$p \longrightarrow \partial\mathcal{Q}$
Π_0 is $O(\sigma_j^\alpha e^{-k_0/\sigma_j^2})$	Π_1 is $O(\sigma_j^{2\,w_j\,\frac{\gamma_j}{\gamma_u}})$
$\downarrow$	$\downarrow$
Exponential	Polynomial

where α, k_0 are constant.

We note that :

- For $\delta = 0$, the prior decreases to 0 when p approaches the boundary of singularity $\partial\,\mathcal{Q}$ with an **exponential** term leading to an inverse Gamma prior for the variance.
- For $\delta = 1$, the prior decreases to 0 when p approaches the boundary of singularity $\partial\,\mathcal{Q}$ with a **polynomial** term leading to a Gamma prior for the variance. We note the presence of the parameter w_i in the power term.

This kind of behaviour pushes us to use the 0 prior in that it is able to eliminate the degeneracy of the likelihood function.

VII.6 Examples

In this section we develop the δ prior in 2 learning problems : Multivariate Gaussian mixture and blind source separation and segmentation.

VII.6.1 Multivariate Gaussian mixture

The multivariate Gaussian mixture distribution of $\boldsymbol{x} \in \mathbb{R}^n$ is :

$$p(\boldsymbol{x}_i) = \sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{x}_i \,;\, \boldsymbol{m}_k, \boldsymbol{R}_k) \tag{VII.13}$$

where w_k, $\boldsymbol{m}_k$ and $\boldsymbol{R}_k$ are the weight, mean and covariance of the cluster k. This can be interpreted as an incomplete data problem where the missing data are the labels $(z_i)_{i=1..T}$ of the clusters. Therefore, the mixture (VII.13) is considered as a marginalization over z :

$$p(\boldsymbol{x}_i) = \sum_{z_i} p(z_i)\, \mathcal{N}(\boldsymbol{x}_i \,|\, z_i, \boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ is the set of the unknown means and covariances. Our objective is the prediction of the future observations given the trained data $\boldsymbol{x}_i$, $i = 1..T$. The whole parameter characterizing the statistical model is $\boldsymbol{\eta} = (\boldsymbol{\theta}, \boldsymbol{w})$. We consider now the derivation of the δ prior for $\delta \in \{0\,,\, 1\}$ and compare the two resulting priors.

The δ prior has the following form :

$$\Pi_\delta(\boldsymbol{\eta}) \propto e^{-\frac{\gamma_e}{\gamma_u} D_\delta(p_{\boldsymbol{\eta}}, p_0)} \sqrt{g(\boldsymbol{\eta})}$$

Therefore, we have to compute the D_δ divergence and the Fisher information matrix. As noted in the previous section and following [Rodríguez, 2001], the computation is considered in the complete data space $(\mathcal{X} \times \mathcal{Z})^T$ of observations $\boldsymbol{x}_i$ and labels z_i, T is the number of observations. In fact, we mean the number of virtual observations as the construction of the prior precedes the real observations. We have :

$$\begin{cases} D_0(\boldsymbol{\eta} : \boldsymbol{\eta}^0) = \underset{\boldsymbol{x}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\eta}_0}{E} \left[\log \frac{p(\boldsymbol{x}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\eta}_0)}{p(\boldsymbol{x}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\eta})} \right] \\ D_1(\boldsymbol{\eta} : \boldsymbol{\eta}^0) = \underset{\boldsymbol{x}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\eta}}{E} \left[\log \frac{p(\boldsymbol{x}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\eta})}{p(\boldsymbol{x}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\eta}_0)} \right] \\ g_{ij}(\boldsymbol{\eta}) = - \underset{\boldsymbol{x}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\eta}}{E} \left[\frac{\partial^2}{\partial_i \partial_j} \log p(\boldsymbol{x}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\eta}) \right] \end{cases}$$

By classifying the labels $\boldsymbol{z}_{1..T}$ and using the sequential Bayes rule between $\boldsymbol{x}_{1..T}$ and $\boldsymbol{z}_{1..T}$, the δ divergences become :

$$\begin{cases} D_0(\boldsymbol{\eta} : \boldsymbol{\eta}^0) = T \sum_{i=1}^{k} w_i^0 \left(D_0(\mathcal{N}_i : \mathcal{N}_i^0) + \log \frac{w_i^0}{w_i} \right) \\ D_1(\boldsymbol{\eta} : \boldsymbol{\eta}^0) = T \sum_{i=1}^{k} w_i \left(D_1(\mathcal{N}_i : \mathcal{N}_i^0) + \log \frac{w_i}{w_i^0} \right) \end{cases}$$

where $D_0(\mathcal{N}_i : \mathcal{N}_i^0) = D_1(\mathcal{N}_i^0 : \mathcal{N}_i)$ is the 0 divergence between two multivariate Gaussians :

$$\begin{cases} D_0(\mathcal{N}_i \,\|\, \mathcal{N}_i^0) = \frac{1}{2} \left(\log |\boldsymbol{R}_i \boldsymbol{R}_{i0}^{-1}| + \operatorname{Tr}\left(\boldsymbol{R}_{i0} \boldsymbol{R}_i^{-1} \right) - n + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i0})^* \boldsymbol{R}_i^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i0}) \right) \\ D_1(\mathcal{N}_i \,\|\, \mathcal{N}_i^0) = D_0(\mathcal{N}_i^0 \,\|\, \mathcal{N}_i) \end{cases}$$

The Fisher matrix is block diagonal with K diagonal blocks corresponding to the components of the mixture. Each block g_i with size $(n + n^2 + 1)$ has also a diagonal form (n is the dimension of the vector $\boldsymbol{x}_t$) :

$$g = \begin{bmatrix} [g_1] & & \\ & \ddots & \\ & & [g_K] \end{bmatrix} \quad , \quad g_i = \begin{bmatrix} w_i\, g_{\mathcal{N}}(\boldsymbol{m}_i,\, \boldsymbol{R}_i) & [0] \\ [0] & 1/w_i \end{bmatrix}$$

where $g_{\mathcal{N}}$ is the Fisher matrix of the multivariate Gaussian and has the following expression :

$$g_{\mathcal{N}}(\boldsymbol{m}, \boldsymbol{R}) = \begin{bmatrix} \boldsymbol{R}^{-1} & [0] \\ [0] & -\frac{1}{2}\frac{\partial \boldsymbol{R}^{-1}}{\partial \boldsymbol{R}} \end{bmatrix}$$

whose determinant is :

$$|g_{\mathcal{N}}(\boldsymbol{m}, \boldsymbol{R})| = |\boldsymbol{R}|^{-(n+2)}$$

Thus, the determinant of the block g_i is :

$$|g_i(w_i, \boldsymbol{m}_i, \boldsymbol{R}_i)| = \left(\frac{1}{2}\right)^{n^2} w_i^{(n^2+n-1)} |\boldsymbol{R}_i|^{-(n+2)} \tag{VII.14}$$

The additional form of the $\{0\,,\,1\}$ divergences (implying the multiplicative form of their exponentials) and the multiplicative form of the determinant of the Fisher matrix (due to its block diagonal form) lead to an independent priors of the components $\boldsymbol{\eta}_i = (w_i, \boldsymbol{m}_i, \boldsymbol{R}_i)$: $\Pi(\boldsymbol{\eta}) = \prod_{k=1}^{K} \Pi(\boldsymbol{\eta}_i)$. The two values of $\delta = \{0\,,\,1\}$ lead to two different priors Π_δ :

– $\delta = 0$:

$$\begin{aligned} \Pi_0(\boldsymbol{\eta}_i) &\propto \exp\left[-\tfrac{\gamma_e}{\gamma_u}\left(w_i^0 D_0(\mathcal{N}_i : \mathcal{N}_i^0) + w_i^0 \log \tfrac{w_i^0}{w_i}\right)\right] \sqrt{|g_i(\boldsymbol{\eta}_i)|} \\ &\propto \mathcal{N}\left(\boldsymbol{m}_i\,;\,\boldsymbol{m}_0, \tfrac{\boldsymbol{R}_i}{\alpha\, w_i^0}\right) \mathcal{W}_n\left(\boldsymbol{R}_i^{-1}\,;\,\nu_0, \boldsymbol{R}_0^{-1}\right) w_i^{\beta_0} \end{aligned} \tag{VII.15}$$

with,

$$\alpha = \tfrac{\gamma_e}{\gamma_u}, \quad \nu_0 = \alpha\, w_i^0, \quad \beta_0 = \alpha w_i^0 + \tfrac{n^2+n-1}{2}$$

$\mathcal{W}_n$ is the wishart distribution of an $n \times n$ matrix :

$$\mathcal{W}_n(\boldsymbol{R}\,;\,\nu, \boldsymbol{\Sigma}) \propto |\boldsymbol{R}|^{\frac{\nu-(n+1)}{2}} \exp\left[-\frac{\nu}{2}\mathrm{Tr}\left(\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\right)\right]$$

The 0-prior is Normal Inverse Wishart for the mean and covariance $(\boldsymbol{m}_i, \boldsymbol{R}_i)$ and Dirichlet for the weight w_i, that is the **conjugate** prior.

– $\delta = 1$:

$$\begin{aligned} \Pi_1(\boldsymbol{\eta}_i) &\propto \exp\left[-\tfrac{\gamma_e}{\gamma_u}\left(w_i D_1(\mathcal{N}_i : \mathcal{N}_i^0) + w_i \log \tfrac{w_i}{w_i^0}\right)\right] \sqrt{|g_i(\boldsymbol{\eta}_i)|} \\ &\propto \mathcal{N}\left(\boldsymbol{m}_i\,;\,\boldsymbol{m}_0, \tfrac{\boldsymbol{R}_i}{\alpha\, w_i}\right) \mathcal{W}_n\left(\boldsymbol{R}_i\,;\,\alpha w_i - 1, \tfrac{\alpha w_i - 1}{\alpha w_i}\boldsymbol{R}_0\right) \\ &\quad w_i^{\frac{n^2+n-1}{2} - (1+\frac{n}{2})\alpha w_i} \left(w_i^0\right)^{\alpha w_i} \Gamma_n\left(\tfrac{\alpha w_i - 1}{2}\right) \end{aligned} \tag{VII.16}$$

where Γ_n is the generalized Gamma function of dimension n ([Box et Tiao, 1972] page 427) :

$$\Gamma_n(b) = \left[\Gamma(\frac{1}{2})\right]^{\frac{1}{2}n(n-1)} \prod_{i=1}^{n} \Gamma(b + \frac{i-n}{2}), \quad b > \frac{n-1}{2}$$

The 1-prior Π_1 (VII.16) is the generalized entropic prior [Rodríguez, 2001] to the multivariate case. We see that the prior Π_1 is a **Wishart** function of the covariance matrices $\boldsymbol{R}_i$ and the prior Π_0 is an **inverse Wishart** function of the covariances. This leads to a difference of the behaviour of these functions on the boundary of singularity (the set of singular matrices).

VII.6.2 Source separation

The second example deals with the source separation problem. The observations $\boldsymbol{x}_{1..T}$ are T samples of m-vectors. At each time t, the vector data $\boldsymbol{x}_t$ is supposed to be a noisy instantaneous mixture of an observed

n-vector source $\boldsymbol{s}_t$ with unknown mixing coefficients forming the mixing matrix $\boldsymbol{A}$. This is simply modeled by the following equation :

$$\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{s}_t + \boldsymbol{n}_t,\ t = 1..T$$

where given the data $\boldsymbol{x}_{1..T}$, our objective is the recovering of the original sources $\boldsymbol{s}_{1..T}$ and the unknown matrix $\boldsymbol{A}$. The Bayesian approach taken to solve this inverse problem [Knuth, 1999; Mohammad-Djafari, 1999; Snoussi et Mohammad-Djafari, 2002b] needs also the estimation of the noise covariance matrix $\boldsymbol{R}_n$ and the learning of the statistical parameters of the original sources $\boldsymbol{s}_{1..T}$. In the following, we suppose that the sources are statistically independent and that each source is modeled by a mixture of univariate Gaussians, so that we have to learn each set of source j parameters $\boldsymbol{\eta}^j$ which contains the weights, means and variances composing the mixture j :

$$\begin{cases} \boldsymbol{\eta}^j = \left(\boldsymbol{\eta}_i^j\right)_{i=1..K_j} \\ \\ \boldsymbol{\eta}_i^j = (w_i^j, m_i^j, \sigma_i^j) \end{cases}$$

The index j indicates the source j and i indicates the Gaussian component i of the distribution of the source j. Therefore we don't have a multidimensional Gaussian mixture but instead independent unidimensional Gaussian mixtures.

In the following, our parameter of interest is $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{R}_n, \boldsymbol{\eta})$: the mixing matrix $\boldsymbol{A}$, the noise covariance $\boldsymbol{R}_n$ and $\boldsymbol{\eta}$ contains all the parameters of the sources model. Our objective is the computation of the δ priors for $\delta \in \{0\,,\,1\}$. We have an incomplete data problem with two hierarchies of hidden variables, the sources $\boldsymbol{s}_{1..T}$ and the labels $\boldsymbol{z}_{1..T}$ so that the complete data are $(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T})$. We begin by the computation of the Fisher information matrix which is common to the both geometries.

a/ Fisher information matrix

The Fisher matrix $\mathcal{F}(\boldsymbol{\theta})$ is defined as :

$$\mathcal{F}_{ij}(\boldsymbol{\theta}) = -\underset{\boldsymbol{x}_{1..T},\boldsymbol{s}_{1..T},\boldsymbol{z}_{1..T}}{E}\left[\frac{\partial^2}{\partial_i \partial_j}\log p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta})\right]$$

The factorization of the joint distribution $p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta})$ as :

$$p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta}) = p(\boldsymbol{x}_{1..T} \,|\, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T}, \theta)\, p(\boldsymbol{s}_{1..T} \,|\, \boldsymbol{z}_{1..T}, \theta)\, p(\boldsymbol{z}_{1..T} \,|\, \theta)$$

and the corresponding expectations as

$$\underset{\boldsymbol{x}_{1..T},\boldsymbol{s}_{1..T},\boldsymbol{z}_{1..T}}{E}[.] = \underset{\boldsymbol{z}_{1..T}}{E}[.]\ \underset{\boldsymbol{s}_{1..T}\,|\,\boldsymbol{z}_{1..T}}{E}[.]\ \underset{\boldsymbol{x}_{1..T}\,|\,\boldsymbol{s}_{1..T},\boldsymbol{z}_{1..T}}{E}[.]$$

and taking into account the conditional independencies $((\boldsymbol{x}_{1..T} \,|\, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T}) \Leftrightarrow (\boldsymbol{x}_{1..T} \,|\, \boldsymbol{s}_{1..T})$ and $(\boldsymbol{s}_{1..T} \,|\, \boldsymbol{z}_{1..T}) \Leftrightarrow \prod \boldsymbol{s}_{1..T}^j \,|\, \boldsymbol{z}_{1..T}^j)$, the Fisher information matrix will have a block diagonal structure as follows :

$$g(\boldsymbol{\theta}) = \begin{bmatrix} g(\boldsymbol{A}, \boldsymbol{R}_n) & \dots & [0] \\ \vdots & g(\boldsymbol{\eta}^1) & \\ & & \ddots \\ [0] & & \dots \quad g(\boldsymbol{\eta}^n) \end{bmatrix}$$

a.1/ $(\boldsymbol{A}, \boldsymbol{R}_n)$-block

The Fisher information matrix of $(\boldsymbol{A}, \boldsymbol{R}_n)$ is :

$$\mathcal{F}_{ij}(\boldsymbol{A}, \boldsymbol{R}_n) = -\underset{\boldsymbol{s}}{E}\ \underset{\boldsymbol{x}\,|\,\boldsymbol{s}}{E}\left[\frac{\partial^2}{\partial_i \partial_j}\log p(\boldsymbol{x}_{1..T} \,|\, \boldsymbol{s}_{1..T}, \boldsymbol{A}, \boldsymbol{R}_n)\right]$$

which is very similar to the Fisher information matrix of the mean and covariance of a multivariate Gaussian distribution. The obtained expression is

$$g(\boldsymbol{A}, \boldsymbol{R}_n) = \begin{bmatrix} \left(\underset{\boldsymbol{s}_{1..T}}{E} \boldsymbol{R}_{ss} \right) \otimes \boldsymbol{R}_n^{-1} & [0] \\ [0] & -\frac{1}{2} \frac{\partial \boldsymbol{R}_n^{-1}}{\partial \boldsymbol{R}_n} \end{bmatrix}$$

where $\boldsymbol{R}_{ss} = \frac{1}{T} \sum \boldsymbol{s}_t \boldsymbol{s}_t^*$ and $\otimes$ is the Kronecker product.

We note the block diagonality of the $(\boldsymbol{A}, \boldsymbol{R}_n)$-Fisher matrix. The term corresponding to the mixing matrix $\boldsymbol{A}$ is the signal to noise ratio as can be expected. Thus, the amount of information about the mixing matrix is proportional to the signal to noise ratio. The induced volume of $(\boldsymbol{A}, \boldsymbol{R}_n)$ is then :

$$|g(\boldsymbol{A}, \boldsymbol{R}_n)|^{1/2} \, d\,\boldsymbol{A} \, d\boldsymbol{R}_n = \frac{|\underset{\boldsymbol{\eta}}{E} \boldsymbol{R}_{ss}|^{m/2}}{|\boldsymbol{R}_n|^{\frac{m+n+1}{2}}} d\,\boldsymbol{A} \, d\boldsymbol{R}_n$$

a.2/ $(\boldsymbol{\eta}^j)$**-block**

Each $g(\boldsymbol{\eta}^j)$ is the Fisher information of a one-dimensional Gaussian distribution. Therefore, it is obtained by setting $n = 1$ in the expression (VII.14) of the previous section :

$$\left|g(\boldsymbol{\eta}^j)\right|^{1/2} d\,\boldsymbol{\eta}^j = \left\{ \prod_{i=1}^{K_j} \frac{w_i^{1/2}}{v_i^{3/2}} \right\} d\,\boldsymbol{\eta}^j$$

b/ δ**-Divergence (**$\delta = 0,\, 1$**)**

The δ-divergence between two parameters $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{R}_n, \boldsymbol{\eta})$ and $\boldsymbol{\theta}^0 = (\boldsymbol{A}^0, \boldsymbol{R}_n^0, \boldsymbol{\eta}^0)$ for the complete data likelihood $p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta})$ is :

$$\begin{cases} D_0(\boldsymbol{\theta} : \boldsymbol{\theta}^0) = \underset{x,s,z|\theta^0}{E} \log \frac{p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta}^0)}{p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta})} \\ \\ D_1(\boldsymbol{\theta} : \boldsymbol{\theta}^0) = \underset{x,s,z|\theta}{E} \log \frac{p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta})}{p(\boldsymbol{x}_{1..T}, \boldsymbol{s}_{1..T}, \boldsymbol{z}_{1..T} \,|\, \boldsymbol{\theta}^0)} \end{cases}$$

Similar developments of the above equation as in the computation of the Fisher matrix based on the conditional independencies, lead to an affine form of the divergence, which is a sum of the expected divergence between the $(\boldsymbol{A}, \boldsymbol{R}_n)$ parameters and the divergence between the sources parameters $\boldsymbol{\eta}$:

$$\begin{cases} D_0(\boldsymbol{\theta} : \boldsymbol{\theta}^0) = \underset{s|\eta^0}{E} \underset{|s}{D_0}(\boldsymbol{A}, \boldsymbol{R}_n \,:\, \boldsymbol{A}^0, \boldsymbol{R}_n^0) + D_0(\boldsymbol{\eta} \,:\, \boldsymbol{\eta}^0) \\ \\ D_1(\boldsymbol{\theta} : \boldsymbol{\theta}^0) = \underset{s|\eta}{E} \underset{|s}{D_1}(\boldsymbol{A}, \boldsymbol{R}_n \,:\, \boldsymbol{A}^0, \boldsymbol{R}_n^0) + D_1(\boldsymbol{\eta} \,:\, \boldsymbol{\eta}^0) \end{cases}$$

where $\underset{|s}{D_\delta}$ means the divergence between the distributions $p(\boldsymbol{x}_{1..T} \,|\, \boldsymbol{A}, \boldsymbol{R}_n, \boldsymbol{s}_{1..T})$ and $p(\boldsymbol{x}_{1..T} \,|\, \boldsymbol{A}^0, \boldsymbol{R}_n^0, \boldsymbol{s}_{1..T})$ keeping the sources $\boldsymbol{s}_{1..T}$ fixed.

The δ-divergence between $\boldsymbol{\eta}$ and $\boldsymbol{\eta}_0$ is the sum of the δ-divergences between each source parameter $\boldsymbol{\eta}^j$ and $\boldsymbol{\eta}_0^j$ due to the *a priori* independence between the sources. Then, the divergence between $\boldsymbol{\eta}^j$ and $\boldsymbol{\eta}_0^j$ is obtained as a particular case ($n = 1$) of the general expression derived in the multivariate case. Therefore we have the same form of the prior as in equations (VII.15) and (VII.16).

The expressions of the averaged divergences between the $(\boldsymbol{A}, \boldsymbol{R}_n)$ parameters are :

$$\left\{\begin{array}{ll} \underset{s|\eta^0}{E}\, D_0(\boldsymbol{A}, \boldsymbol{R}_n \,:\, \boldsymbol{A}_0, \boldsymbol{R}_{n0}) = & \frac{1}{2}\left(\log\left|\boldsymbol{R}_n \boldsymbol{R}_{n0}^{-1}\right| + \mathrm{Tr}\left(\boldsymbol{R}_n^{-1}\boldsymbol{R}_{n0}\right)\right. \\ & \left. +\mathrm{Tr}\left(\boldsymbol{R}_n^{-1}(\boldsymbol{A}-\boldsymbol{A}_0)\underset{s|\eta^0}{E}[\boldsymbol{R}_{ss}](\boldsymbol{A}-\boldsymbol{A}_0)^*\right)\right) \\ \underset{s|\eta}{E}\, D_1(\boldsymbol{A}, \boldsymbol{R}_n \,:\, \boldsymbol{A}_0, \boldsymbol{R}_{n0}) = & \frac{1}{2}\left(\log\left|\boldsymbol{R}_{n0} \boldsymbol{R}_{n}^{-1}\right| + \mathrm{Tr}\left(\boldsymbol{R}_{n0}^{-1}\boldsymbol{R}_{n}\right)\right. \\ & \left. +\mathrm{Tr}\left(\boldsymbol{R}_{n0}^{-1}(\boldsymbol{A}-\boldsymbol{A}_0)\underset{s|\eta}{E}[\boldsymbol{R}_{ss}](\boldsymbol{A}-\boldsymbol{A}_0)^*\right)\right) \end{array}\right.$$

leading to the following δ priors on $(\boldsymbol{A}, \boldsymbol{R}_n)$:

$$\left\{\begin{array}{lll} \Pi_0(\boldsymbol{A}, \boldsymbol{R}_n^{-1}) & \propto & \mathcal{N}\left(\boldsymbol{A}\,;\, \boldsymbol{A}_0, \frac{1}{\alpha}{\boldsymbol{R}_{ss}^0}^{-1} \otimes \boldsymbol{R}_n\right) \mathcal{W}_{im}\left(\boldsymbol{R}_n^{-1}\,;\, \alpha, {\boldsymbol{R}_n^0}^{-1}\right) |\underset{s|\eta}{E}[\boldsymbol{R}_{ss}]|^{\frac{m}{2}} \\ \Pi_1(\boldsymbol{A}, \boldsymbol{R}_n) & \propto & \mathcal{N}\left(\boldsymbol{A}\,;\, \boldsymbol{A}_0, \frac{1}{\alpha}\underset{s|\eta}{E}[\boldsymbol{R}_{ss}]^{-1} \otimes \boldsymbol{R}_n^0\right) \mathcal{W}_{im}\left(\boldsymbol{R}_n\,;\, \alpha - n, \frac{\alpha - n}{\alpha}\boldsymbol{R}_n^0\right) \end{array}\right.$$

Therefore, the 0-prior is a normal inverse Wishart prior (conjugate prior). The mixing matrix and the noise covariance are not *a priori* independent. In fact, the covariance matrix of $\boldsymbol{A}$ is the noise to signal ratio $\frac{1}{\alpha}{\boldsymbol{R}_{ss}^0}^{-1} \otimes \boldsymbol{R}_n$. We note a multiplicative term which is a power of the determinant of the *a priori* expectation of the source covariance $\underset{s|\eta}{E}[\boldsymbol{R}_{ss}]$. This term can be injected in the prior $p(\boldsymbol{\eta})$ and thus the $(\boldsymbol{A}, \boldsymbol{R}_n)$ parameters and the $\boldsymbol{\eta}$ parameters are *a priori* independent.

The 1-prior (entropic prior) is normal Wishart. The mixing matrix and the noise covariance are *a priori* independent since the noise to signal ratio $\frac{1}{\alpha}\underset{s|\eta}{E}[\boldsymbol{R}_{ss}]^{-1} \otimes \boldsymbol{R}_n^0$ depend on the reference parameter $\boldsymbol{R}_n^0$. However, we have in counterpart the dependence of $\boldsymbol{A}$ and $\boldsymbol{\eta}$ through the term $\underset{s|\eta}{E}[\boldsymbol{R}_{ss}]^{-1}$ present in the covariance matrix of $\boldsymbol{A}$. In practice, we prefer to replace the expected covariance $\underset{s|\eta}{E}[\boldsymbol{R}_{ss}]$, in the two priors, by its reference value $\boldsymbol{R}_{ss}^0$.

We note that the precision matrix for the mixing matrix $\boldsymbol{A}$ ($\alpha \boldsymbol{R}_{ss}^0 \otimes \boldsymbol{R}_n^{-1}$ for Π_0 and $\alpha \underset{s|\eta}{E}[\boldsymbol{R}_{ss}] \otimes {\boldsymbol{R}_n^0}^{-1}$ for Π_1) is the product of the confidence term $\alpha = \frac{\gamma_e}{\gamma_u}$ in the reference parameters and the signal to noise ratio. Therefore, the resulting precision of the reference matrix $\boldsymbol{A}_0$ is not only our *a priori* coefficient γ_e but the product of this coefficient and the signal to noise ratio.

VII.7 Conclusion and discussion

In this paper, we have shown the importance of providing a geometry (a measure of distinguishibility) to the space of distributions. A different geometry will give a different learning rule mapping the training data to the space of predictive distributions. The prior selection procedure established in a statistical decision framework needs to be taken in a specified geometry. We have tried to elucidate the interaction between the parametric and non parametric modeling. The notion of "projected mass" gives to the restricted parametric modelization a non parametric sense and shows the role of the relative geometry of the parametric model in the whole space of distributions. The same investigations are considered in the interaction between a curved family and the whole parametric model containing it. Exact expressions are shown in a simple case of auto-parallel families and we are working on the more abstract space of distributions.

References

[Amari, 1985] S. Amari. *Differential-Geometrical Methods in Statistics*. Volume 28 of Springer Lecture Notes in Statistics, Springer-Verlag, New York, 1985.

[Amari et Nagaoka, 2000] S. Amari et H. Nagaoka. *Methods of Information Geometry*, volume 191 of Translations of Mathematical Monographs. AMS, OXFORD, University Press, 2000.

[Box et Tiao, 1972] G. E. P. Box et G. C. Tiao. *Bayesian inference in statistical analysis*. Addison-Wesley publishing, 1972.

[Kass et Wasserman, 1994] R. E. Kass et L. Wasserman. Formal rules for selecting prior distributions : A review and annotated bibliography. Technical report no. 583, Department of Statistics, Carnegie Mellon University, 1994.

[Knuth, 1999] K. Knuth. A Bayesian approach to source separation. In *Proceedings of Independent Component Analysis Workshop*, pages 283–288, 1999.

[Mohammad-Djafari, 1999] A. Mohammad-Djafari. A Bayesian approach to source separation. In J. R. G. Erikson et C. Smith, éditeurs, *Bayesian Inference and Maximum Entropy Methods*, Boise, IH, USA, juillet 1999. MaxEnt Workshops, Amer. Inst. Physics.

[Rodríguez, 1991] C. Rodríguez. Entropic priors. *Tech. rep. Electronic form* `http : omega.albany.edu :8008/entpriors.ps`, 1991.

[Rodríguez, 2001] C. Rodríguez. Entropic priors for discrete probabilistic networks and for mixtures of Gaussians models. In R. L. FRY, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 410–432. MaxEnt Workshops, Amer. Inst. Physics, août 2001.

[Snoussi et Mohammad-Djafari, 2001] H. Snoussi et A. Mohammad-Djafari. Penalized maximum likelihood for multivariate gaussian mixture. In R. L. Fry, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 36–46. MaxEnt Workshops, Amer. Inst. Physics, août 2001.

[Snoussi et Mohammad-Djafari, 2002a] H. Snoussi et A. Mohammad-Djafari. Information Geometry and Prior Selection. In C. Williams, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 307–327. MaxEnt Workshops, Amer. Inst. Physics, août 2002.

[Snoussi et Mohammad-Djafari, 2002b] H. Snoussi et A. Mohammad-Djafari. MCMC Joint Separation and Segmentation of Hidden Markov Fields. In *Neural Networks for Signal Processing XII*, pages 485–494. IEEE workshop, septembre 2002.

[Zhu et Rohwer, 1995a] H. Zhu et R. Rohwer. Bayesian invariant measurements of generalisation. In *Neural Proc. Lett.*, volume 2 (6), pages 28–31, 1995.

[Zhu et Rohwer, 1995b] H. Zhu et R. Rohwer. Bayesian invariant measurements of generalisation for continuous distributions. Technical report, NCRG/4352, ftp ://cs.aston.ac.uk/neural/zhuh/continuous.ps.z, Aston University, 1995.

Chapitre VIII

CONCLUSION ET PERSPECTIVES

L'un des points traités dans ce mémoire est l'utilisation de la non stationnarité pour séparer les sources. Dans le cas des images, ceci consiste à effectuer la ségmentation conjointement avec la séparation. Nous avons vu que l'approche bayésienne offre un cadre naturel à l'incorporation de l'étape de la ségmentation. Nous allons maintenant proposer une solution basée sur la maximisation de l'information mutuelle afin d'incorporer la ségmentation dans une approche informationnelle. La comparaison de l'approche entropique avec l'approche bayésienne peut être appréhendée grâce à la notion de dualité entre la logique des propositions et la logique des questions.

VIII.1 Sur la séparation et la ségmentation conjointes

Nous avons vu dans le chapitre (IV) que l'exploitation de la non stationnarité est un outil efficace pour la séparation d'images. On peut expliquer cette efficacité par les faits suivants.

1. La non stationnarité est bien adaptée à la modélisation des images réelles. Elle permet de modéliser l'homogénéité par régions. Chaque région est caractérisée statistiquement par sa propre densité de probabilité. Notons que la classification des pixels d'une image (ségmentation) présente un intérêt en soi comme c'est le cas en imagerie satellitaire (pour l'aide à la décision), en imagerie médicale (pour l'aide au diagnostic), en compression et transmission des images...
2. En tenant compte de la non stationnarité, on peut garantir l'identifiabilité de la matrice de mélange même lorsque les sources sont modélisées par des gaussiennes. Par ailleurs, la gaussiannité des sources nous permet d'avoir des expressions explicites et souvent linéaires offrant ainsi une bonne efficacité algorithmique.
3. L'exploitation de la non stationnarité est optimale lorsqu'on associe au problème de séparation la ségmentation automatique des images sources. Autrement dit, on ne fixe pas *a priori* la partition des pixels des images sources mais on essaie d'estimer cette partition conjointement avec la séparation des images. Heureusement, le problème de la ségmentation est de même nature que le problème de séparation : on sépare les régions en exploitant leur diversité statistique (voir figure (VIII.1)). Cette similarité facilite considérablement le traitement conjoint (conceptuel et algorithmique) des deux problèmes en les considérant tous les deux comme des problèmes à variables cachées. Les sources sont les variables cachées pour l'identification de la matrice de mélange et les étiquettes sont les variables cachées pour la modélisation des sources.

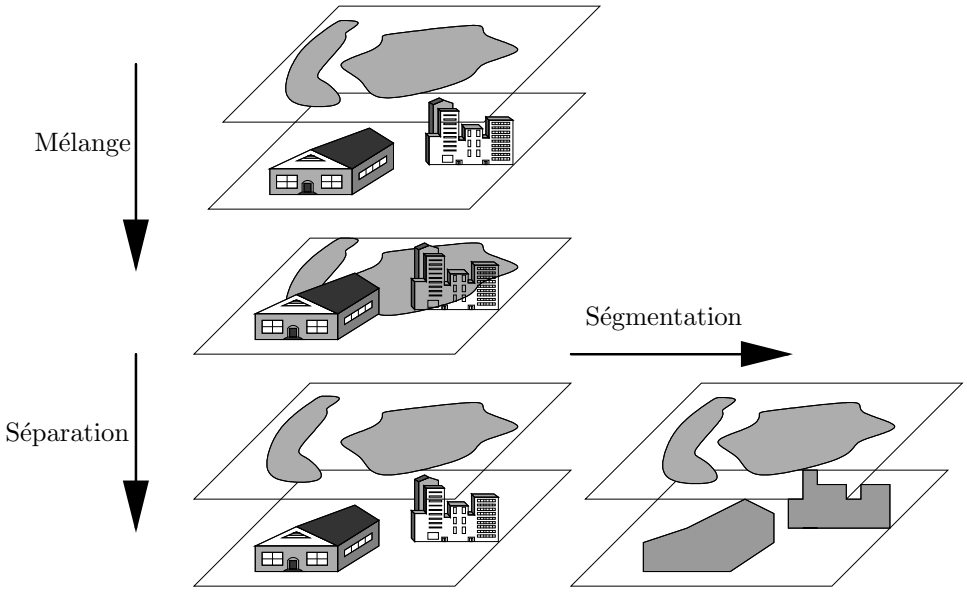


FIG. VIII.1: On distingue deux types de séparation : (**i**) une séparation transversale le long des capteurs, (**ii**) une séparation spatiale le long des pixels

VIII.1.1 Approche bayésienne

On introduit les variables $\boldsymbol{Z}$ représentant les étiquettes des pixels. D'un point de vue logique, ceci revient à expliquer virtuellement le processus qui a généré les sources $\boldsymbol{S}$. En notant $\boldsymbol{I}$ toute l'information *a priori* disponible, la proposition initiale ($\boldsymbol{I} \longrightarrow \boldsymbol{S}$) se transforme en ($\boldsymbol{I} \longrightarrow \boldsymbol{Z} \longrightarrow \boldsymbol{S}$) limitant ainsi le choix des probabilités subjectives des sources. Le problème direct d'obtention des observations $\boldsymbol{X}$ est décrit logiquement

par les propositions suivantes :

$$\boldsymbol{I} \longrightarrow \boldsymbol{Z} \longrightarrow \boldsymbol{S} \text{ puis } \boldsymbol{S} \wedge \boldsymbol{A} \longrightarrow \boldsymbol{X},$$

où $\boldsymbol{A}$ est la matrice de mélange (ou en général le processus de mélange).

Connaissant les données $\boldsymbol{X}$, on peut définir plusieurs problèmes d'inférence : ($\mathcal{I} = \boldsymbol{X} \longrightarrow \boldsymbol{A}$), ($\mathcal{I} = \boldsymbol{X} \longrightarrow \boldsymbol{S}$), ($\mathcal{I} = \boldsymbol{X} \longrightarrow \boldsymbol{Z}$),... Nous allons considérer le problème global :

$$\mathcal{I} = \boldsymbol{X} \longrightarrow \boldsymbol{A} \wedge \boldsymbol{S} \wedge \boldsymbol{Z}.$$

Le degré d'incertitude de la proposition $\mathcal{I}$ est la distribution *a posteriori* :

$$\mathrm{p}(\boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{A} \mid \boldsymbol{X}) \propto \mathrm{p}(\boldsymbol{X} \mid \boldsymbol{S}, \boldsymbol{A}) \times \mathrm{p}(\boldsymbol{S} \mid \boldsymbol{Z}) \times \mathrm{p}(\boldsymbol{Z}) \times \mathrm{p}(\boldsymbol{A})$$

Problème direct du mélange

Régularisation de la classification

a priori sur les cœfficients du mélange

Caractérisation statistique des régions

Nous avons considéré les deux points de vue pour le choix des probabilités.

1. Point de vue subjectif : ceci concerne la modélisation markovienne des étiquettes $\boldsymbol{Z}$ et la modélisation gaussienne des sources $\boldsymbol{S}$ connaissant les étiquettes. Ces choix sont basés sur des considérations pratiques. En effet, dans un cadre non stationnaire, l'identification de la matrice de mélange est souvent possible à partir des statistiques d'ordre deux. Concernant les étiquettes, l'*a priori* markovien est un moyen de régulariser la ségmentation et d'assurer sa robustesse dans un environnement aveugle et bruité.
2. Point de vue logique : nous avons essayé dans le dernier chapitre (VII) de construire des *a priori* en imposant des règles basées sur la théorie géométrique de la prédiction et sur la notion de l'ignorance.

Les techniques bayésiennes d'échantillonnage permettent la mise en œuvre de la séparation et de la ségmentation conjointe. En particulier, l'échantillonnage de Gibbs consiste à échantillonner alternativement les sources, les étiquettes et la matrice de mélange, fournissant, à la convergence, les échantillons des marginales $\mathrm{p}(\boldsymbol{S} \mid \boldsymbol{X})$ et $\mathrm{p}(\boldsymbol{Z} \mid \boldsymbol{X})$ et toute inférence peut être menée à partir de ces échantillons.

```
L'approche bayésienne représente ainsi un cadre naturel à l'incorporation
du problème de la ségmentation. En plus, elle offre les outils techniques
nécessaires pour l'implémentation conjointe de la séparation et de la ség-
mentation.
```

VIII.1.2 Approche InfoMAx

Nous allons essayer dans cette section de formuler le problème de la séparation et de la ségmentation simultanées à l'aide de la théorie de l'information. On suppose que le mélange est linéaire non bruité. On note $\boldsymbol{X} = (\boldsymbol{X}^i)_{i=1}^{i=m}$ l'ensemble des images observées, $\boldsymbol{B}$ la matrice séparatrice, $\boldsymbol{Y} = (\boldsymbol{Y}^j)_{j=1}^{j=n}$ les images

à reconstruire et $\boldsymbol{Z}$ la classification des images[1]. Afin de simplifier les développements qui vont suivre, on va classifier les pixels seulement en deux régions $\mathcal{R}_1$ et $\mathcal{R}_2 = \bar{\mathcal{R}}_1$ (voir figure (VIII.2)). Autrement dit, les étiquettes $\boldsymbol{Z}_r$ ne prennent que deux valeurs possibles 1 et 2.

A chaque pixel $r \in \mathcal{S}$:

$$\boldsymbol{y}_r = \boldsymbol{B}\boldsymbol{x}_r.$$

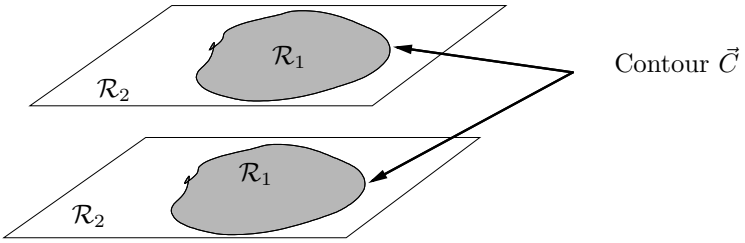


FIG. VIII.2: On suppose une même classification pour toutes les images en deux régions $\mathcal{R}_1$ et $\mathcal{R}_2$. Si $r \in \mathcal{R}_1$ alors Y_r^j suit la loi p_1 et si $r \in \mathcal{R}_2$ alors Y_r^j suit la loi p_2. Les deux régions sont délimitées par un contour $\vec{C}$.

[A] SÉPARATION

La séparation de sources peut être interprétée comme une analyse en composantes indépendantes. L'estimation de la matrice séparatrice $\boldsymbol{B}$ se base sur la minimisation de l'information mutuelle (conditionnelle à Z) entre les composantes Y^j.

$$\begin{aligned} \hat{\boldsymbol{B}} &= \arg\min_{\boldsymbol{B}} \left\{ I(Y^1, ..., Y^n \mid Z) \right\} \\ &= \arg\min_{\boldsymbol{B}} \left\{ Pr(Z=1) I(Y^1, ..., Y^n \mid Z=1) + Pr(Z=2) I(Y^1, ..., Y^n \mid Z=2) \right\} \end{aligned} \tag{VIII.1}$$

Le calcul de l'information mutuelle est approché par une somme empirique **conditionnellement** à la connaissance de la classification $\boldsymbol{Z}$. La résolution du problème d'optimisation (VIII.1) est similaire à celle dans [Pham et Cardoso, 2001; Pham, 2002]. Les densités p_1 et p_2 peuvent être choisies gaussiennes ou non gaussiennes.

D'une manière équivalente, on peut estimer $\boldsymbol{B}$ en maximisant le flux d'information entre $\boldsymbol{X}$ et une transformation non linéaire $\boldsymbol{V}$ de $\boldsymbol{Y}$: $\boldsymbol{V} = \Phi(\boldsymbol{B}\boldsymbol{X})$ (Φ s'applique terme à terme), [Bell et Sejnowski, 1995].

[B] SÉGMENTATION

Le point intéressant est que le problème de la ségmentation peut aussi se résoudre avec une approche informationnelle [Unal *et al.*, 2002; Kim *et al.*, 2002]. En effet, l'information mutuelle $I(Y(r), Z(r))$ entre l'intensité $Y(r)$ (on considère l'image multidimensionnelle) et l'étiquette $Z(r)$ du pixel r est maximale quand le champ $\boldsymbol{Z}$ reflète la vraie classification. Dans ce cas, la seule connaissance de la valeur de l'étiquette est suffisante à la détermination de la densité de l'intensité $Y(r)$ (si elle est p_1 ou p_2) et la localisation r du pixel n'apporte pas d'information supplémentaire. La recherche de la bonne classification est ainsi obtenue en maximisation le flux d'information entre les champs d'intensité $\boldsymbol{Y}$ et la classification $\boldsymbol{Z}$.

L'information mutuelle s'écrit,

$$\begin{aligned} I(Y, Z) &= \mathcal{H}(Y) - \mathcal{H}(Y \mid Z) \\ &= \mathcal{H}(Y) - Pr\,(Z=1)\mathcal{H}(Y \mid Z=1) - Pr\,(Z=2)\mathcal{H}(Y \mid Z=2) \end{aligned} \tag{VIII.2}$$

[1] on suppose une classification commune à toutes les images afin de simplifier les notations. La considération de plusieurs classifications indépendantes ne complique pas la solution finale et améliore les performances de la séparation.

où $\mathcal{H}(p_\xi)$ désigne l'entropie de p_ξ :

$$\mathcal{H}(\xi) = -\int p(\xi) \log\, p(\xi) d\,\xi.$$

Le terme $\mathcal{H}(Y)$ ne dépend pas de la classification. On va donc essayer d'approcher l'expression $\mathcal{H}(Y \mid Z)$ en se basant sur les données $\boldsymbol{Y}$ et $\boldsymbol{Z}$. On considère l'approximation suivante de l'entropie (proposée par [Hyvärinen, 1997]) :

$$\mathcal{H}(p) \approx \mathcal{H}(\nu) - \frac{1}{2}\sum_{l=1}^{L} {e_l}^2,$$

où $\mathcal{H}(\nu)$ est l'entropie d'une gaussienne normalisée $\mathcal{N}(0,1)$ et les quantités e_l représentent des moments de p correspondants à des fonctions G_l :

$$e_l = \int_\xi p(\xi) G_l(\xi) d\,\xi.$$

La quantité $\mathcal{H}(Y \mid Z)$ est donc approchée par :

$$\left\{ \begin{array}{l} \hat{\mathcal{H}}(Y \mid Z) = -\frac{|\mathcal{R}_1|}{|\mathcal{S}|}\left(\mathcal{H}(\nu) - \frac{1}{2}\sum_{l=1}^{L} {e_l}^2\right) - \frac{|\mathcal{R}_2|}{|\mathcal{S}|}\left(\mathcal{H}(\nu) - \frac{1}{2}\sum_{l=1}^{L} {f_l}^2\right) \\ \\ e_l = \dfrac{\int_{\mathcal{R}_1} G_l(Y(r)) d\,r}{|\mathcal{R}_1|}\ ,\ f_l = \dfrac{\int_{\mathcal{R}_2} G_l(Y(r)) d\,r}{|\mathcal{R}_2|} \end{array} \right. \tag{VIII.3}$$

Remarque 25 *On peut montrer l'équivalence de l'approche informationnelle avec le maximum de vraisemblance. En effet, la vraisemblance normalisée s'écrit :*

$$\begin{aligned} \frac{\log\, \mathrm{p}(\boldsymbol{Y} \mid \boldsymbol{Z})}{|\mathcal{S}|} &\approx \frac{|\mathcal{R}_1|}{|\mathcal{S}|}\frac{\int_{\mathcal{R}_1} \log p_1(Y(r)) d\,r}{|\mathcal{R}_1|} + \frac{|\mathcal{R}_2|}{|\mathcal{S}|}\frac{\int_{\mathcal{R}_2} \log p_2(Y(r)) d\,r}{|\mathcal{R}_2|} \\ &\approx -\frac{|\mathcal{R}_1|}{|\mathcal{S}|}\mathcal{H}(Y \mid Z=1) - \frac{|\mathcal{R}_2|}{|\mathcal{S}|}\mathcal{H}(Y \mid Z=2) \\ &\approx -\mathcal{H}(Y \mid Z) \end{aligned}$$

La maximisation de la vraisemblance est ainsi équivalente à la maximization du flux d'information entre Y et Z.

Remarque 26 *On note la similarité des deux problèmes de séparation et de ségmentation au niveau de l'approximation des entropies et des moyennages empiriques. En effet, les mêmes expressions entropiques apparaissent dans les deux cas. Cependant, rien ne nous contraint à choisir les mêmes approximations de l'entropie (les mêmes fonctions G_l).*

[B].1 Approche par contour

On peut paramétriser la classification par une courbe plane $\vec{C}$ qu'on appelle **contour** (voir figure (VIII.2)). Le contour $\vec{C}$ partage l'image en deux régions : les pixels à l'intérieur de $\vec{C}$ appartiennent à $\mathcal{R}_1$ et les pixels à l'extérieur de $\vec{C}$ appartiennent à $\mathcal{R}_2$. Par conséquent, au lieu d'estimer tout le champ d'étiquettes $\boldsymbol{Z}$ (comme c'était le cas dans l'approche bayésienne), on estime juste le contour $\vec{C}$ minimizant une énergie $E(\vec{C})$ définie par :

$$E(\vec{C}) = -\hat{I}(Y,Z) + \alpha \oint_{\vec{C}} d\,s$$

où on a rajouté le terme $\alpha \oint_{\vec{C}} d\,s$ pénalisant la longueur du contour $\vec{C}$. Ce terme est donc l'équivalent de la régularisation markovienne dans l'approche bayésienne.

La minimization de l'énergie est effectuée avec les techniques d'évolution de contour. On fait évoluer $\vec{C}$ dans la direction opposée du gradient [Zhu et Yuille, 1996]. On montre que le gradient de l'énergie $E(\vec{C})$, en considérant les approximations (VIII.3), possède l'expression suivante :

$$\frac{\partial \vec{C}}{\partial t} = \underbrace{\left\{ \left[\sum_{l=1}^{L} (e_l - f_l) \left((G_l(Y) - e_l) + (G_l(Y) - f_l) \right) \right] - \alpha\kappa \right\}}_{\text{vitesse d'évolution du contour}} \vec{N}$$

où κ est la courbure de $\vec{C}$ et $\vec{N}$ est le vecteur unitaire normal à la courbe $\vec{C}$.

L'évolution dynamique du contour est implémentée avec les méthodes d'ensembles de niveau [Osher et Sethian, 1988].

La figure (VIII.3) illustre le schéma de la séparation et de la ségmentation. La machine proposée consiste en deux blocs maximisant des flux d'information. Le premier bloc "Bloc-Sep" fait évoluer la matrice de séparation $\boldsymbol{B}$ pour maximiser le transfert d'information entre les observations $\boldsymbol{X}$ et les images séparées $\boldsymbol{V}$ connaissant la classification $\boldsymbol{Z}$. Le deuxième bloc "Bloc-Seg" fait évoluer le contour $\vec{C}$ pour maximiser le flux d'information entre les images $\boldsymbol{Y}$ et leur classification $\boldsymbol{Z}$.

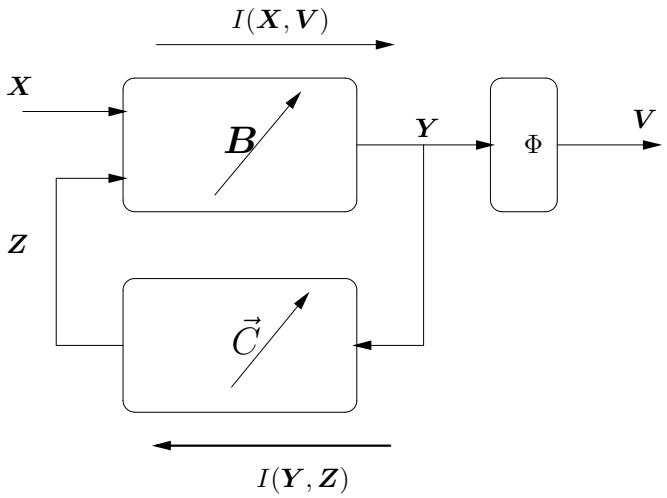


FIG. VIII.3: La machine d'apprentissage est constituée de deux blocs maximisant les flux d'informations.

La séparation et la ségmentation peuvent être menées, d'une manière disjointe, sur des bases informationnelles. Cependant, on n'arrive pas encore à modéliser le problème joint avec cette approche informationnelle. Ceci était possible dans le cadre bayésien grâce à la consistance logique des probabilités. En effet, la séparation connaissant la classification est décrite par la proposition $(\boldsymbol{X} \wedge \boldsymbol{Z} \longrightarrow \boldsymbol{S})$, la ségmentation est décrite par $(\boldsymbol{S} \longrightarrow \boldsymbol{Z})$ et le problème joint est décrit par $(\boldsymbol{X} \longrightarrow \boldsymbol{Z} \wedge \boldsymbol{S})$. L'objectif de la section suivante est d'essayer de trouver l'équivalent de la démarche probabiliste, et tout particulièrement sa consistance logique, dans l'approche informationnelle.

VIII.2 Vers la logique des questions...

Nous pensons que la théorie de la logique des questions proposée par R.T. Cox [Cox, 1979] et récemment reprise par R. Fry [Fry, 2000, 2001] et K. Knuth [Knuth, 2000, 2001, 2002] peut offrir un cadre logique à l'approche informationnelle.

VIII.2.1 Quelques définitions

On définit une question comme l'ensemble des propositions qui représentent des réponses possibles à cette question [Cox, 1979]. Par exemple, la question $\boldsymbol{P}$ = "Dans quelle ville tu préfères vivre ?" est définie par l'ensemble $\{a_i\}$ de toutes les villes du monde. La question $\boldsymbol{F}$ = "Dans quelle ville en France tu préfères vivre ?" est définie par l'ensemble $\{f_j\}$ des villes de France.

On définit la conjonction $\boldsymbol{A} \wedge \boldsymbol{B}$ de deux questions $\boldsymbol{A}$ et $\boldsymbol{B}$ comme étant la question posée conjointement par $\boldsymbol{A}$ et $\boldsymbol{B}$. Son système de propositions est définie par la conjonction des deux systèmes $\{a_i\}$ et $\{b_i\}$:

$$\boldsymbol{A} \wedge \boldsymbol{B} = \{a_i \wedge b_j\}$$

On définit aussi la disjonction $\boldsymbol{A} \vee \boldsymbol{B}$ dont le système de propositions est l'union des deux systèmes de propositions correspondants à $\boldsymbol{A}$ et $\boldsymbol{B}$:

$$\boldsymbol{A} \vee \boldsymbol{B} = \{a_i, b_j\}$$

On peut munir l'ensemble des questions d'une relation d'ordre telle que $\boldsymbol{A} \longrightarrow \boldsymbol{B}$ siginifie $\boldsymbol{A} \wedge \boldsymbol{B} = \boldsymbol{A}$ et $\boldsymbol{A} \vee \boldsymbol{B} = \boldsymbol{B}$. Dans l'exemple défini plus haut, on a $\boldsymbol{P} \longrightarrow \boldsymbol{F}$. En effet, la question posée conjointement est $\boldsymbol{P} \wedge \boldsymbol{F}$ = "Dans quelle ville tu préfères vivre ?" et la question commune posée est $\boldsymbol{P} \vee \boldsymbol{F}$ = "Dans quelle ville en France tu préfères vivre ?". La question $\boldsymbol{P}$ inclut la question $\boldsymbol{F}$.

La proposition $\boldsymbol{A} \longrightarrow \boldsymbol{B}$ ne prend que deux valeurs possibles 0 (Faux) ou 1 (Vraie). On introduit alors une mesure d'incertitude $b(\boldsymbol{A} \longrightarrow \boldsymbol{B}) = b(\boldsymbol{A} \mid \boldsymbol{B})$. On peut interpréter $b(.)$ comme la pertinence de la question $\boldsymbol{A}$ par rapport à la question $\boldsymbol{B}$. Cette mesure représente l'équivalent de la probabilité mesurant le degré d'implication entre deux propositions. Les règles de produit et de somme pour la mesure $b(.)$ sont :

$$b(\boldsymbol{A} \vee \boldsymbol{B} \mid \boldsymbol{C}) = b(\boldsymbol{A} \mid \boldsymbol{B} \vee \boldsymbol{C})\, b(\boldsymbol{B} \mid \boldsymbol{C})$$

$$b(\boldsymbol{A} \mid \boldsymbol{B}) + b(\sim \boldsymbol{A} \mid \boldsymbol{B}) = 1$$

En utilisant la commutativité de la disjonction, on arrive à l'équivalent de la règle de Bayes :

$$b(\boldsymbol{B} \mid \boldsymbol{A} \vee \boldsymbol{C}) = \frac{b(\boldsymbol{A} \mid \boldsymbol{B} \vee \boldsymbol{C})\, b(\boldsymbol{B} \mid \boldsymbol{C})}{b(\boldsymbol{A} \mid \boldsymbol{C})}.$$

La conjecture proposée dans [Cox, 1979] est que l'entropie convient comme une mesure $b(.)$ de la pertinence. En effet, les relations vérifiées par le calcul de l'entropie coincident avec celles vérifiées par l'algèbre des questions. A titre illustratif, on considère deux questions $\boldsymbol{B}$ et $\boldsymbol{C}$ et une question $\boldsymbol{H}$ représentant toutes les questions possibles. La relation suivante :

$$b(\boldsymbol{B} \vee \sim \boldsymbol{C} \mid \boldsymbol{H}) = b(\boldsymbol{C} \vee \sim \boldsymbol{B} \mid \boldsymbol{H}) + b(\boldsymbol{B} \mid \boldsymbol{H}) - b(\boldsymbol{C} \mid \boldsymbol{H})$$

coincide avec la relation vérifiée par l'entropie :

$$\mathcal{H}(b \mid c) = \mathcal{H}(c \mid b) + \mathcal{H}(b) - \mathcal{H}(c).$$

En particulier l'information mutuelle $I(b, c)$ entre deux variables b et c représente la pertinence de la question $\boldsymbol{B} \vee \boldsymbol{C}$ sur la question *a priori* $\boldsymbol{H}$:

$$I(b, c) = b(\boldsymbol{B} \vee \boldsymbol{C} \mid \boldsymbol{H}).$$

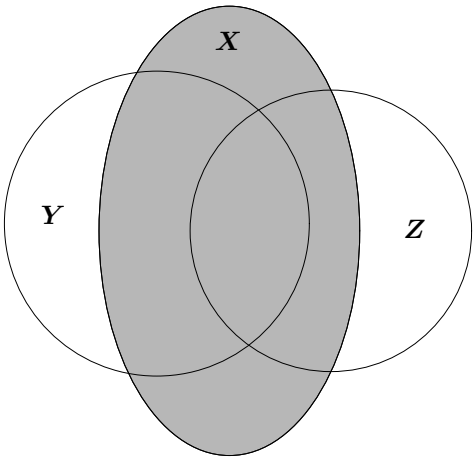
VIII.2.2 Interprétation de l'InfoMax

En revenant au problème de la séparation de sources, on va essayer de donner une formulation basée sur la logique des questions expliquant le fonctionnement de la machine d'apprentissage (VIII.3) proposée plus haut.

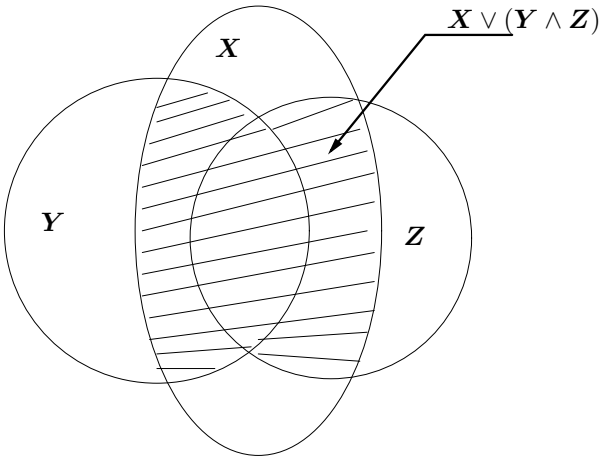
On définit les questions suivantes :
$\boldsymbol{Y}_q$ ="Quelle est l'intensité y des images sources ?"
$\boldsymbol{Z}_q$ ="Quelle est la classification z des images sources ?"
$\boldsymbol{X}_q$ ="Quelle est l'intensité x des images observées ?"
Le problème de la séparation et de la ségmentation consiste à maximiser la pertinence de la question $\boldsymbol{X}_q \vee (\boldsymbol{Y}_q \wedge \boldsymbol{Z}_q)$ sur la question d'intérêt $\boldsymbol{Y}_q \wedge \boldsymbol{Z}_q$. La matrice séparatrice $\boldsymbol{B}$ et le contour $\vec{C}$ sont alors solutions du problème d'optimisation suivant :

$$(\boldsymbol{B}, \vec{C}) = \arg\max \ \{b(\boldsymbol{X}_q \vee (\boldsymbol{Y}_q \wedge \boldsymbol{Z}_q) \mid \boldsymbol{Y}_q \wedge \boldsymbol{Z}_q)\} \tag{VIII.4}$$

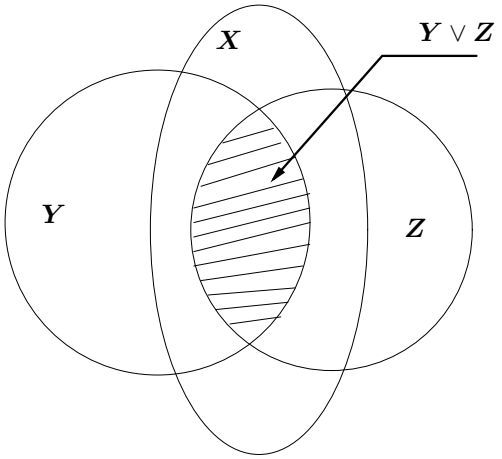
Le diagramme suivant (VIII.4) donne un aperçu sur la dynamique de l'évolution des quantités informationnelles lors de l'évolution de la matrice $\boldsymbol{B}$ et du contour $\vec{C}$.



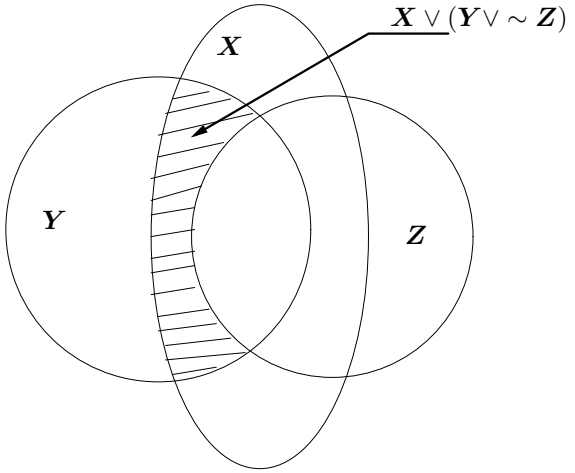
Mesure de l'information $\boldsymbol{X}$.



Maximisation de la projection de l'information $\boldsymbol{X}$ sur $\boldsymbol{Y} \wedge \boldsymbol{Z}$.



Maximisation de la projection de l'information $\boldsymbol{Y}$ sur $\boldsymbol{Z}$: Ségmentation.



Maximisation de la projection de l'information $\boldsymbol{X}$ sur $\boldsymbol{Y} \wedge \sim \boldsymbol{Z}$: Séparation de sources.

FIG. VIII.4: Diagramme expliquant les maximisations des flux d'information sur la base de la théorie de la logique des questions.

Nous avons donné ainsi un cadre logique à l'opération conjointe de la séparation et de la ségmentation. Cependant, notre pratique avec la logique des questions reste à un stade prématuré. Ainsi, le calcul et l'interprétation de certains des termes issus de la décomposition de l'expression (VIII.4) ne sont pas encore

claires. Une bonne maîtrise (parfois difficile à cause des "contre-intuitions") de ce domaine va certes ouvrir un champ de recherche et d'applications très important.

Bibliographie

[Bell et Sejnowski, 1995] A. J. Bell et T. J. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7 (6) : 1129–1159, 1995.

[Cox, 1979] R. Cox. On inference and inquiry. In *Proc. Maximum Entropy Formalism Conference, MIT Press*, pages 119–167, 1979.

[Fry, 2000] R. Fry. Cybernetic systems based on inductive logic. In A. Mohammad-Djafari, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 106–119, Gif-sur-Yvette, juillet 2000. Proc. of MaxEnt, Amer. Inst. Physics.

[Fry, 2001] R. Fry. The engineering of cybernetic systems. In R. L. Fry, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 497–528. MaxEnt Workshops, Amer. Inst. Physics, août 2001.

[Hyvärinen, 1997] A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. Report, Helsinki University, 1997.

[Kim *et al.*, 2002] J. Kim, J. Fisher, A. Yezzi, M. Cetin et A. Willsky. Nonparametric methods for image segmentation using information theory and curve evolution. In *Proc. IEEE ICIP*, volume 3, pages 797–800, Rochester, New York, septembre 2002.

[Knuth, 2000] K. Knuth. Source separation as an exercise in logical induction. In A. Mohammad-Djafari, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 340–349, Gif-sur-Yvette, juillet 2000. Proc. of MaxEnt, Amer. Inst. Physics.

[Knuth, 2001] K. Knuth. Inductive logic : From experimental design to data analysis. In R. L. Fry, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 392–404. MaxEnt Workshops, Amer. Inst. Physics, août 2001.

[Knuth, 2002] K. Knuth. What is a question ? In C. J. Williams, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 227–242, Moscow, Idaho, août 2002. MaxEnt Workshops, Amer. Inst. Physics.

[Osher et Sethian, 1988] S. Osher et J. A. Sethian. Fronts propagating with curvature-dependent speed : Algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics*, 79 : 12–49, 1988.

[Pham, 2002] D.-T. Pham. Blind Separation of Non Stationary Non Gaussian Sources. In *Proceeding of EUSIPCO'02*, Toulouse, septembre 2002.

[Pham et Cardoso, 2001] D.-T. Pham et J. Cardoso. Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. Signal Processing*, 49, 9 (11) : 1837–1848, 2001.

[Unal *et al.*, 2002] G. Unal, H. Krim et A. Yezzi. A vertex-based representation of objects in an image. In *Proc. IEEE ICIP*, volume 1, pages 896–899, Rochester, New York, septembre 2002.

[Zhu et Yuille, 1996] S. Zhu et A. Yuille. Region competition : Unifying snakes, region growing, and bayes/mdl for multiband image segmentation. *IEEE PAMI*, 18 : 884–900, 1996.

RÉFÉRENCES BIBLIOGRAPHIQUES

[Amari, 1985] S. Amari. *Differential-Geometrical Methods in Statistics*. Volume 28 of Springer Lecture Notes in Statistics, Springer-Verlag, New York, 1985.

[Amari et Nagaoka, 2000] S. Amari et H. Nagaoka. *Methods of Information Geometry*, volume 191 of Translations of Mathematical Monographs. AMS, OXFORD, University Press, 2000.

[Amari et Cardoso, 1997] S.-I. Amari et J.-F. Cardoso. Blind source separation — semiparametric statistical approach. *IEEE Trans. Signal Processing*, 45 (11) : 2692–2700, novembre 1997.

[Amari *et al.*, 1996] S.-I. Amari, A. Cichocki et H. Yang. A new learning algorithm for blind source separation. In *Advances in Neural Information Processing Systems 8*, pages 757–763. MIT Press, 1996.

[Ans *et al.*, 1985] B. Ans, J. Hérault et C. Jutten. Adaptive neural architectures : detection of primitives. In *Proc. of COGNITIVA'85*, pages 593–597, Paris, France, 1985.

[Attias, 1999] H. Attias. Blind separation of noisy mixture : An EM algorithm for independent factor analysis. *Neural Computation*, 11 : 803–851, 1999.

[Baccigalupi *et al.*, 2000] C. Baccigalupi, L. Bedini, C. Burigana, G. De Zotti, A. Farusi, D. Maino, M. Maris, F. Perrotta, E. Salerno, L. Toffolatti et A. Tonazzini. *Monthly Notices of the Royal Astronomical Society*, 318 : 769–780, novembre 2000.

[Bell et Sejnowski, 1995] A. J. Bell et T. J. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7 (6) : 1129–1159, 1995.

[Belouchrani *et al.*, 1997] A. Belouchrani, K. Abed Meraim, J.-F. Cardoso et Éric Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. Signal Processing*, 45 (2) : 434–44, février 1997.

[Belouchrani et Amin, 1997] A. Belouchrani et M. Amin. Blind source separation using time-frequency distributions : algorithm and asymptotic performance. In *Proc. ICASSP*, pages 3469 – 3472, Munchen, 1997.

[Belouchrani et Cardoso, 1995] A. Belouchrani et J.-F. Cardoso. Maximum likelihood source separation by the expectation-maximization technique : deterministic and stochastic implementation. In *Proc. NOLTA*, 1995.

[Bennet et *al*, 2003] C. Bennet et *al*. First year Wilkinson Microwave Anisotropy Probe (WMAP) observations : Preliminary maps and basic results. *submitted to ApJ*, 2003.

[Benoît et *al*, 2003] A. Benoît et *al*. The cosmic microwave background anisotropy power spectrum measured by Archeops. *Astronomy and Astrophysics*, 399 : L19–L23, mars 2003.

[Bermond, 2000] O. Bermond. *Méthodes statistiques pour la séparation de sources*. thèse de doctorat, Ecole Nationale Supérieure des Télécommunications, 2000.

[Bouchet et Gispert, 1999] F. R. Bouchet et R. Gispert. *New Astronomy*, 4 : 443–479, novembre 1999.

[Box et Tiao, 1972] G. E. P. Box et G. C. Tiao. *Bayesian inference in statistical analysis*. Addison-Wesley publishing, 1972.

[Boyles, 1983] R. A. Boyles. On the convergence of the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 45 : 47–50, 1983.

[Brewer, 1978] J. W. Brewer. Kronecker products and matrix calculus in system theory. *IEEE Trans. Circ. Syst.*, CS-25 (9) : 772–781, 1978.

[Brooks et Roberts, 1995] S. Brooks et G. Roberts. Diagnosing convergence of Markov chain Monte Carlo algorithms. Technical report no. 95-12, Stat. Lab., U. of Cambridge, 1995.

[Burg, 1982] J. P. Burg. Estimation of structured covariance matrices. *Proceeding of IEEE*, 70 (9) : 963–974, septembre 1982.

[Cardoso et Labeld, 1996] J. Cardoso et B. Labeld. Equivariant adaptative source separation. *Signal Processing*, 44 : 3017–3030, 1996.

[Cardoso *et al.*, 2002] J. Cardoso, H. Snoussi, J. Delabrouille et G. Patanchon. Blind separation of noisy gaussian stationary sources. application to cosmic microwave background imaging. In *Eusipco*, Toulouse, septembre 2002.

[Cardoso, 1989] J.-F. Cardoso. Source separation using higher order moments. In *Proc. ICASSP*, pages 2109–2112, 1989.

[Cardoso, 1997] J. F. Cardoso. Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing*, 4 : 112–114, avril 1997.

[Cardoso et Souloumiac, 1993] J.-F. Cardoso et A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140 (6) : 362–370, décembre 1993.

[Celeux et Diebolt, 1985] G. Celeux et J. Diebolt. The SEM algorithm : A probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Comput. Statist. Quat.*, 2 : 73–82, 1985.

[Celeux et Diebolt, 1990] G. Celeux et J. Diebolt. Une version de type recuit simulé de l'algorithme EM. *Compte-rendus de l'académie des sciences*, 310 : 2, 1990.

[Choi et Cichocki, 2000] S. Choi et A. Cichocki. Blind separation of nonstationary sources in noisy mixtures. *Electronics Letters*, 36(9) : 848–849, apr 2000.

[Cichocki et Moszczynski, 1992] A. Cichocki et L. Moszczynski. A new learning algorithm for blind separation of sources. *Electronics Letters*, 28(21) : 1986–1987, 1992.

[Cichocki *et al.*, 1994] A. Cichocki, R. Unbehauen et E. Rummert. Robust learning algorithm for blind separation of signals. *Electronics Letters*, 30(17) : 1386–1387, 1994.

[Cichocki et Unbehaunen, 1996] A. Cichocki et R. Unbehaunen. Robust neural networks with on-line learning for blind identification and blind separation of sources. *IEEE Trans. on Circuits and Systems*, 43 (11) : 894–906, 1996.

[Comon, 1994] P. Comon. Independent Component Analysis, a new concept ? *Signal processing, Special issue on Higher-Order Statistics, Elsevier*, 36 (3) : 287–314, avril 1994.

[Cox, 1946] R. Cox. Probability, frequency and reasonable expectation. *Am. J. Physics*, 14 : 1–13, 1946.

[Cox, 1961] R. Cox. *The Algebra of Probable Inference.* Johns Hopkins University Press, Baltimore, MD, USA, 1961.

[Cox, 1979] R. Cox. On inference and inquiry. In *Proc. Maximum Entropy Formalism Conference, MIT Press*, pages 119–167, 1979.

[Darmois, 1953] G. Darmois. Analyse Générale des Liaisons Stochastiques. *Rev. Inst. Internat. Stat.*, 21 : 2–8, 1953.

[Day, 1969] N. Day. Estimating the components of a mixture of normal distributions. *Biometrika*, 56 : 463–474, 1969.

[De Bernardis et *al*, 2000] P. De Bernardis et *al.* A flat Universe from high-resolution maps of the cosmic microwave background radiation. *Nature*, 404 : 955–959, 2000.

[De Finetti, 1974] B. De Finetti. *Theory of Probability,* 2 *vols.* Translated by A. Machi and A.F.M. Smith, Wiley, London, 1974.

[De Zotti *et al.*, 1999] G. De Zotti, L. Toffolatti, F. Argüeso, R. D. Davies, P. Mazzotta, R. B. Partridge, G. F. Smoot et N. Vittorio. In *AIP Conf. Proc. 476 : 3K cosmology*, page 204, 1999.

[Delabrouille *et al.*, 2001] J. Delabrouille, G. Patanchon et E. Audit. *Monthly Notices of the Royal Astronomical Society*, 2001.

[Delfosse et Loubaton, 1995] N. Delfosse et P. Loubaton. Adaptive blind separation of independent sources : a deflation approach. *Signal Processing*, 45 : 59–83, 1995.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird et D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39 : 1–38, 1977.

[Descombes, 1993] X. Descombes. *Champs Markoviens en analyse d'image*. thèse de doctorat, École Nationale Supérieure des Télécommunications, Paris, décembre 1993.

[Eke *et al.*, 1998] V. R. Eke, J. F. Navarro et C. S. Frenk. The evolution of X-Ray Clusters in a Low Density Universe. *APJ*, 503 : 569, novembre 1998.

[Feller, 1968] W. Feller. *An introduction to probability theory and its applications, Volume 1, 3rd edition*. Wiley, New York, NY, USA, 1968.

[Fry, 2000] R. Fry. Cybernetic systems based on inductive logic. In A. Mohammad-Djafari, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 106–119, Gif-sur-Yvette, juillet 2000. Proc. of MaxEnt, Amer. Inst. Physics.

[Fry, 2001] R. Fry. The engineering of cybernetic systems. In R. L. Fry, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 497–528. MaxEnt Workshops, Amer. Inst. Physics, août 2001.

[Gaeta et Lacoume, 1990] M. Gaeta et J.-L. Lacoume. Source separation without prior knowledge : the maximum likelihood solution. In *Proc. EUSIPCO'90*, pages 621–624, 1990.

[Geweke, 1989] J. Geweke. Bayesian inference in econometric models using monte carlo integration. *Econometrika*, 57 : 1317–1339, 1989.

[Ghahramani et Jordan, 1997] Z. Ghahramani et M. Jordan. Factorial Hidden Markov Models. *Machine Learning*, (29) : 245–273, 1997.

[Gostiaux, 1993a] B. Gostiaux. *Cours de mathématiques spéciales. Algèbre*. Presses Universitaires de France, Paris, 1993.

[Gostiaux, 1993b] B. Gostiaux. *Cours de mathématiques spéciales. Analyse fonctionnelle et calcul différentiel*. Presses Universitaires de France, Paris, 1993.

[Grainge et *al*, 2003] K. Grainge et *al*. The CMB power spectrum out to l=1400 measured by the VSA. *MNRAS in press*, 2003.

[Green, 1990] P. J. Green. Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Trans. Medical Imaging*, 9 (1) : 84–93, mars 1990.

[Hammersley et Clifford, 1968] J. M. Hammersley et P. Clifford. Markov fields of finite graphs and lattices. Rapport interne, University of California-Berkeley, preprint, 1968.

[Hanany et *al*, 2000] S. Hanany et *al*. MAXIMA-1 : A Measurement of the cosmic microwave background anisotropy on angular scales of 10 arcminutes to 5 degrees. *ApJ Letters*, 545 : L5–L9, décembre 2000.

[Hastings, 1970] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57 : 97, janvier 1970.

[Hathaway, 1986] R. J. Hathaway. A constrained EM algorithm for univariate normal mixtures. *J. Statist. Comput. Simul.*, 23 : 211–230, 1986.

[Hérault, 1985] J. Hérault. Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé. In *Actes 10e coll. GRETSI*, pages 1017–1022, Nice, France, 1985.

[Hérault et Ans, 1984] J. Hérault et B. Ans. Circuits neuronaux à synapses modifiables : décodage de messages composites par apprentissage non supervisé. *C. R. de l'Académie des Sciences*, 299 (III-13) : 525–528, 1984.

[Hero et Fessler, 1993] A. O. Hero et J. A. Fessler. Asymptotic convergence properties of EM-type algorithms. Preprints 85-T-21, Dept. of Electrical Engineering and Computer Science, University of Michigan, 1993.

[Hobson *et al.*, 1998] M. Hobson, A. W. Jones, A. N. Lasenby et F. R. Bouchet. *Monthly Notices of the Royal Astronomical Society*, 300 : 1–29, octobre 1998.

[Hu et Sugiyama, 1996] W. Hu et N. Sugiyama. *APJ*, 471 : 542, novembre 1996.

[Hunt, 1971] B. R. Hunt. A matrix theory proof of the discrete convolution theorem. *IEEE Trans. Automat. Contr.*, AC-19 : 285–288, 1971.

[Hyvärinen, 1997] A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. Report, Helsinki University, 1997.

[Hyvärinen, 1999] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3) : 626–634, 1999.

[Hyvärinen et Oja, 1997] A. Hyvärinen et E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7) : 1483–1492, 1997.

[Jungman *et al.*, 1996] G. Jungman, M. Kamionkowski, A. Kosowsky et D. N. Spergel. *Physical Review Letters*, 76 : 1007–1010, février 1996.

[Jutten, 2000] C. Jutten. Source separation : from dusk till dawn. In *Proc. of 2nd Int. Workshop on Independent Component Analysis and Blind Source Separation (ICA'2000)*, pages 15–26, Helsinki, Finland, 2000.

[Jutten et Herault, 1991] C. Jutten et J. Herault. Blind separation of sources .1. an adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24 (1) : 1–10, 1991.

[Kass *et al.*, 1988] M. Kass, A. P. Witkin et D. Terzopoulos. Snakes : Active contour models. *Int. J. Computer Vision*, 1 (4) : 321–331, 1988.

[Kass et Wasserman, 1994] R. E. Kass et L. Wasserman. Formal rules for selecting prior distributions : A review and annotated bibliography. Technical report no. 583, Department of Statistics, Carnegie Mellon University, 1994.

[Kiefer et Wolfowitz, 1956] J. Kiefer et J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.*, 27 : 887–906, 1956.

[Kim *et al.*, 2002] J. Kim, J. Fisher, A. Yezzi, M. Cetin et A. Willsky. Nonparametric methods for image segmentation using information theory and curve evolution. In *Proc. IEEE ICIP*, volume 3, pages 797–800, Rochester, New York, septembre 2002.

[Knuth, 1999] K. Knuth. A Bayesian approach to source separation. In *Proceedings of Independent Component Analysis Workshop*, pages 283–288, 1999.

[Knuth, 2000] K. Knuth. Source separation as an exercise in logical induction. In A. Mohammad-Djafari, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 340–349, Gif-sur-Yvette, juillet 2000. Proc. of MaxEnt, Amer. Inst. Physics.

[Knuth, 2001] K. Knuth. Inductive logic : From experimental design to data analysis. In R. L. Fry, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 392–404. MaxEnt Workshops, Amer. Inst. Physics, août 2001.

[Knuth, 2002] K. Knuth. What is a question ? In C. J. Williams, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 227–242, Moscow, Idaho, août 2002. MaxEnt Workshops, Amer. Inst. Physics.

[Kuo et *al*, 2002] C. Kuo et *al.* High resolution observations of the CMB power spectrum with ACBAR. *ApJ, available at astro-ph/0212289*, 2002.

[Lehmann et Casella, 1996] E. Lehmann et G. Casella. *Theory of point estimation (revised edition)*. Chapman and Hall, New York, NY, USA, 1996.

[Lindley, 1980] D. Lindley. Apporximate Bayesian methods (with discussion). In *Bayesian Statistics, J.M. Bernardo et al. editor. Valencia University Press*, pages 223–245, 1980.

[Liu et Pierce, 1994] Q. Liu et D. A. Pierce. A note on Gauss-Hermite quadrature. *J. Amer. Statist. Assoc.*, 81 (3) : 624–629, 1994.

[Malouche et Macchi, 1998] Z. Malouche et O. Macchi. Adaptive unsupervised extraction of one component of a linear mixture with a single neuron. *IEEE Trans. on Neural Networks*, 9(1) : 123–138, 1998.

[Matsuoka *et al.*, 1995] K. Matsuoka, M. Ohya et M. Kawamoto. A neural net for blind separation of nonstationary sources. *Neural Networks*, 8(3) : 411–419, 1995.

[McLachlan et Basford, 1987] G. J. McLachlan et K. E. Basford. *Mixture Models, inference and applications to clustering*, volume 84 de *statistics*. Dekker, 1987.

[McLachlan et Krishnan, 1997] G. J. McLachlan et T. Krishnan. *The EM Algorithm and Extensions*. Wiley series in probability and statistics. John Wiley and Sons, Inc., 1997.

[McLachlan et Peel, 2000] G. J. McLachlan et D. Peel. *Finite Mixture Models*. Wiley series in probability and statistics. Wiley, 2000.

[Metropolis *et al.*, 1953] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller et E. Teller. Equations of state calculations by fast computing machines. *Journal of chemical physics*, 21 : 1087–1092, juin 1953.

[Meyn et Tweedie, 1993] S. Meyn et R. Tweedie. *Markov chains and stochastic stability*. Springer-Verlag, London, 1993.

[Mohammad-Djafari, 1999] A. Mohammad-Djafari. A Bayesian approach to source separation. In J. R. G. Erikson et C. Smith, éditeurs, *Bayesian Inference and Maximum Entropy Methods*, Boise, IH, USA, juillet 1999. MaxEnt Workshops, Amer. Inst. Physics.

[Moreau et Macchi, 1996] E. Moreau et O. Macchi. High-order contrasts for self-adaptative source separation. In *Adaptative Control Signal Process. 10*, pages 19–46, 1996.

[Moulines *et al.*, 1997] E. Moulines, J. Cardoso et E. Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *ICASSP-97*, Munich, Allemagne, avril 1997.

[Müller, 1991] P. Müller. A generic approach to posterior integration and Gibbs sampling. Rapport technique 91-09, Purdue Uni. West Lafayette, Indiana, Indiana, 1991.

[Müller, 1992] P. Müller. Alternatives to the Gibbs sampling scheme. Rapport technique, Institute of Statistics and Decision Sciences, Duke Uni., 1992.

[Ormoneit et Tresp, 1998] D. Ormoneit et V. Tresp. Averaging, maximum penalized likelihood and Bayesian estimation for improving Gaussian mixture probability density estimates. *IEEE Transactions on Neural Networks*, 9 (4) : 639–649, juillet 1998.

[Osher et Sethian, 1988] S. Osher et J. A. Sethian. Fronts propagating with curvature-dependent speed : Algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics*, 79 : 12–49, 1988.

[Patanchon *et al.*, 2003] G. Patanchon, H. Snoussi, J. Cardoso et J. Delabrouille. Component separation for cosmic microwave background data : a blind approach based on spectral diversity. In *PSIP*, Grenoble, janvier 2003.

[Pearson et *al*, 2003] T. Pearson et *al*. The anisotropy of the microwave background to l = 3500 : Mosaic observations with the Cosmic Background Imager. *Accepted by The Astrophysical Journal*, 2003.

[Pham, 1996] D.-T. Pham. Blind separation of instantaneous mixture sources via independent component analysis. *IEEE Trans. Signal Processing*, 44, 1996.

[Pham, 2002] D.-T. Pham. Blind Separation of Non Stationary Non Gaussian Sources. In *Proceeding of EUSIPCO'02*, Toulouse, septembre 2002.

[Pham et Cardoso, 2001] D.-T. Pham et J. Cardoso. Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. Signal Processing*, 49, 9 (11) : 1837–1848, 2001.

[Pham *et al.*, 1992] D.-T. Pham, P. Garrat et C. Jutten. Separation of a mixture of independent sources through a maximum likelihood approach. In *Proc. EUSIPCO'92*, pages 771–774, 1992.

[Rabiner et Juang, 1986] L. R. Rabiner et B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Mag.*, pages 4–16, 1986.

[Rabiner, 1989] R. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77 (2) : 257–286, février 1989.

[Rahbar et Reilly, 2001] K. Rahbar et J. Reilly. Blind source separation of convolved sources by joint approximate diagonalization of cross-spectral density matrices. In *Proc. ICASSP*, 2001.

[Ridolfi et Idier, 1999] A. Ridolfi et J. Idier. Penalized maximum likelihood estimation for univariate normal mixture distributions. In *Actes 17e coll. GRETSI*, pages 259–262, Vannes, septembre 1999.

[Robert, 1996] C. Robert. *Méthodes de Monte-Carlo par chaînes de Markov*. Economica, Paris, 1996.

[Rodríguez, 1991] C. Rodríguez. Entropic priors. *Tech. rep. Electronic form* `http : omega.albany.edu :8008/entpriors.ps`, 1991.

[Rodríguez, 2001] C. Rodríguez. Entropic priors for discrete probabilistic networks and for mixtures of Gaussians models. In R. L. FRY, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 410–432. MaxEnt Workshops, Amer. Inst. Physics, août 2001.

[Roeder et Wasserman, 1997] K. Roeder et L. Wasserman. Practical bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.*, 92 : 894–902, 1997.

[Rowe, 1998] D. Rowe. *Correlated Bayesian Factor analysis*. thèse de doctorat, Department of Statistics, University of California, Riverside, 1998.

[Rubinstein, 1981] R. Rubinstein. *Simulation and the Monte Carlo Method*. J. Wiley, New York, 1981.

[Seljak et Zaldarriaga, 2000] U. Seljak et M. Zaldarriaga. *ApJ. Suppl. ser.*, 129 : 431, 2000.

[Senecal, 2000] P. Senecal, S.Amblard. MCMC methods for discrete source separation. In *Bayesian Inference and Maximum Entropy Methods*, pages 350–360, Gif-sur-Yvette, juillet 2000. Proc. of MaxEnt, Amer. Inst. Physics.

[Senecal, 2002] S. Senecal. *Méthodes de simulation Monte-Carlo par chaînes de Markov pour l'estimation de modèles. Applications en séparation de sources et en égalisation.* thèse de doctorat, INPG (Grenoble), 2002.

[Snoussi et Mohammad-Djafari, 2000] H. Snoussi et A. Mohammad-Djafari. Bayesian source separation with mixture of Gaussians prior for sources and Gaussian prior for mixture coefficients. In A. Mohammad-Djafari, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 388–406, Gif-sur-Yvette, juillet 2000. Proc. of MaxEnt, Amer. Inst. Physics.

[Snoussi et Mohammad-Djafari, 2001] H. Snoussi et A. Mohammad-Djafari. Penalized maximum likelihood for multivariate gaussian mixture. In R. L. Fry, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 36–46. MaxEnt Workshops, Amer. Inst. Physics, août 2001.

[Snoussi et Mohammad-Djafari, 2002a] H. Snoussi et A. Mohammad-Djafari. Bayesian unsupervised learning for source separation with mixture of gaussians prior. *To appear in Int. Journal of VLSI Signal Processing Systems*, 2002.

[Snoussi et Mohammad-Djafari, 2002b] H. Snoussi et A. Mohammad-Djafari. Information Geometry and Prior Selection. In C. Williams, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 307–327. MaxEnt Workshops, Amer. Inst. Physics, août 2002.

[Snoussi et Mohammad-Djafari, 2002c] H. Snoussi et A. Mohammad-Djafari. MCMC Joint Separation and Segmentation of Hidden Markov Fields. In *Neural Networks for Signal Processing XII*, pages 485–494. IEEE workshop, septembre 2002.

[Snoussi et Mohammad-Djafari, 2003] H. Snoussi et A. Mohammad-Djafari. Fast joint separation and segmentation of mixed images. *To appear in Journal of Electronic Imaging*, 2003.

[Snoussi *et al.*, 2001] H. Snoussi, G. Patanchon, J. Macías-Pérez, A. Mohammad-Djafari et J. Delabrouille. Bayesian blind component separation for cosmic microwave background observations. In R. L. Fry, éditeur, *Bayesian Inference and Maximum Entropy Methods*, pages 125–140. MaxEnt Workshops, Amer. Inst. Physics, août 2001.

[Souloumiac, 1995] A. Souloumiac. Blind source detection and separation using second order nonstationarity. In *Proc. ICASSP*, pages 1912–1915, 1995.

[Stolyarov *et al.*, 2002] V. Stolyarov, M. P. Hobson, M. A. J. Ashdown et A. N. Lasenby. *Monthly Notices of the Royal Astronomical Society*, 336 : 99–111, 2002.

[Tegmark et Estathiou, 1996] M. Tegmark et G. Estathiou. A method for subtracting foregrounds from multifrequency CMB sky maps. *Monthly Notices of the Royal Astronomical Society*, 281 : 1297, 1996.

[Tierney, 1994] L. Tierney. Markov chain for exploring posterior distribution. *Annals Statist.*, 22 (4) : 1701–1762, décembre 1994.

[Tierney et Kadane, 1986] L. Tierney et J. Kadane. Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Ass.*, (81) : 82–86, 1986.

[Tierney *et al.*, 1986] L. Tierney, R. Kass et J. Kadane. Approximate marginal densities of nonlinear functions. *Biometrika*, (76) : 425–433, 1986.

[Tikhonov et Arsenin, 1977] A. Tikhonov et V. Arsenin. *Solutions of Ill-Posed Problems*. Winston, Washington, DC, USA, 1977.

[Unal *et al.*, 2002] G. Unal, H. Krim et A. Yezzi. A vertex-based representation of objects in an image. In *Proc. IEEE ICIP*, volume 1, pages 896–899, Rochester, New York, septembre 2002.

[Winkler, 1995] G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*. Springer Verlag, Berlin, Allemagne, 1995.

[Wu, 1983] C. F. J. Wu. On the convergence of the EM algorithm. *Ann. Statist.*, 11 (1) : 95–103, 1983.

[Yu et Mykland, 1994] B. Yu et P. Mykland. Looking at Markov samplers through Cusum Path Plots : A simple diagnostic idea. Technical report no. 9413, Department of Statistics, U. of California, Berkeley, 1994.

[Zhu et Rohwer, 1995a] H. Zhu et R. Rohwer. Bayesian invariant measurements of generalisation. In *Neural Proc. Lett.*, volume 2 (6), pages 28–31, 1995.

[Zhu et Rohwer, 1995b] H. Zhu et R. Rohwer. Bayesian invariant measurements of generalisation for continuous distributions. Technical report, NCRG/4352, ftp ://cs.aston.ac.uk/neural/zhuh/continuous.ps.z, Aston University, 1995.

[Zhu et Yuille, 1996] S. Zhu et A. Yuille. Region competition : Unifying snakes, region growing, and bayes/mdl for multiband image segmentation. *IEEE PAMI*, 18 : 884–900, 1996.

Nom : SNOUSSI **Prénom :** Hichem

Titre : Approche bayésienne en séparation de sources. Applications en imagerie.

Title : Bayesian approach to source separation. Applications in imagery.

Résumé :
Mon travail de thèse consiste à développer l'approche bayésienne en séparation de sources. Mes contributions sont à la fois méthodologiques et algorithmiques illustrées par des applications en imagerie satellitaire et en cosmologie observationnelle.

- Au niveau méthodologique :
  - nous avons proposé une modélisation pertinente des sources. L'aspect hiérarchique de ce modèle est bien adapté à la structure cachée naturelle du problème de séparation de sources.
  - Nous avons étudié le problème de dégénérescence du maximum de vraisemblance dans le cas vectoriel et dans le contexte de séparation de sources.
  - Nous avons proposé une approche originale pour la sélection d'*a priori* avec les outils de la géométrie différentielle.
- Au niveau Algorithmique :
  - Nous avons proposé des algorithmes de séparation et de ségmentation dont le principe est l'exploitation de la non stationnarité dans le domaine temporel, spatial, spectral, temps-fréquence...
  - Nous avons mis en œuvre la solution bayésienne avec une impémentation parallèle de l'échantillonneur de Gibbs ainsi que d'autres approximations stochastiques de l'EM.
  - Ces algorithmes sont illustrés par une application en imagerie satellitaire et une application en cosmologie observationnelle.

Enfin, j'ouvre des perspectives théoriques sur la dualité de l'approche bayésienne et de l'approche informationnelle dans le cadre de la séparation et de la ségmentation conjointes des sources.

Abstract :
My thesis consists in applying the Bayesian approach to the source separation problem. My contributions have both theoretic and algorithmic aspects that are illustrated by applications in satellite imaging and in astrophysics.

- Theoretic contributions :
  - We have proposed a pertinent source model. The hierarchical aspect of this model is well adapted to the natural hidden structure of the source separation problem.
  - We have characterized the degeneracy of the maximum likelihood in the multivariate and the source separation context.
  - We have proposed an original contribution for *a priori* selection with the differential geometry tools.
- Algorithmic contributions :
  - We have proposed algorithms for the simultaneous separation and segmentation of sources. The common principle of this algorithms is the exploitation of the non stationarity in the temporal, spatial, spectral or time-frequency domains.
  - We have implemented a parallel Gibbs sampling and other stochastic approximations of the EM algorithm.
  - The performances of our method is illustrated with applications in satellite imaging and in astrophysics for the estimation of the CMB (Cosmic Microwave Background).

Finally, I open some theoretic perspectives on the duality between the Bayesian approach and the information theoretic approach in the context of the simultaneous separation and segmentation of sources.

Mots clés : Séparation de sources, Approche bayésienne, Géométrie de l'information, Cosmologie observationnelle, MCMC.

Keywords : Source separation, Bayesian approach, Information geometry, Astrophysics, MCMC.