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 excité 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 leur principes. Cette introduction débouche sur les principales contributions de l'auteur 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 ($t = 1..T$), le vecteur des m observations $\boldsymbol{x}(t) = [x_1(t), ..., x_m(t)]^*$ est lié aux sources par l'équation suivante :

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

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 **bruit**. On a alors la relation suivante entre les sources et les observations :

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

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

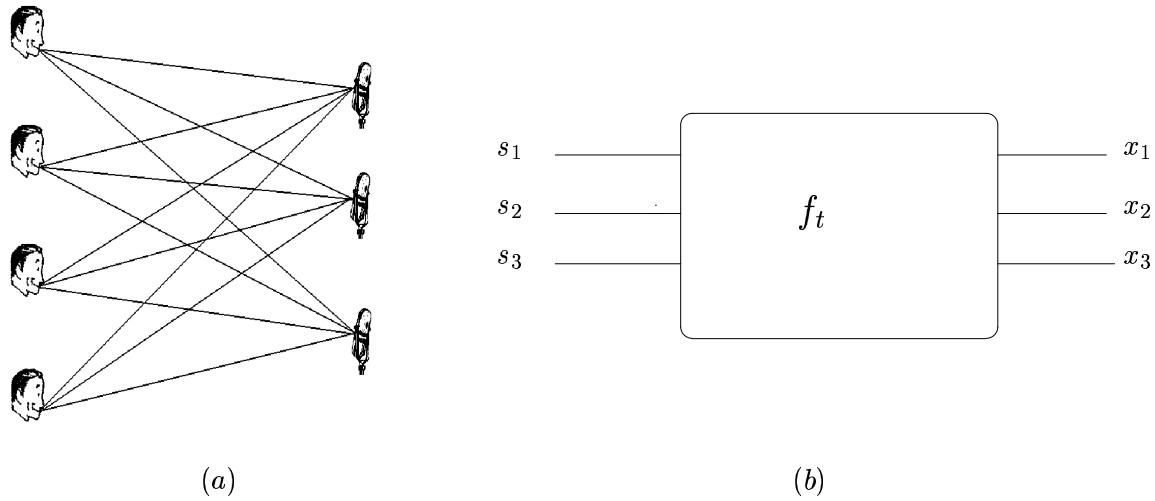


Fig. I.1: L'exemple traditionnellement repris dans la litérature est celui du "cocktail party" où plusieurs personnes parlent au même temps et les signaux sont mélangés sur les micros, à droite la modélisation de ce mélange

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

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

fonctions f_t qu'on suppose connues (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 **?**. 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 :

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

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 :

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

L'objectif est de restaurer les sources à partir des observations. La non connaissance de la matrice de mélange rend le problème mal posé (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 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, la décomposition en ondelettes..., dans l'objectif d'éliminer une certaine redondance et 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 est la recherche d'une base de signaux décorrélés. L'analyse en composantes indépendantes 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 eux 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 :

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

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

[3] ici, le temps indique un indice général, ça peut être le pixel d'une image ou le pixel d'une transformation temps fréquence...

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

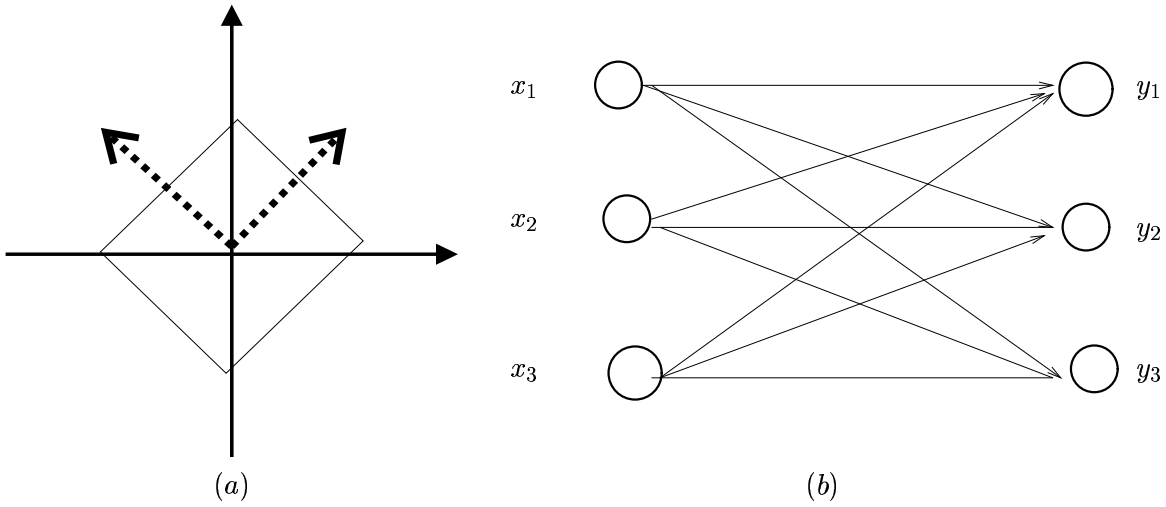


FIG. I.2: Recherche des directions indépendantes

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 système de contrôle où on cherche à produire une action (les sorties $\boldsymbol{y}_{1..T}$) connaissant les informations mesurées $\boldsymbol{x}_{1..T}$.

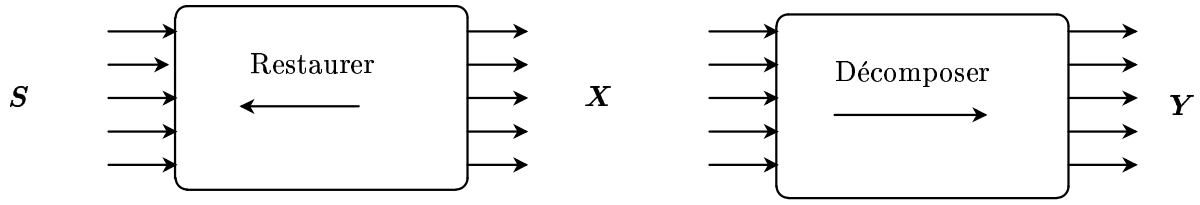


FIG. I.3: Restaurer ou décomposer ?

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

[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 échelle près à l'estimation des sources dans le modèle de mélange III.1 à condition que ces sources soient indépendantes et au plus l'une d'entre elles est 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 :*

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

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 III.1 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 applicant le théorème de Darmois pour des sources $\boldsymbol{s}$ indépendantes ayant au plus une composante gaussienne, 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) $\boldsymbol{C} = \boldsymbol{\Lambda}\boldsymbol{P}$, Λ une matrice diagonale et $\boldsymbol{P}$ une matrice de permutation

En assurant ainsi l'indépendance mutuelle (ou, moins restrictivement, 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 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 :

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t) + \boldsymbol{b}(t),$$

En effet, les données i.i.d $[\boldsymbol{x}(1), ..., \boldsymbol{x}(T)]$ suivent la loi p_x^* et on cherche une matrice $\boldsymbol{B}$ telle que les nouvelles données construites $[\boldsymbol{y}(t) = \boldsymbol{B}\boldsymbol{x}(t)]$ soient le plus possible indépendantes. 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 donc sous la contrainte que les $\boldsymbol{y}_{1..T}$ soient aussi 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 $\boldsymbol{B}$:

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

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 est alors la minimisation de la distance entre les deux ensembles $\mathcal{Q}^*$ et $\mathcal{P}_\Pi$ (voir figure I.4) :

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

L'existence et l'unicité des solutions sont alors directement liées aux géométries des deux surfaces $\mathcal{Q}^*$ et $\mathcal{P}_{Pi}$. Dans le cas où les observations $\boldsymbol{x}_{1..T}$ suivent un modèle de mélange linéaire non bruité, le théorème de Darmois assure l'existence et l'unicité de cette solution aux indéterminations d'échelle et de permutation près.

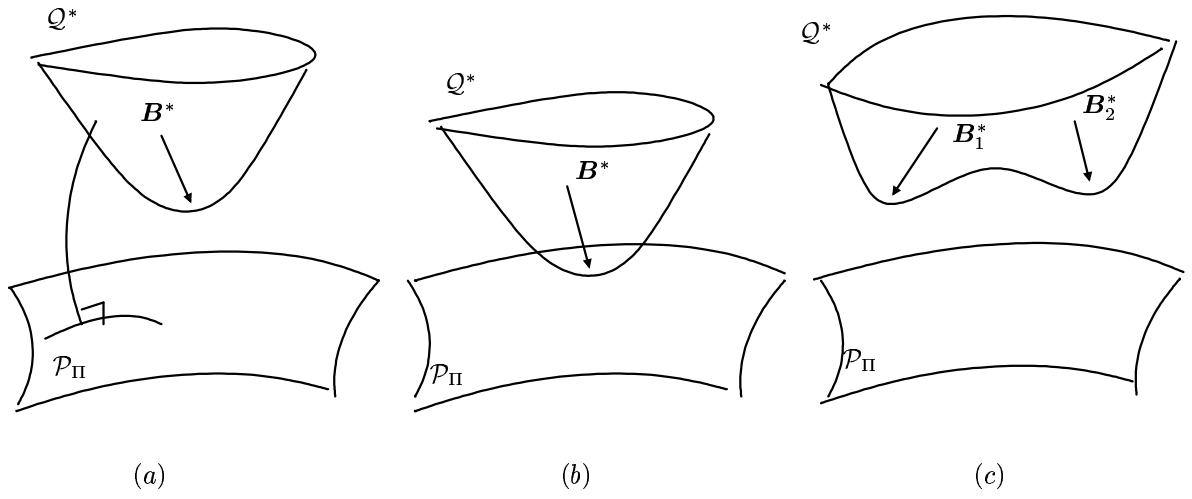


FIG. I.4: Le problème de minimisation de la distance entre les ensembles $\mathcal{Q}^*$ et $\mathcal{P}_\Pi$ dépend de la géométrie de la surface $\mathcal{Q}^*$ (qui dépend de la vraie loi p^* de $\boldsymbol{x}$)

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

[A].1 Minimisation de l'information mutuelle

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

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

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

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. La minimisation 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.

La minimisation de l'information mutuelle est alors équivalente à la minimisation de la somme des entropies marginales sous la contrainte que les composantes de $\boldsymbol{y}$ soient décorrélées. 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 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 connait pas la distribution $p^*_j(y_j)$ et on est alors amené à approximer $\mathrm{E}\big[\log p^*_j(y_j)\big]$. 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 pionier 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. Avec la minimisation de l'information mutuelle on doit aussi choisir une seule fonction $G = \log p_j^*(y_j)$ (qui n'est pas forcément optimale). 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, 2002a,b; Fry, 2002].

Les observations $\boldsymbol{x}$ subissent une opération linéaire $\boldsymbol{B}$ suivie d'une transformation non linéaire composant par composant Φ (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 qn'en prenant la fonction Φ comme la distribution cumulative des sources on retrouve le même algorithme de séparation en utilisant le maximum de vraisemblance [Cardoso, 1997].

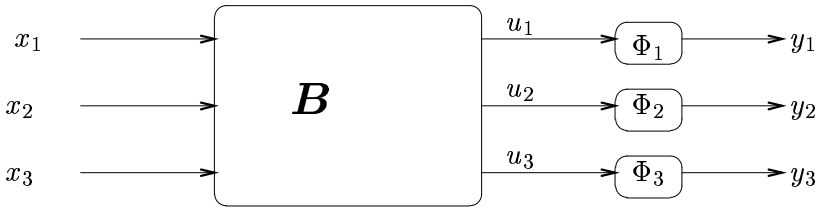


FIG. I.5: Infomax : maximiser le flux d'informations entre les entrées et les sorties du système

[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 non bruité :

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

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.

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

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

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

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'une 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[5]. 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.

[5] on trouve dans Pham et Cardoso [Pham et Cardoso, 2001] une discussion dans la section dans le cas gaussien qui clarifie ce point

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 présentent des avantages qui le distingue 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 :

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

   où p_b est la loi du bruit additif.
2. Les sources ne sont pas forcément indépendantes et toute information peut être contenue 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 $p(\boldsymbol{x}_{1..T} \mid \boldsymbol{A})$. On peut aussi considérer la matrice $\boldsymbol{A}$ comme une variable aléatoire et former la loi jointe :

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

où $\mathcal{H}$ représente notre connaissance *a priori* comme par exemple la forme des distributions des sources, leur indépendance...

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 chaines de Markov (MCMC).

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

[B].2 Méthodes tensorielles : 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 $\mathrm{E}\big[f(\boldsymbol{A}\boldsymbol{s})\big]$. Exploitant l'indépendance des sources et la linéarité du mélange il est souvent possible de se contenter aux cumulants d'ordre supérieure comme les cumulants d'ordre 4 pour identifier la matrice $\boldsymbol{A}$: c'est le principe des méthodes 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 $kurt(s_j)$. 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 [Cardoso et Souloumiac, 1993]. Elle se base sur le fait que la matrice $\boldsymbol{B}$ diagnalise toutes les matrices $\mathrm{C}(\boldsymbol{M})$ $\forall \boldsymbol{M}$. Un choix judicieux est de choisir 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 carrées 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 deux points :

1. L'équivalent des méthodes tensorielles en maximum de vraisemblance 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 Leibler-Kullback 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 aussi fixer les valeurs des statistiques des sources mais il faut les estimer dans l'algorithme de séparation.

I.2.2 Exploitation de la corrélation

Dans les méthodes précédentes, on a supposé que les sources sont 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'autocovariance. 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 sont différentes et non réduites à la distribution de Dirac δ.

[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 carrées de tous les termes est constante
[8] Ici l'hypothèse de l'i.i.d est considérée pour la dimension temporelle

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 :

   $$\boldsymbol{W}\boldsymbol{C}_j\boldsymbol{W}^* \approx \Lambda_j$$

   où $\boldsymbol{W}$ est la matrice de Fourier et Λ_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.*, 2002; 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 sont estimées. Nous allons revenir en détail sur cette méthode dans le chapitre **??**.

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 Leibler-Kullback), on peut essayer de se contenter de calculer les matrices d'autocovariance $\boldsymbol{R}_{xx}(\tau) = \mathrm{E}\left[\boldsymbol{x}(t)\boldsymbol{x}(t+\tau)^*\right]$ pour quelques retards temporels $\tau = 1..K$ 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 connait 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 stationarité

On peut aussi se baser sur la non stationarité des sources. Plus présicément, en supposant que les sources sont temporellement indépendantes mais que leurs variances varient en fonction du temps, on peut se limiter aux statistiques du second ordre pour séparer les sources. En effet, à chaque instant t, la matrice de covariance théorique du vecteur $\boldsymbol{x}$ se met sous la forme :

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

[9] En passant dans le domaine de Fourier, on va plutôt exploiter la non stationarité spectrale.

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 tensorielle (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 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 stationarité 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, on expose l'approche bayésienne en séparation de sources. On distingue l'aspect théorique de l'aspect technique de cette approche. Sur le plan théorique, cette approche présente plusieurs avantages :

1. En considérant la matrice de mélange comme une variable aléatoire munie d'une loi *a priori*, on peut exprimer nos connaissances physiques du système via cette loi et régulariser ainsi le problème de séparation. On peut aussi intégrer par rapport à cette matrice pour obtenir la loi marginale des sources.

[10] On rappelle que le temps est un indice générique qui peut aussi désigner le pixel d'une image, la fréquence, l'indice temps fréquence...

[11] après une étape de blanchiement

2. En considérant les hyperparamètres comme des variables aléatoires, on peut les régler automatiquement.
3. En introduisant des variables cachées, on peut enrichir la modélisation des sources.
4. On tient compte 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, on va considérer un mélange linéaire instantané bruité. Le point commun de ces chapitres est l'exploitation de la non stationarité que ce soit dans le domaine temporel, spatial, fréquentiel et temps fréquence. Les algorithmes proposés intègre implicitement la reconstruction des sources contrairement aux méthodes estimant la matrice de mélange en ajustant les statistiques.

Dans le chapitre III, on considère des sources 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. 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 présentons alors l'algorithme EM pénalisé pour l'estimation de la matrice de mélange et les paramètres des sources. Nous étudions aussi le cas des sources modélisées par des chaînes de Markov cachées (les étiquettes forment une chaîne de Markov) en implémentant l'algorithme EM exact. Des versions sous optimales ont été implémentées pour accélérer l'EM. La représentation hiérarchique des sources par l'introduction des variables cachées peut être interprétée comme une exploitation de la non stationarité des variances 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, on va exploiter la non stationarité 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[12]. 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. Après avoir étudié l'identifiabilité de la matrice de mélange ainsi que les autres paramètres intervenant dans le problème, 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.

Dans le chapitre **??**, on va exploiter la non stationarité 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 de divergence 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 en 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

[12] 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

séparer des composantes astrophysiques en l'implémentant dans le domaine de Fourier et dans le domaine des harmoniques sphériques.

Dans le chapitre **??**, on va exploiter la non stationarité dans le domaine temps fréquence en séparant des sources localement stationnaires. On aura deux sortes de classifications :

1. Classification temporelle : On divise l'intervalle $\mathcal{I} = [1\ T]$ en trames $\mathcal{I}_\tau$ ($\tau = 1..T'$) qui se chevauchent. Dans chaque trame, le signal est supposé stationnaire et donc admet un spectre. On suppose qu'il y a K spectres possibles. Donc on va classifier les trames en K classes.
2. Classification fréquentielle : De la même manière que dans le chapitre précédent, avec l'hypothèse circulante, on divise l'intervalle fréquentiel $[1\ |\mathcal{I}_\tau|]$ en sous intervalles où les variances sont fixes mais qui varient d'un sous-intervalle à l'autre.

Les applications concernent la séparation des signaux de musique.

Le chapitre **??** traîte en détail le problème de dégénérescence du maximum de vraisemblance. On généralise des résultats qui existent déjà dans le cas de mélange de gaussiennes monovariées au cas multivariable. La dégénérescence est produite quand les matrices de covariance approchent des matrices non régulières (points de singularité). L'élimination de cette dégénérescence est guarantie par l'utilisation d'un *a priori* Wishart inverse 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 les modèles de Markov cachés). La pénalisation par un *a priori* inverse Wishart élimine également cette dégénérescence.

Le chapitre **??** est consacré au problème de la sélection de la loi *a priori* dans un contexte bayésien. Nous présentons une approche originale 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 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 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. on Sig. Proc.*, 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. on Sig. Proc.*, 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, 2002] 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 2002.

[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, 2002a] 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 2002.

[Knuth, 2002b] K. Knuth. What is a question ? In *Bayesian Inference and Maximum Entropy Methods*. MaxEnt Workshops, à paraître dans Amer. Inst. Physics, août 2002.

[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.

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

[Pham et Cardoso, 2001] D.-T. Pham et J. Cardoso. Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. on Sig. Proc.*, 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.

[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, 2002] 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.*, 2002] 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 2002.

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

[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

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 autoconsistance. La règle de Bayes est 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 ont permis de mettre en oeuvre la méthodologie bayésienne. Parmi ces techniques, on décrit l'algorithme *Expectation-Maximization* (EM) et les méthodes de Monte Carlo par Chaînes de Markov (MCMC). 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 $\boldsymbol{x}_{1..T}$. Le processus générant les données est en effet géré par les lois de la physique qui sont des lois universelles. On suppose de plus que ce processus physique est une composition de transformations de signaux d'origine $\boldsymbol{s}_{1..T}$. On peut alors modéliser ce processus par un système physique $\mathcal{M}$. Les données observées $\boldsymbol{x}_{1..T}$ sont les signaux de sortie de $\mathcal{M}$. Elles résultent d'une transformation $\mathcal{F}$ des signaux d'entrées $\boldsymbol{s}_{1..T}$:

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

La distinction entre les entrées et les sorties d'un système est parfois ambigue. Cette ambiguité est en général soulevée en considérant le sens entrées sorties le sens dans lequel on perd de l'information. Le système n'est 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. L'information recherchée $\boldsymbol{\theta}$.
3. Toutes les informations $\boldsymbol{h}$ 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 $\mathcal{M}$. 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 optimale profite de la présence de tels paramètres en assurant que le résultat final n'en dépend pas.

$\boldsymbol{h}$ représente toutes les informations *a priori* qu'on possède sur le problème. On distingue ce paramètre des données $\boldsymbol{x}_{1..T}$. En effet, $\boldsymbol{h}$ 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 $\mathcal{M}$ comme la nature de la famille paramétrique $f_{\boldsymbol{\theta}}$, 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 $\boldsymbol{h}$. Si $\boldsymbol{\theta}$ représente les signaux d'entrée alors l'inférence est une reconstruction. Si $\boldsymbol{\theta}$ représente les paramètres du système $\boldsymbol{\theta}$ alors l'inférence est une sorte de prédiction où les données $\boldsymbol{x}_{1..T}$ servent à prédire le vecteur $\boldsymbol{x}_{T+1}$ à l'instant $T+1$. La prédiction se distingue de la reconstruction dans le sens où son objectif n'est pas la reconstruction 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. On considère 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 recherchées sont $\boldsymbol{\theta}$! "
- $\boldsymbol{h}$ est la proposition : " l'information *a priori* est $\boldsymbol{h}$! "

Dans le cas parfait, la valeur de la proposition $\mathcal{I} = (\boldsymbol{x} \wedge \boldsymbol{h} \longrightarrow \boldsymbol{\theta})$ peut être construite avec les règles de la logique[1] et vaudra 0 ou 1. Si $\boldsymbol{x}$, $\boldsymbol{h}$ et $\boldsymbol{\theta}$ varient dans leurs espaces respectifs $\mathcal{X}$, $\mathcal{H}$ et Θ,

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

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 $\boldsymbol{h}$, la solution du problème d'inférence est la valeur $\hat{\boldsymbol{\theta}}$ telle que la proposition $\mathcal{I} = (\boldsymbol{x} \wedge \boldsymbol{h} \longrightarrow \hat{\boldsymbol{\theta}})$ soit vraie (vaut 1). Malheureusement, la proposition $\mathcal{I} = (\boldsymbol{x} \wedge \boldsymbol{h} \longrightarrow \boldsymbol{\theta})$ ne peut pas être évaluée avec exactitude (on ne peut pas affirmer si elle est vraie ou fausse). Ceci revient à, au moins, trois raisons :

1. L'information *a priori* $\boldsymbol{h}$ 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 $\mathcal{I} = (\boldsymbol{x} \wedge \boldsymbol{h} \longrightarrow \boldsymbol{\theta})$ 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 [**???**]). Le problème d'inférence (ou de prédiction) est alors complètement décrit par la fonction $\mathrm{P}_r(\boldsymbol{x} \wedge \boldsymbol{h} \longrightarrow \boldsymbol{\theta})$. Autrement dit, la quantité $\mathrm{P}_r(\boldsymbol{x} \wedge \boldsymbol{h} \longrightarrow \boldsymbol{\theta})$ 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. $\boldsymbol{x}_{1..T}$, $\boldsymbol{h}$ et $\boldsymbol{\theta}$ représentent trois propositions et on peut 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* $\mathrm{P}(A)$ *n'existe pas. Cette notion existe au sens de la théorie fréquentiste où* A *n'est pas une proposition mais plutôt un événement et* $\mathrm{P}(A)$ *est la fréquence de cet événement dans une infinité de réalisations. On trouve parfois dans la litérature de l'inférence logique des termes comme* $\mathrm{P}(A)$ *mais rigoureusement ça représente* $\mathrm{P}(H \longrightarrow A)$ *où* H *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 applicant la règle de produit (conséquence de l'associativité) :

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

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

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

L'incertitude sur la proposition de l'inférence $\mathcal{I} =: (\boldsymbol{x}_{1..T} \wedge \boldsymbol{h} \longrightarrow \boldsymbol{\theta})$ 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 $\mathrm{d}\,\mathrm{P}(B \longrightarrow A) = \mathrm{P}(B \longrightarrow A \in \mathcal{V}(A))$. En changeant les notations $\mathrm{d}\,\mathrm{P}(B \longrightarrow A)$ par $\mathrm{d}\,\mathrm{P}(A \mid B)$ et la conjonction $(A \wedge B)$ par $(A,\, B)$, le théorème de Bayes II.1 s'écrit :

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

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) :

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

$\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{h})$ est la densité *a priori* de $\boldsymbol{\theta}$ et $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}, \boldsymbol{h})$ 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 $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{h})$ est l'évidence des données. Il peut être interprété comme un coefficient de normalisation en imposant que $\int \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{h}) d\,\boldsymbol{\theta} = 1$ [2].

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

1. Au niveau fondamentale :
   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 language, on qualifie $\boldsymbol{\theta}$ de variable aléatoire mais ceci ne présente aucun jugement sur sa nature alétoire 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* $\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{h})$ 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

[2] le fait que la somme totale $\int p$ soit une constante est une conséquence de la consistance logique des probabilités
[3] ce qui va aboutir à une discussion sur la nature de ce monde.

coût $\mathrm{C}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ et l'estimateur $\hat{\boldsymbol{\theta}}$ est le minimiseur de l'espérance *a posteriori* de ce coût :

$$\hat{\boldsymbol{\theta}} = \arg\min \int \mathrm{C}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \mathrm{p}(\boldsymbol{\theta}^* \mid \boldsymbol{x}_{1..T}, \boldsymbol{h}) d\boldsymbol{\theta}^*$$

ce qui revient à considérer une caractéristique particulière de la distribution *a posteriori* comme la moyenne, la médiane, le maximum,...

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* $\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{h})$. Autrement dit, ayant l'information *a priori* $\boldsymbol{h}$, quelle est le degré de plausibilité de la proposition $\boldsymbol{h} \longrightarrow \boldsymbol{\theta}$?. 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 la question : quelle vraisemblance $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}, \boldsymbol{h})$ doit-on choisir ?.

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 $p(A \mid B)$ est l'incertitude de la proposition $B \longrightarrow A$ et la question qui se pose est quelle est la valeur de cette incertitude ?

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 $p(A \longrightarrow B)$ 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 à A. Cependant, cette théorie n'est pas déconnectée de la démarche scientifique. L'utilisateur doit respecter la consistance du calcul des probabilités. Par exemple, si $p(A \mid B)$ est fixée alors $p(\overline{A} \mid B)$ ne peut pas être fixée librement et doit être égale à $1 - p(A \mid B)$.
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 $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}, \boldsymbol{h})$ et $\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{h})$. 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 $\boldsymbol{\theta}$. Le choix est en pratique 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 [**???**].

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 :

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

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

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

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 $(\boldsymbol{x}_{1..T} \longrightarrow \boldsymbol{\theta} \wedge \boldsymbol{c})$ est plus simple à modéliser. D'un point de vue logique, ceci revient à intercaler une proposition $\boldsymbol{c}$ entre l'information *a priori* $\boldsymbol{h}$ et le paramètre $\boldsymbol{\theta}$ pour faciliter le raisonnement logique ou à compléter l'information dans $\boldsymbol{\theta}$ pour expliquer les données $\boldsymbol{x}_{1..T}$:

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

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 $\boldsymbol{\theta}$:

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

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 :

$$(\boldsymbol{h} \longrightarrow \boldsymbol{\theta}) \rightsquigarrow (\boldsymbol{h} \longrightarrow \boldsymbol{c}_0 \longrightarrow \boldsymbol{\theta}) \rightsquigarrow (\boldsymbol{h} \longrightarrow \ldots \longrightarrow \boldsymbol{c}_1 \longrightarrow \boldsymbol{c}_0 \longrightarrow \boldsymbol{\theta})$$

Cette structure hiérarchique facilite le choix des probabilités intervenant dans le calcul de la distribution *a posteriori*. Ainsi, au lieu de choisir la vraisemblance $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}, \boldsymbol{h})$, on construit plutôt la fonction $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{h})$ en profitant de l'augmentation de l'information. Le choix de l'*a priori* est aussi facilité par la décomposition logique de $(\boldsymbol{h} \longrightarrow \boldsymbol{\theta})$ en $(\boldsymbol{h} \longrightarrow \boldsymbol{c} \longrightarrow \boldsymbol{\theta})$. 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. Bien qu'on réussisse à

obtenir une forme analytique de la distribution *a posteriori*, le calcul de ses caractéristiques (comme le maximum, 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'approximer 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 recherchée $\boldsymbol{\theta}$. L'estimé de $\boldsymbol{\theta}$ est alors le minimiseur de la moyenne *a posteriori* du coût $C(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$:

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

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] :

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

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

Le calcul et l'optimisation de la vraisemblance $\mathrm{p}(\boldsymbol{x}_{1..T} \mid \boldsymbol{\theta}, \boldsymbol{h})$ sont en général compliqués du fait de la présence de l'intégration. L'algorithme EM[**?**] est un algorithme itératif qui permet d'approximer numériquement la solution de II.6. Le principe de l'EMest 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 $(\boldsymbol{\theta}, \boldsymbol{c})$:

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

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

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** (Maximisation) :
   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 EMest que la distribution *a posteriori* marginalisée $\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{h})$ croit à chaque itération (monotonie) :

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

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

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

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 :

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

Dans le cas du maximum de vraisemblance ($\mathrm{p}(\boldsymbol{\theta}) \propto cte$), des conditions de continuité [**?**] de la fonctionnelle $\mathcal{Q}$ guarantissent 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 [**???**] 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) [**??**].

[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).

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'approximer la densité *a posteriori* par des fonctions simples. L'approximation normale est la plus naturelle du fait de sa validité asymptotique [**??**]. 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 $(h(\boldsymbol{\theta})f(\boldsymbol{\theta}))$ [**???**]. Cependant, ces approximations ne peuvent 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 $h(\boldsymbol{\theta})$:

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

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 [**?**] des conditions suffisantes sur g pour guarantir 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 guarantir 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 **?**.

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 :

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

$\mathrm{P}(\boldsymbol{Y}_n \mid \boldsymbol{Y}_{n-1}$ est appelée le noyau de transition $\mathcal{K}_n(. \mid .)$ de la chaîne. La chaîne de markov est dite homogène si le noyau de transition ne varie pas avec l'indice n, on le note $\mathcal{K}(. \mid .)$. Le noyau

[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 $g(\boldsymbol{\theta}) = \frac{|h(\boldsymbol{\theta})| f(\boldsymbol{\theta})}{\int |h(\boldsymbol{\theta})| f(\boldsymbol{\theta}) d\boldsymbol{\theta}}$ [**?**]

$\mathcal{K}(. \mid .)$ peut être considérée 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 litérature [**?**]. Le chapitre 3 de l'ouvrage [**?**] et [**?**] représentent une bonne synthèse de ce domaine. L'annexe B de [**?**] 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 guarantit 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 [**?**] 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 :

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

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 $\mathcal{K}^m(. \mid \boldsymbol{y}_0)$:

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

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.

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 les 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

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

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 :

$$\mathcal{K}_G(\boldsymbol{y}^{m+1} \mid \boldsymbol{y}^m) = \pi_1^*(y_1^{m+1} \mid y_2^m, y_3^m, ..., y_p^m)\, \pi_2^*(y_2^{m+1} \mid y_1^{m+1}, y_3^m, y_4^m, ..., y_p^m)$$

$$... \pi_p^*(y_p^{m+1} \mid y_1^{m+1}, y_2^{m+1}, ..., y_{p-1}^{m+1})$$

[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.

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 traîtés 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 :

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

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 [**??**] 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 $\boldsymbol{z}$ ($\boldsymbol{y}_{m+1} = \boldsymbol{z}$) ou on le rejette en gardant la valeur $\boldsymbol{y}_m$ ($\boldsymbol{y}_{m+1} = \boldsymbol{y}_m$) 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 :

   - 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 $\mathcal{K}_{MH}$ [**?**]. L'irréductibilité et l'apériodicité sont à étudier selon le contexte (la distribution π^* et la distribution instrumentale $g(. \mid .)$) [**?**].

Version hybride : L'échantillonnage de Gibbs présente des avantages et des inconvénients par rapport à l'algorithme de Metropolis-Hastings. Son principale 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 $\boldsymbol{y}$.

[11] on peut parfois corriger cet inconvénient par une reparamétrisation ou un repartitionnement 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 [**??**] consiste à reprendre l'algorithme de Gibbs mais en applicant 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

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

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.

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 [**?**] 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 pré-définis. 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

instantanné 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 $(m \times n)$. t est un indexe générique. Il désigne le temps dans le chapitre III, le pixel d'une image dans le chapitre IV, la fréquence dans le chapitre **??** et l'indice temps-fréquence dans le chapitre **??**.

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 : définition du problème d'inférence, construction de la distribution *a posteriori* avec la règle de Bayes, choix des probabilités (vraisemblance et *a priori*), choix du critère et finalement choix de l'algorithme d'optimisation.

Estimation de $\boldsymbol{A}$: Le problème d'inférence considéré est $\mathcal{I} := (\boldsymbol{h} \wedge \boldsymbol{x}_{1..T} \longrightarrow \boldsymbol{A})$ où $\boldsymbol{h}$ 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 :

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

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

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

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

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 III.3, on doit choisir les probabilités p_ϵ, p_s et p_A. Bien qu'en pratique, on adopte le point de vue subjectif 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 EMou l'échantillonnage de Gibbs.

Remarque 5
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 III.3 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 $\mathcal{I} := (\boldsymbol{x}_{1..T}, \boldsymbol{h} \longrightarrow \boldsymbol{s}_{1..T})$. Connaissant les données $\boldsymbol{x}_{1..T}$ et l'information *a priori* $\boldsymbol{h}$ (incluant le fait que les données $\boldsymbol{x}_{1..T}$ suivent le modèle de mélange linéaire instantanné bruité III.1), l'objectif est de reconstruire les sources $\boldsymbol{s}_{1..T}$.

[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 [?] peut être conservée pour une certaine classe de lois *a priori* sur la matrice $\boldsymbol{B}$

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

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

Le modèle direct ($\boldsymbol{s}_{1..T} \wedge \boldsymbol{h} \longrightarrow \boldsymbol{x}_{1..T}$) 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 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 :

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

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

Estimation conjointe : Le problème d'inférence est défini par $\mathcal{I} := (\boldsymbol{x}_{1..T} \wedge \boldsymbol{h} \longrightarrow \boldsymbol{s}_{1..T} \wedge \boldsymbol{A})$. 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 approxime 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 7 *L'un des avantages de l'algorithme* MCMC *II.6 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 litérature de la séparation de sources[15] est le choix de la densité *a priori* des sources $\mathrm{p}_s(\boldsymbol{s}_{1..T} \mid \boldsymbol{h})$. Ce choix doit tenir compte de deux 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 guarantir l'identifiabilité du modèle (dont 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 :

$$\boldsymbol{h} \longrightarrow \boldsymbol{s}_{1..T} \rightsquigarrow \boldsymbol{h} \longrightarrow \boldsymbol{z}_{1..T} \longrightarrow \boldsymbol{s}_{1..T}$$

[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.

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

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

où la densité conditionnelle $\mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{z}_{1..T}, \boldsymbol{h})$ 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. [???]. 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,

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 $\mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{z}_{1..T}, \boldsymbol{h})$ gaussiennes et un bruit blanc gaussien, la séparation va se baser sur des statistiques d'ordre deux en exploitant la non stationarité des sources.

Prédiction : Le problème d'inférence est défini par $\mathcal{I} := (\boldsymbol{x}_{1..T} \wedge \boldsymbol{h} \longrightarrow \mathrm{p}(\boldsymbol{x}_{t+1}))$. Connaissant les données $\boldsymbol{x}_{1..T}$, notre objectif est de prédire la densité de l'observation $\boldsymbol{x}_{t+1}$ à l'instant $(t+1)$. La variable à manipuler est donc de type 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 $\mathrm{P}_r(\boldsymbol{x}_{1..T} \mid \mathrm{p}) = \mathrm{p}(\boldsymbol{x}_{1..T})$ 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 [?]. 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 [?],

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

[16] ceci nécessite la manipulation des outils de la géométrie différentielle. Une grande partie du chapitre ?? est consacrée à ces notions

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 Leibler-Kullback),* $\hat{\mathrm{q}}$ *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 :

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 fondamentale 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 EMou les méthodes MCMC assurent l'implémentation efficace de la méthode bayésienne lorsque l'étape 3 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 supportée par l'approche bayésienne au niveau pratique et technique, va jouer un rôle important pour simplifier l'implémentation des algorithmes de séparation et pour guarantir l'identifiabilité du mélange.

Bibliographie

Chapitre III

SÉPARATION DE SOURCES MONOVARIÉES : NON STATIONARITÉ 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 [??] 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. Le modèle markovien peut être considéré comme un terme de régularisation de 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 [?]. 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 [**???**] 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 [**?**] est de contraindre les variances à appartenir à un intervalle strictement positive. Cette contrainte complique la mise en oeuvre de l'estimation des variances. Récemment, une solution se basant sur l'approche bayésienne a été proposée dans [**?**] 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 [**?**], 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 **??** 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 cœfficients 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. On commence par établir les équations de ré-estimation de l'algorithme EMen utilisant la

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

procédure de Baum-Welsh et discuter le coût de son implémentation. Ensuite, on présente 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'EMdû à la structure temporelle de la 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 EMa été utilisé dans [**??**] pour séparer des sources modélisées par des mélanges de gaussiennes. Dans ces travaux, on peut montrer que :

- L'algorithme EMn'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'EMest 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 essayé d'apporter les éléments suivants :

- 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 cœfficients 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 tous 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) :

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

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.22 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 :

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

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 $P_{lk} = \left[P(z_{t+1}^j = k \,|\, z_t^j = l)\right]_{l,k=1..K_j}$. Conditionnellement à cette chaîne, la source s^j est temporellement blanche :

$$p(s_{1..T}^j \,|\, z_{1..T}^j) = \prod_{t=1}^{T} p(s_t^j \,|\, z_t^j) \tag{III.4}$$

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

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.
- Elle est appliquée avec succès dans le traitement des signaux de parole. Des modèles de HMMplus élaborés peuvent être trouvés dans [**?**].

La loi *a priori* de la j^{me} source s'écrit alors :

$$\mathrm{p}_s(\boldsymbol{s}_{1..T}^j \mid \boldsymbol{\eta}_s^j) = \sum_{\boldsymbol{z}_{1..T}} Pr(\boldsymbol{z}_{1..T}^j \mid \boldsymbol{\eta}_p^j) \prod_{t=1}^{t=T} p(s_t^j \mid z_t^j, \boldsymbol{\eta}_g^j)$$

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

[B] Modélisation de la matrice de mélange

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

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

Ce choix est motivé par les raisons suivantes :

- On peut interpréter facilement cette loi. C'est comme on connaît la valeur M_{ij} du cœfficient A_{ij} de la matrice de mélange avec une incertitude $\sigma_{a,ij}^2$.
- 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 **??** est entièrement consacré au choix de l'*a priori* avec les outils de la géométrie de l'information.

Exemple 3 *Dans certaines applications [?], 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 cœfficient 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 [?] 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 **??** 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 $\boldsymbol{z}_{1..T}$:

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

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 :

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

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 $\mathcal{T}_k = \{t \mid \boldsymbol{z}_t = \boldsymbol{k}\}$:

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

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. Le 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 $\hat{\boldsymbol{R}}_k = \sum_{\mathcal{T}_k} \boldsymbol{x}_t \boldsymbol{x}_t^* / T_k$,

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

où les α_k représentent les proportions des classes.

Le 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) :

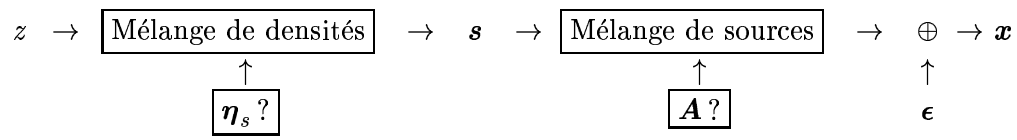
$$\frac{\log \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x}_{1..T}, \boldsymbol{z}_{1..T})}{T} = \underbrace{\sum_{k=1}^{K} \alpha_k D_{KL}(\boldsymbol{R}_k, \hat{\boldsymbol{R}}_k)}_{\text{Ajustement des covariances}} + \underbrace{\log \mathrm{p}(\boldsymbol{\theta}) \;/\; T}_{\text{Terme de régularisation}}$$

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 op'erations 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[**?**]. La propriété fondamentale de cet algorithme est qu'il assure la croissance monotone de la distribution *a posteriori* incomplète. Toute valeur $\boldsymbol{\theta}$ feasant 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 EMsont donnés dans le chapitre II ou [**?**].
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 EMet de l'algorithme SEM (Stochastic EM[**?**]). 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 EMexact

La fonctionnelle $\mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^k) = \mathrm{E}\left[\log p(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{z} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \mid \boldsymbol{x}, \boldsymbol{\theta}^k\right]$, 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 :

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

où $(*)$ désigne le 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 [**?**], $\odot$ 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 obtenues :

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

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 [**?**] 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 :

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

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

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

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 :

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

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 $diag\left(\sigma^2_{i_1}, \sigma^2_{i_2} \dots \sigma^2_{i_n}\right)$.

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

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

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 :

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

Dans la deuxième expression de (III.18), on note que la pénalisation de la vraisemblance par un inverse gamma 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 :

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

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

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

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 marginalisation 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, $\widehat{z}^k_t$ maximise $p(z_t \,|\, \boldsymbol{x}_{1..T},\ \widehat{z}^k_{i<t},\ \widehat{z}^{k-1}_{i>t})$, ce qui donne pour $t = 1..T$:

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

et

$$\begin{cases} z^k_1 = \underset{l=1..K}{\arg\max}\, \phi(\boldsymbol{x}_1 \,|\, \boldsymbol{\theta}_l, \boldsymbol{A}^k) \boldsymbol{T}_{[l, z^{k-1}_2]} \\ \\ z^k_T = \underset{l=1..K}{\arg\max}\, \boldsymbol{T}_{[z^k_{T-1}, l]} \phi(\boldsymbol{x}_T \,|\, \boldsymbol{\theta}_l, \boldsymbol{A}^k) \end{cases}$$

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'EMsont 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, $\widehat{z}^k_t \sim p(z_t \,|\, \boldsymbol{x}_{1..T},\ \widehat{z}^k_{i<t},\ \widehat{z}^{k-1}_{i>t})$, ce qui donne pour $t = 1..T$:

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

et

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

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 $[p(z_t = i \mid \boldsymbol{x}_{1..T}, z_{s\neq t})]_{i=1..K}$.

Le but des versions Viterbi et Gibbs de l'algorithme EMest 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^2T ($K = \prod_{l=1}^n K_l$) de la procédure Forward-Backward est réduite à KT (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 $P(\boldsymbol{z}_t = \boldsymbol{i} \mid \boldsymbol{x}_{1..T}, \boldsymbol{\theta})$ dans les trois algorithmes EM, Viterbi-EMet 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 :

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

On note, d'après la deuxième ligne de l'équation plus haut que c'est la distribution $p(\boldsymbol{x} \mid \boldsymbol{s}, \boldsymbol{\theta})$ 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 :

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

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 :

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

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 $p(\boldsymbol{z}_t \mid \boldsymbol{x}_t, \boldsymbol{\theta})$ dans la version Viterbi ou Gibbs, nous avons, avec la stratégie de relaxation, seulement $n \times k$ probabilités $\left(p(z_j(t) \mid \boldsymbol{x},\, \boldsymbol{\theta}', \widehat{s}_{l \neq j})\right)_{z=1..k}^{j=1..n}$ 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-vari'ees 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*

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

$$et$$

$$\begin{cases} z_j(1)^k = \underset{l=1..K_j}{\arg\max}\, \phi(\boldsymbol{x}_1 \mid s_{l \neq j}, \boldsymbol{\theta}_l, \boldsymbol{A}^k) \boldsymbol{T}_{[l, z^{k-1}_{j,2}]} \\ z_j(T)^k = \underset{l=1..K_j}{\arg\max}\, \boldsymbol{T}_{[z^k_{j,T-1}, l]}\, \phi(\boldsymbol{x}_T \mid s_{l \neq j}, \boldsymbol{\theta}_l, \boldsymbol{A}^k) \end{cases} \tag{III.23}$$

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*

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

$$et$$

$$\begin{cases} z_j(1) \sim \phi(\boldsymbol{x}_1 \mid s_{l\neq j}, \boldsymbol{\theta}_z, \boldsymbol{A}^k)T_{z_1 z_2} \\ z_j(T) \sim T_{z_{T-1}z_T}\phi(\boldsymbol{x}_T \mid s_{l\neq j}, \boldsymbol{\theta}_z, \boldsymbol{A}^k) \end{cases} \tag{III.24}$$

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 garder le même état. Cependant, la matrice $\boldsymbol{T}_2$ possède la même ligne. Ce qui signifie que le mélange est i.i.d.. La figure IV.5 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 IV.6 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 [**?**] :

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

La figure IV.7 illustre l'évolution, au cours des itérations, des estimées par l'algorithme EMexact des cœfficients de la matrice de mélange. La ligne horizontale indique la vraie valeur du cœfficient. 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 EMexact 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 IV.11). La figure (b) de IV.7 illustre la convergence de l'indice de performance de l'algorithme EMvers une valeur satisfeasante de -31 dB. La figure IV.8 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 IV.9 et IV.10 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 EMsur la figure gauche de IV.7). 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 IV.11 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'EMau cours des itérations [**?**]. 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 EMexact : 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

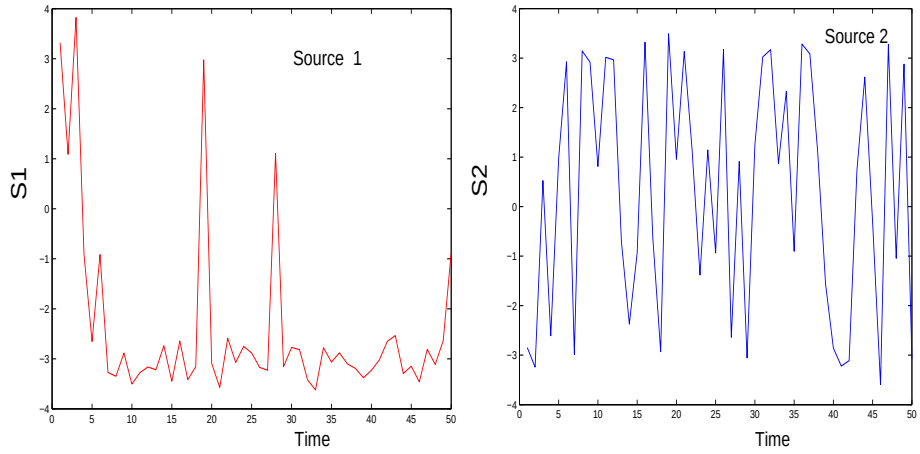


FIG. III.1: Graphes des sources s_1 et s_2. Seuls les 50 premiers échantillons sont montrés.

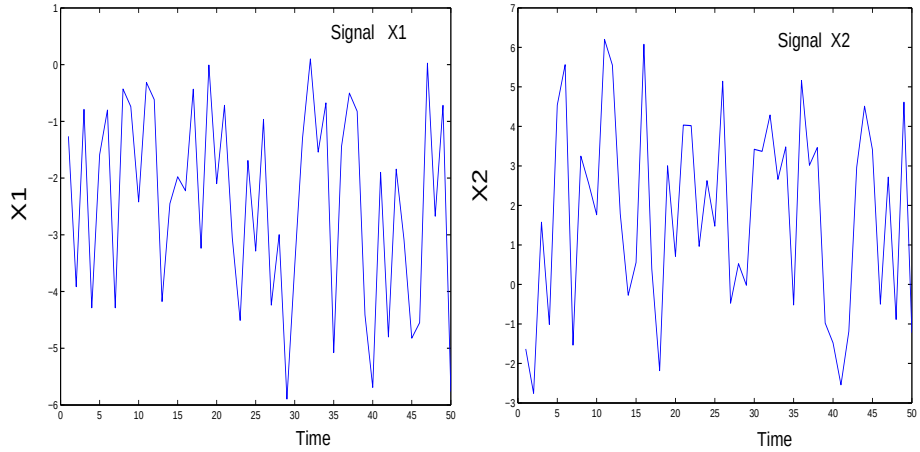
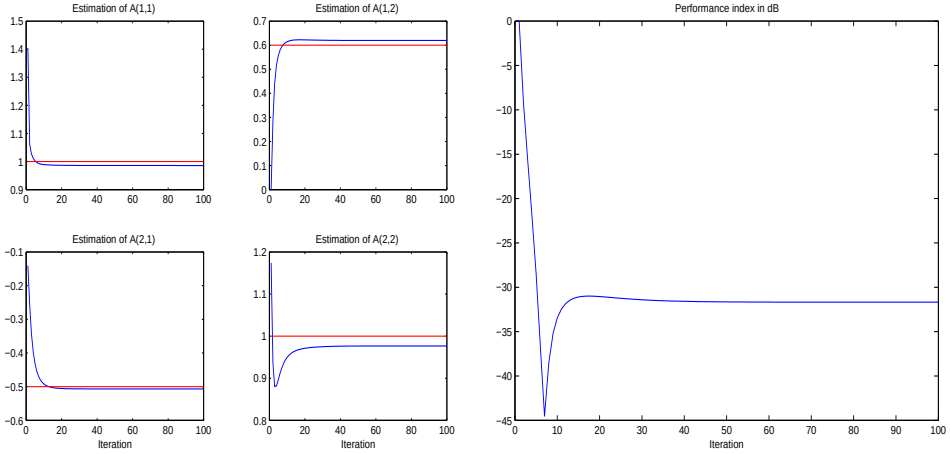


FIG. III.2: Graphes des sources mélangées $X_1 = a_{11}S_1 + a_{12}S_2$ et $X_2 = a_{21}S_1 + a_{22}S_2$



a b

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

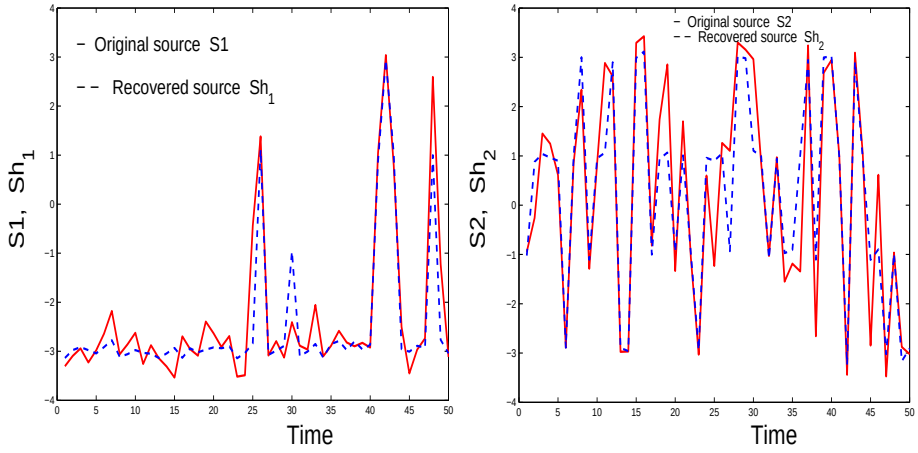
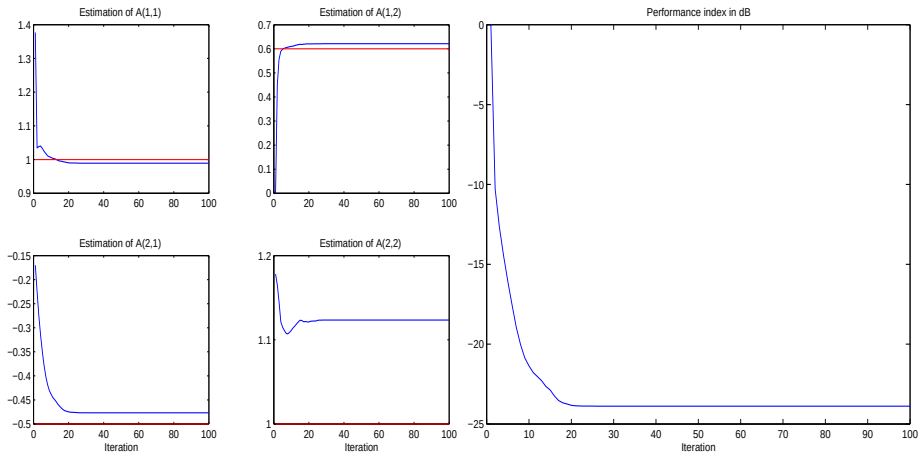


FIG. III.4: Résultats de reconstruction des sources avec l'algorithme EM



a b

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

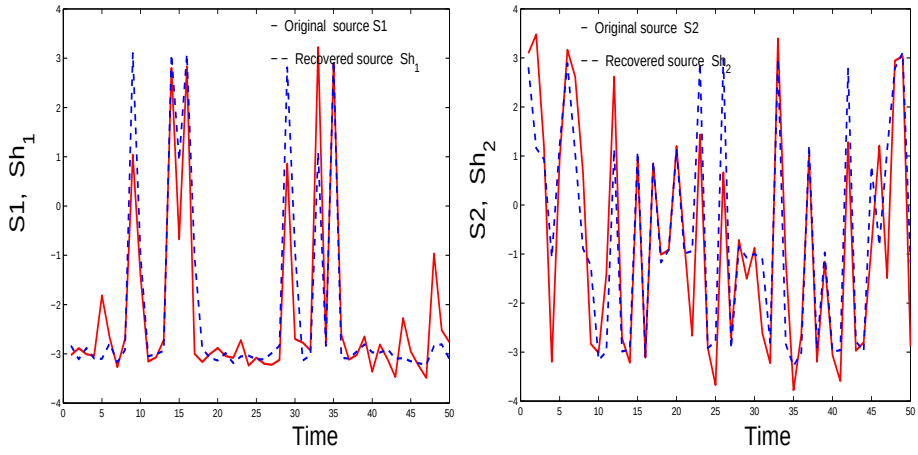
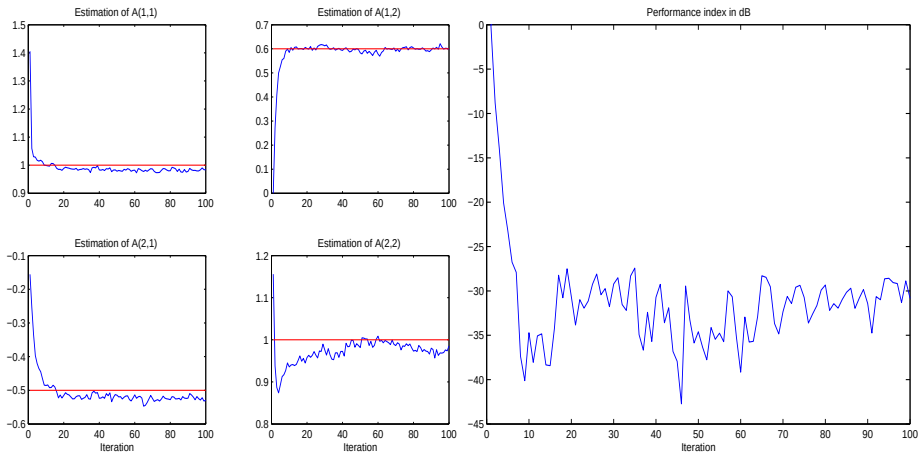


FIG. III.6: Résultats de reconstruction des deux sources avec l'algorithme *Viterbi-EM*.



a b

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

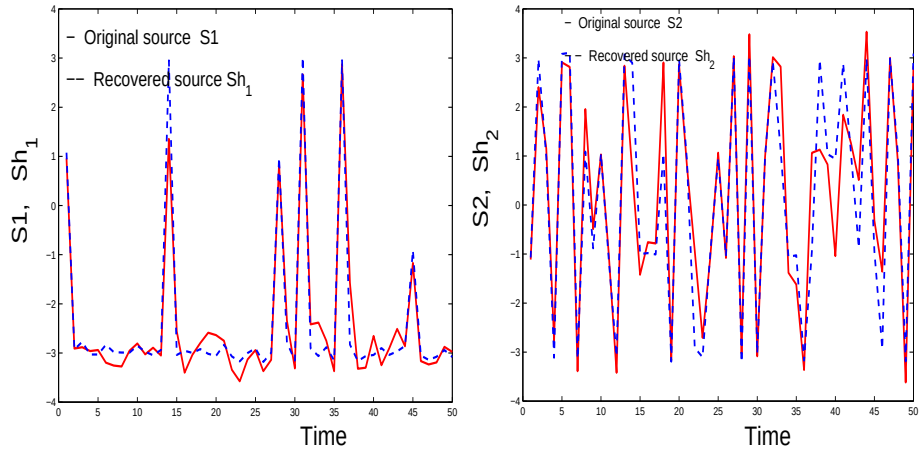


FIG. III.8: Résultats de reconstruction des deux sources avec l'algorithme *Gibbs-EM*.

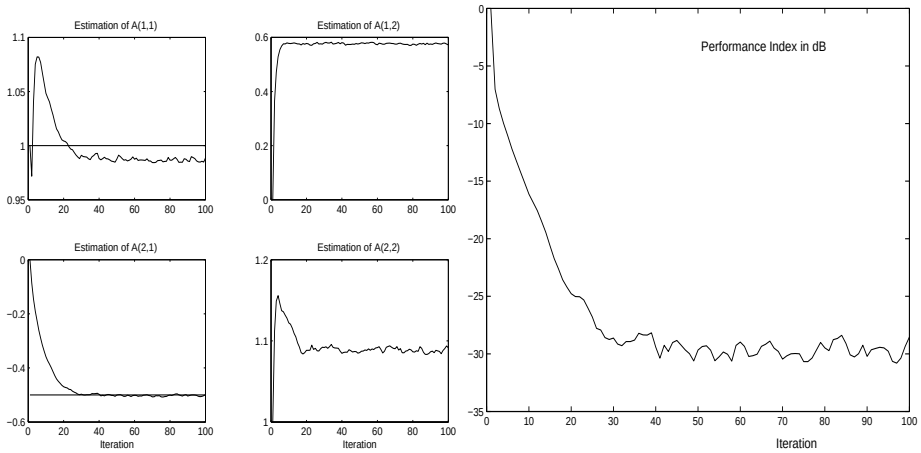


FIG. III.9: (a) Evolution au cours des itérations des estimées des cœfficients de mélange avec l'algorithme *Fast-Viterbi-EM*, (b) Evolution de l'indice de performance avec *Fast-Viterbi-EM*.

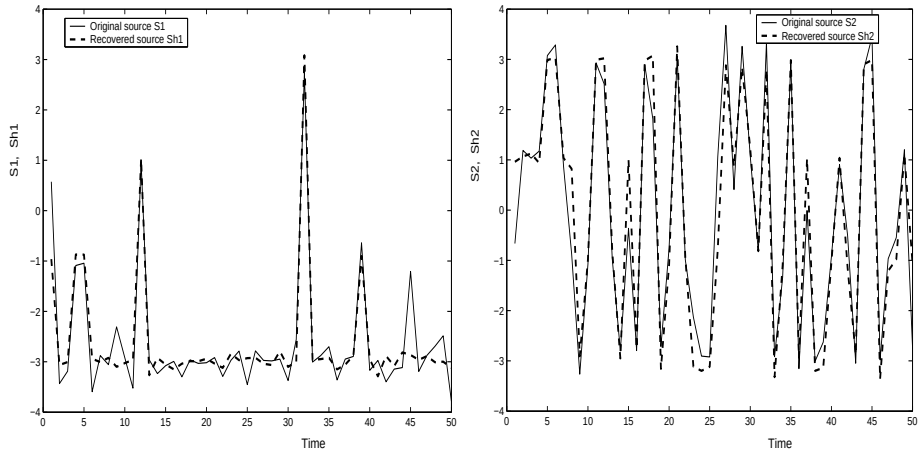
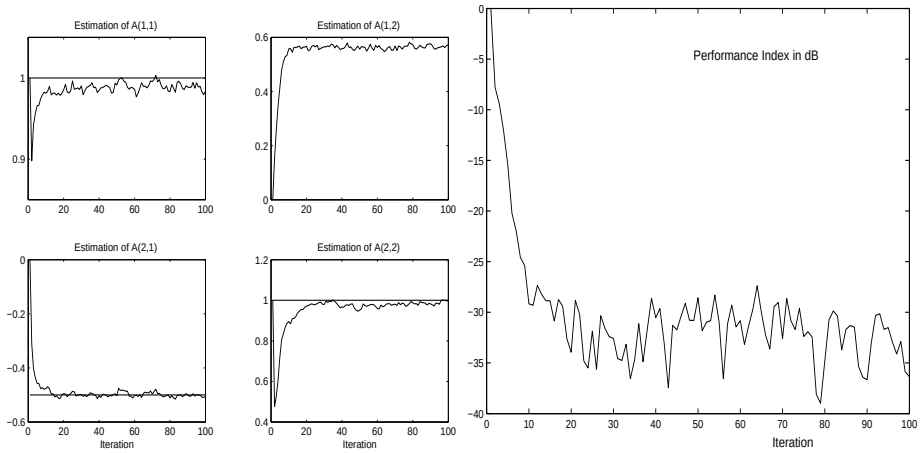


FIG. III.10: Résultats de reconstruction des deux sources avec l'algorithme *Fast-Viterbi-EM*.



a b

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

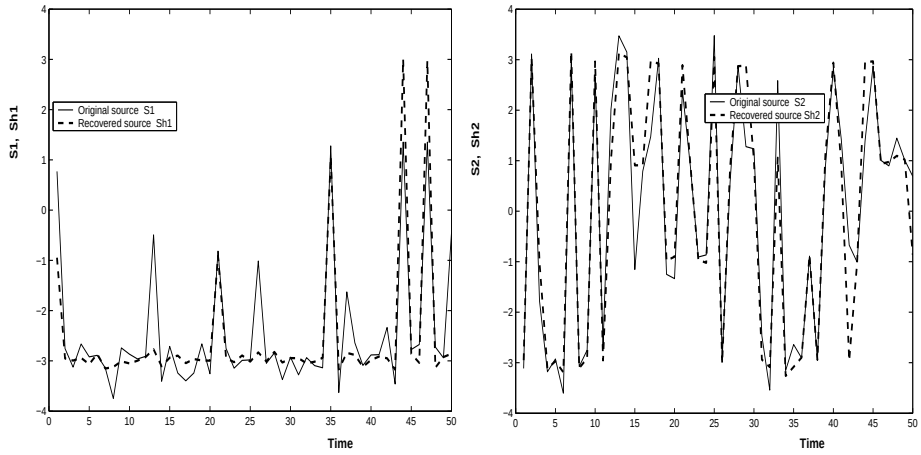


FIG. III.12: Résultats de reconstruction des deux sources avec l'algorithme *Fast-Gibbs-EM*.

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 $\mathcal{S}$:

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

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 :

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

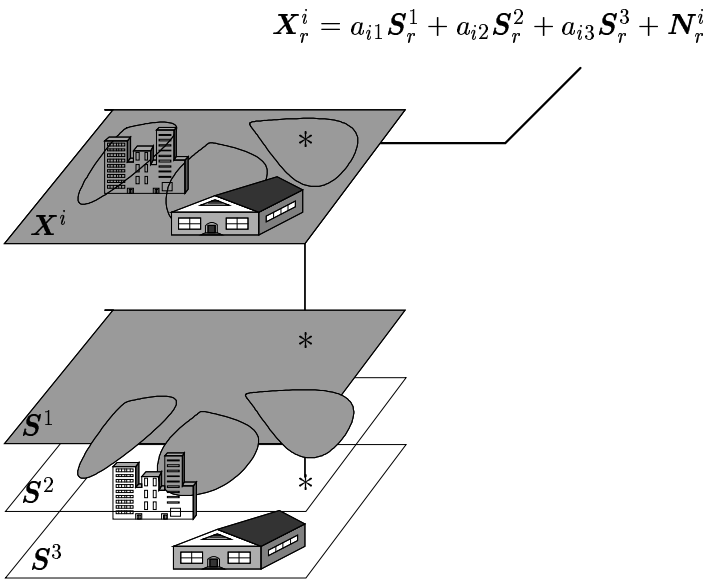


FIG. IV.1: Mélange de sources : l'image observée sur le capteur i est une combinaison linéaire bruitée des images sources. Les cœfficients de la combinaison forment la $i^{\text{ème}}$ ligne de la matrice de mélange $\boldsymbol{A}$.

[A] MODÉLISATION DES SOURCES ET DU BRUIT

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

$$\mathrm{E}\left[\boldsymbol{n}(r)\,\boldsymbol{n}(s)^*\right] = \delta_{r=s}\boldsymbol{R}_\epsilon$$

où * désigne la transposée d'un vecteur.

[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))

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 :

- le mélange de gaussienne 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 ([**??**], 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 temporelle,
- mettre l'accent sur la structure cachée de ce modèle qui fait apparaître une étape de classification. En effet, la modélisation des étiquettes par une chaîne de Markov (dans le cas monodimensionnel) 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 [**?**].

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.*

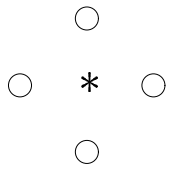
Exemple 4 *On suppose que $\mathcal{S}$ est sous-graphe de $\mathbb{Z} \times \mathbb{Z}$,*

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

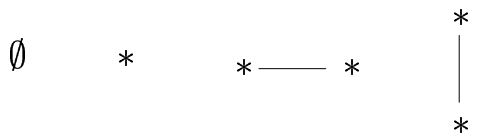
et le système de voisinage est défini par :

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

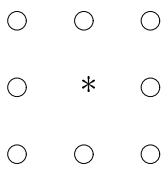
où c est une constante qui mesure l'étendu du voisinage. Pour $c = 1$, chaque site $*$ possède 4 voisins $\circ$ (voisinage d'ordre 1) :



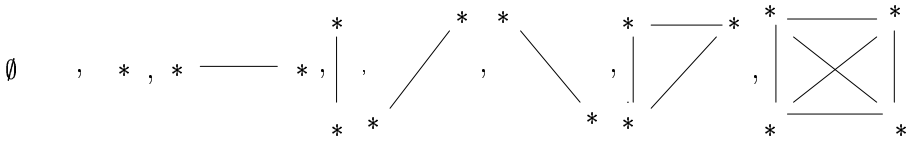
Les cliques correspondants sont :



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 [**?**], 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 paramètre 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 $\boldsymbol{\eta}^j = (\mu_{jk}, \sigma_{jk}^2)_{k=1..K}$ 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 ségmenter 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 et
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 ségmentation 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 ségmentation est une séparation le long de la dimension spatiale. Un traitement optimal ne consiste pas à effectuer une séparation suivie d'une ségmentation 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.

[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 ségmentation est une généralisation à deux sens :

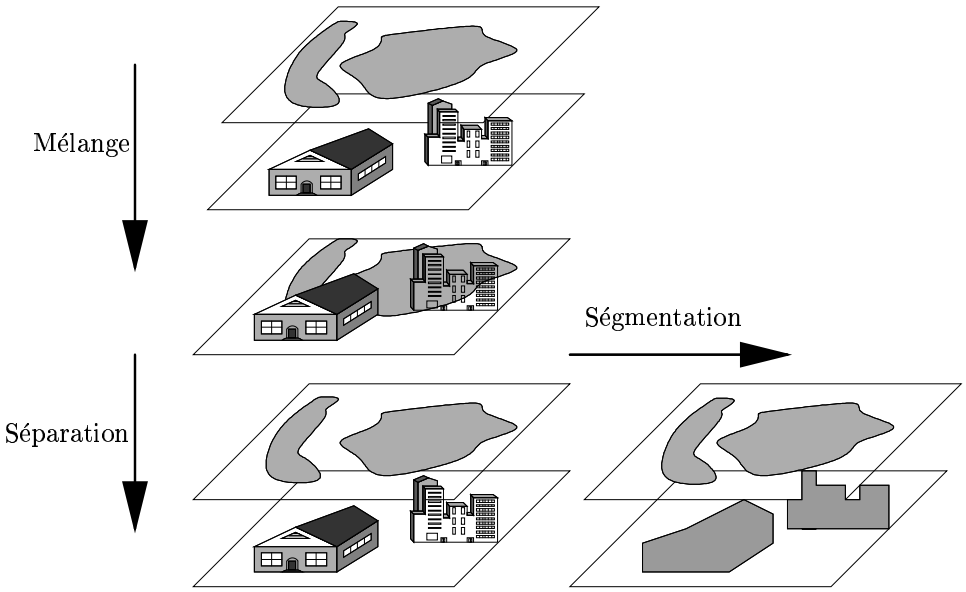


FIG. IV.2: On distingue deux types de séparation : (**i**) une séparation transversale le long des capteurs, (**ii**) une séparation spatiale le long des pixels

1. vis-à-vis du problème de séparation de sources, la ségmentation 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 ségmentation, ce travail est une extension de la ségmentation non supervisée au cas plus difficile où les images à ségmenter 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 [**?**]. On donne les expressions des δ-*a priori* ainsi que les expressions des *a posteriori* résultants 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 accèlèrent 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

[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.

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 :

$$p(\boldsymbol{\theta} \mid \boldsymbol{X}) \propto p(\boldsymbol{X} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

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

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

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, $\boldsymbol{\mu}_{\boldsymbol{z}_r} = [\mu_{1z_1}, ..., \mu_{nz_n}]^t$ 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, le logarithme de la vraisemblance complétée $\log \mathrm{p}(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\theta})$ est, à une constante additive près, une somme pondérée de divergences de Kullback-Leibler entre les covariances empiriques et les covariances théoriques de chaque région :

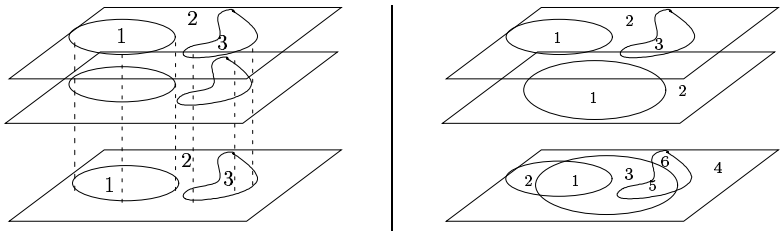
$$\frac{\log \mathrm{p}(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\theta})}{|\mathcal{S}|} = \sum_{k=1}^{K} \alpha_k D_{KL}(\boldsymbol{R}_k, \hat{\boldsymbol{R}}_k) \; + \; cte$$

où $\alpha_k = \frac{|\mathcal{S}_k|}{|\mathcal{S}|}$ est la proportion de la région $\mathcal{S}_k$ appartenant à la classe k. $\hat{\boldsymbol{R}}_k = \sum_{\mathcal{S}_k} \boldsymbol{x}_t \boldsymbol{x}_t^* / |\mathcal{S}_k|$ est la covariance empirique et $\boldsymbol{R}_k = \boldsymbol{A}\boldsymbol{R}_k\boldsymbol{A}^* + \boldsymbol{R}_\epsilon$ 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 alors égale au nombre des étiquettes du champ $\boldsymbol{Z}$ commun à toutes les sources (voir figure (**??**)). 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 [**?**], sauf qu'on suppose 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. Le profil des moyennes et variances peut être le même pour toutes les sources sans altérer les perfomances de séparation.



(a) (b)

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 : $K = K_1 \times K_2 = 6$

[aussi l'identifiabilité (passage de puissance), dégénérescence..]

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 :

$$\begin{cases} \boldsymbol{\eta}^j = \left(\boldsymbol{\eta}_k^j\right)_{k=1..K_j} \\ \\ \boldsymbol{\eta}_k^j = (\mu_k^j, v_k^j = (\sigma_k^j)^2) \end{cases}$$

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 [**?**]. La construction de ce critère est inspirée de la théorie de l'information et fait l'objet du chapitre (**??**). 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 **?** :

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

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 approximer l'expression (IV.5) en travaillant directement sur les vraisemblances complétées $p(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Z} \mid \boldsymbol{\theta})$.

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 multivariabe. 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 :

$$|g(\boldsymbol{A}, \boldsymbol{R}_\epsilon)|^{1/2}\, d\,\boldsymbol{A}\, d\boldsymbol{R}_\epsilon = \frac{|\underset{\boldsymbol{\eta}}{E} \boldsymbol{R}_{ss}|^{m/2}}{|\boldsymbol{R}_\epsilon|^{\frac{m+n+1}{2}}} d\,\boldsymbol{A}\, d\boldsymbol{R}_\epsilon \qquad \text{(IV.6)}$$

[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 [**?**] 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 [**?**]. On obtient ainsi un *a priori* **normal gamma inverse** pour $\boldsymbol{\eta}$:

$$\Pi_0(\boldsymbol{\eta}) \;=\; \prod_{k=1}^K \Pi_0(\boldsymbol{\eta}_k) \;=\; \prod_{k=1}^K \mathcal{N}\left(\mu_k \,;\, \mu^0, \tfrac{v_k}{\nu^0}\right) \mathcal{G}\left(v_k^{-1} \,;\, \tfrac{\nu^0}{2}, \tfrac{\nu^0}{2} v^0\right) \qquad \text{(IV.7)}$$

avec $\nu^0 = \alpha\, w_k^0$, $\alpha = \frac{\gamma_e}{\gamma_u}$, w_k^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 :

$$\begin{aligned} \underset{s|\eta^0}{E} \underset{|s}{D_0}(\boldsymbol{A}, \boldsymbol{R}_\epsilon : \boldsymbol{A}^0, \boldsymbol{R}_\epsilon{}^0) = \tfrac{1}{2} \Big(& \log \left| \boldsymbol{R}_\epsilon \boldsymbol{R}_\epsilon^{0\,-1} \right| + \mathrm{Tr}\left(\boldsymbol{R}_\epsilon^{-1} \boldsymbol{R}_\epsilon{}^0\right) \\ & + \mathrm{Tr}\left(\boldsymbol{R}_\epsilon^{-1} (\boldsymbol{A} - \boldsymbol{A}^0) \underset{s|\eta^0}{E}[\boldsymbol{R}_{ss}] (\boldsymbol{A} - \boldsymbol{A}^0)^* \right) \Big) \end{aligned} \qquad \text{(IV.8)}$$

En combinant (IV.8) avec (IV.6), on obtient l'expression de la distribution 0-*a priori* de $(\boldsymbol{A}, \boldsymbol{R}_\epsilon)$:

$$\Pi_0(\boldsymbol{A}, \boldsymbol{R}_\epsilon^{-1}) = \mathcal{N}\left(\boldsymbol{A}\,;\, \boldsymbol{A}^0, \tfrac{1}{\alpha}{\boldsymbol{R}_{ss}^0}^{-1} \otimes \boldsymbol{R}_\epsilon\right) \mathcal{W}_{im}\left(\boldsymbol{R}_\epsilon^{-1}\,;\, \alpha, {\boldsymbol{R}_\epsilon^0}^{-1}\right) |\underset{s|\eta}{E}[\boldsymbol{R}_{ss}]|^{\frac{m}{2}} \tag{IV.9}$$

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 EMn'est pas implémentable[3]. On s'oriente alors vers les techniques d'échantillonnage en considérant deux types d'algorithmes : des algorithmes de type EMet 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 :

$$\tilde{\mathcal{Q}}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}^{(k)}) = \frac{1}{M}\mathrm{E}_{\boldsymbol{S}}\left[\log\, \mathrm{p}(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Z}^{(m)} \mid \boldsymbol{\theta})\right] + \log \mathrm{p}(\boldsymbol{\theta}) \tag{IV.10}$$

   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 $\boldsymbol{\theta}$:

$$\tilde{\boldsymbol{\theta}}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} \tilde{\mathcal{Q}}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}^{(k)})$$

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 EMexact.

[3]plus précisemment c'est l'intégration par rapport au champ d'étiquettes $\boldsymbol{Z}$ qui n'est pas possible

2. $M < \infty$: on obtient un algorithme de type RB-EM, étudié dans le chapitre précédent (III). Seulement des résultats asymptotiques (lorsque le nombre d'échantillons tend vers l'infini) peuvent être dérivés dans ce cas. Ces propriétés guarantissent 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 inifie 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 est alors effectué 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 entre 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

repeat until convergence,

1.draw$(\tilde{\boldsymbol{Z}}^{(h)}, \tilde{\boldsymbol{S}}^{(h)}) \sim p(\boldsymbol{Z}, \boldsymbol{S} \mid \boldsymbol{X}, \tilde{\boldsymbol{\theta}}^{(h-1)})$

2.draw$\tilde{\boldsymbol{\theta}}^{(h)} \sim p(\boldsymbol{\theta} \mid \boldsymbol{X}, \tilde{\boldsymbol{Z}}^{(h)}, \tilde{\boldsymbol{S}}^{(h)})$

(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 $p(\boldsymbol{\theta} \mid \boldsymbol{X})$. D'après le théorème (**??**) du chapitre (II), les sommes empiriques $\sum_{h=1}^{H} f(\tilde{\boldsymbol{\theta}}^{(h)} \Big/ H$ tendent vers les espérances *a posteriori* $\mathrm{E}\big[f(\boldsymbol{\theta}) \mid \boldsymbol{X}\big]$ 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* $p(\boldsymbol{\theta} \mid \boldsymbol{X})$ et on approxime les espérances *a posteriori* par :

$$\mathrm{E}\big[f(\boldsymbol{\theta}) \mid \boldsymbol{X}\big] \approx \frac{1}{H} \sum_{h=1}^{H} f(\tilde{\boldsymbol{\theta}}^{(h_0+h)}) \tag{IV.12}$$

Echantillonnage de $(\boldsymbol{Z}, \boldsymbol{S})$: D'après la règle séquentielle de Bayes,

$$p(\boldsymbol{Z}, \boldsymbol{S} \mid \boldsymbol{X}, \boldsymbol{\theta}) = p(\boldsymbol{S} \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\theta})\, p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})$$

l'échantilonnage exact de la distribution *a posteriori* jointe est obtenu par un échantillonnage la loi marginale $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})$ suivi par un échantillonnage de la distribution conditionnelle $p(\boldsymbol{S} \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\theta})$.

1. On simule $\tilde{\boldsymbol{Z}}$ selon sa distribution *a posteriori* marginale (en intégrant par rapport aux sources),
$$p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}) \propto p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\theta})\, P_M(\boldsymbol{Z}) \tag{IV.13}$$
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, $\boldsymbol{\mu}_{\boldsymbol{z}_r} = [\mu_{1z_1}, ..., \mu_{nz_n}]^t$ 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 [**?**],
$$\begin{aligned} \boldsymbol{V}_r^{apost} &= \left[\boldsymbol{A}^* \boldsymbol{R}_\epsilon^{-1} \boldsymbol{A} + \boldsymbol{R}_{\boldsymbol{z}_r}^{-1}\right]^{-1} \\ \boldsymbol{m}_r^{apost} &= \boldsymbol{V}_r^{apost} \left(\boldsymbol{A}^* \boldsymbol{R}_\epsilon^{-1} \boldsymbol{x}_r + \boldsymbol{R}_{\boldsymbol{z}_r}^{-1} \boldsymbol{\mu}_{\boldsymbol{z}_r}\right) \end{aligned} \tag{IV.14}$$

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 :
- wishart inverse pour la covariance du bruit et gamma inverse 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 86. On note que la forme wishart inverse des lois des matrices de covariance élimine le risque de dégénérescence mentionné dans le chapitre précédent (III) et repris en détail dans le chapitre (**??**). On donne, ci-après, les expressions correspondantes aux paramètres $(\boldsymbol{A}, \boldsymbol{R}_\epsilon)$ dans le cas particulier $\alpha = 0$ (*a priori* de Jeffreys) :

$$\begin{cases} \boldsymbol{R}_\epsilon^{-1} \sim Wi_m(\nu_p, \boldsymbol{\Sigma}_P),\ \nu_p = \frac{|\mathcal{S}|-n}{2},\ \boldsymbol{\Sigma}_p = \frac{|\mathcal{S}|}{2}(\boldsymbol{R}_{xx} - \boldsymbol{R}_{xs}\boldsymbol{R}_{ss}^{-1}\boldsymbol{R}_{xs}^*) \\ p(\boldsymbol{A} \mid \boldsymbol{R}_\epsilon) \sim \mathcal{N}(\boldsymbol{A}_p, \boldsymbol{\Gamma}_p), \boldsymbol{A}_p = \boldsymbol{R}_{xs}\boldsymbol{R}_{ss}^{-1},\ \boldsymbol{\Gamma}_p = \frac{1}{|\mathcal{S}|}\boldsymbol{R}_{ss}^{-1} \otimes \boldsymbol{R}_\epsilon \end{cases} \tag{IV.15}$$

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 8 *La distribution a posteriori de* $\boldsymbol{Z}$ $\mathrm{p}(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})$ *est un champ de Gibbs avec le même voisinage* ∂ *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 :

$$\begin{cases} \text{pour tout } r \in \mathcal{S}, \\ Z_r \sim \mathrm{p}(Z_r \mid \boldsymbol{Z}_{\mathcal{S}\setminus r}\boldsymbol{X}, \tilde{\boldsymbol{\theta}}^{(k-1)}) \end{cases} \tag{IV.16}$$

On va essayer de résumer 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 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 approximant 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 repartionner 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 une 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 [?].*

Remarque 9 *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 [?]. 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* $\boldsymbol{\theta}^{(h)} \sim \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{S}^{(h)}, \boldsymbol{Z}^{(h)})$

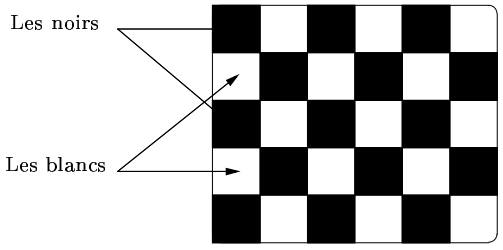


FIG. IV.4: Implémentation parallèle en échiquier

IV.3.3 CONTRÔLE DE CONVERGENCE

Le contrôle de convergence d'une chaîne de Markov est une question délicate [**?**]. Beaucoup d'outils de contrôle ont été développés dans la litérature des méthodes MCMC . Cependant, aucune méthode n'est préconisée [**?**]. 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 traîté. 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'oeil si la série tend vers un comportement stationnaire. On constate qu'avec cette méthode on ne peut 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 approximée 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 [**?**].

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 [**?**] que l'estimateur S_H^{rb} domine 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,

$$\begin{aligned} S_H^{rb}(r) &= \tfrac{1}{H} \textstyle\sum_{h=1}^{H} \mathrm{E}\left[\boldsymbol{s}(r) \mid \boldsymbol{X}, \boldsymbol{Z}^{(h)}, \tilde{\boldsymbol{\theta}}^{(h)}\right] \\ &= \tfrac{1}{H} \textstyle\sum_{h=1}^{H} \boldsymbol{m}_r^{apost} \\ &= \tfrac{1}{H} \textstyle\sum_{h=1}^{H} \left[\boldsymbol{A}^* \boldsymbol{R}_\epsilon^{-1} \boldsymbol{A} + \boldsymbol{R}_{\boldsymbol{z}_r}^{-1}\right]^{-1} \left(\boldsymbol{A}^* \boldsymbol{R}_\epsilon^{-1} \boldsymbol{x}_r + \boldsymbol{R}_{\boldsymbol{z}_r}^{-1} \boldsymbol{\mu}_{\boldsymbol{z}_r}\right) \end{aligned}$$

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'aprio Π_0 (voir l'expression (IV.9)), sa distribution *a posteriori* est gaussienne ((IV.18) de l'annexe 1 page (86)). 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,

$$\begin{aligned} S_H^{rb} &= \tfrac{1}{H} \textstyle\sum_{h=1}^{H} \mathrm{E}\left[\boldsymbol{A} \mid \boldsymbol{X}, \boldsymbol{Z}^{(h)}, \boldsymbol{S}^{(h)}\right] \\ &= \tfrac{1}{H} \textstyle\sum_{h=1}^{H} \boldsymbol{R}_{xs} \boldsymbol{R}_{ss}^{-1} \end{aligned}$$

Remarque 10 *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 :

$$\hat{S}_H = \sum_{h=h_0+1}^{H} \left[T(\tilde{\boldsymbol{\theta}}^{(h)}) - \hat{\mu} \right], \quad \hat{\mu} = (H - h_0)^{-1} \sum_{h=h_0+1}^{H} T(\tilde{\boldsymbol{\theta}}^{(h)})$$

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 [?] consiste à tracer la série $\{\hat{S}_H\}$ en fonction de H et de connecter les points successifs par des ségments. 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 de 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 données 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 ségmentation sur la dernière ligne.

On note la qualité de la séparation des sources. Les résultats de la ségmentation sont presque les mêmes que si on ségmente 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.

IV.5.1 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).

IV.5.2 Séparation et ségmentation 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 ségmentation 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 ségmentation 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 ségmentation au cas plus difficile où les images à classifier ne sont pas directement accessibles et ont subi un mélange linéaire bruité.

La formulation bayésienne offre un cadre naturel à la séparation et la ségmentation 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.

IV.5.3 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 ségmentation simulanées n'est pas plus compliqué que celui d'une séparation avec une classification connue ou une ségmentation directe des sources. A titre d'exemple, l'échantillonneur de Gibbs parallèle (*IV*.17) peut être utilisé exclusivement pour la ségmentation 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

Annexe 1 : *a posteriori* distributions

a posteriori DE $(\boldsymbol{A}, \boldsymbol{R}_\epsilon)$

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 wishart inverse** :

$$p(\boldsymbol{A}, \boldsymbol{R}_\epsilon \,|\, \boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Z}) = \mathcal{N}(\boldsymbol{A}\,;\, \boldsymbol{A}_p, \boldsymbol{\Gamma}_p)\mathcal{W}_{im}(\boldsymbol{R}_\epsilon^{-1}\,;\, \nu_p, \boldsymbol{\Sigma}_p) \tag{IV.18}$$

dont les paramètres sont mis à jour selon les équations suivantes :

$$\begin{cases} \nu_p = K + \alpha,\ (K = |\mathcal{S}|) \\ Vec(\boldsymbol{A}_p) = \left[\boldsymbol{R}_v^{-1} + \boldsymbol{R}_a^{-1}\right]^{-1}\left[\boldsymbol{R}_v^{-1}\, Vec(\boldsymbol{A}_v) + \boldsymbol{R}_a^{-1}\, Vec(\boldsymbol{A}_0)\right] \\ \boldsymbol{\Gamma}_p^{-1} = \boldsymbol{R}_v^{-1} + \boldsymbol{R}_a^{-1} \\ \boldsymbol{R}_v = K^{-1}\boldsymbol{R}_{ss}^{-1} \otimes \boldsymbol{R}_\epsilon \\ \boldsymbol{R}_a = \alpha^{-1}{\boldsymbol{R}_{ss}^0}^{-1} \otimes \boldsymbol{R}_\epsilon \\ \boldsymbol{A}_v = \boldsymbol{R}_{xs}\,\boldsymbol{R}_{ss}^{-1} \\ \boldsymbol{\Sigma}_p^{-1} = \frac{1}{K+\alpha}\left[k\hat{\boldsymbol{R}}_\epsilon + \alpha\boldsymbol{R}_0 + (\boldsymbol{A}_0 - \boldsymbol{A}_v)(K^{-1}\boldsymbol{R}_{ss}^{-1} + \alpha^{-1}{\boldsymbol{R}_{ss}^0}^{-1})^{-1}(\boldsymbol{A}_0 - \boldsymbol{A}_v)^T\right] \\ \hat{\boldsymbol{R}}_\epsilon = \boldsymbol{R}_{xx} - \boldsymbol{R}_{xs}\,\boldsymbol{R}_{ss}^{-1}\,\boldsymbol{R}_{sx} \end{cases}$$

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 $\boldsymbol{R}_{ss}$:

$$\boldsymbol{R}_{ss}^0 = \underset{s|\eta^0}{E}[\boldsymbol{R}_{ss}].$$

a posteriori DE $(\mu_k, v = \sigma_k^2)$

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 :

$$\begin{cases} \mu_p = \frac{N_k\bar{s} + \alpha w_k^0\mu_0}{N_k^\dagger \alpha w_k^0} \\ v_p = \frac{v_k}{N_k^\dagger \alpha w_k^0} \\ \eta_p = \frac{N_k^\dagger \alpha w_k^0}{2} \\ \beta_p = \frac{\alpha w_k^0 v_0}{2} + \frac{s^2}{2} + \frac{1}{2}\frac{N_k \alpha w_k^0}{N_k^\dagger \alpha w_k^0}(\bar{s} - \mu_0)^2 \\ \bar{s} = \frac{\sum_{r\in\mathcal{S}_k} s(r)}{N_k} \\ s^2 = \sum_{r\in\mathcal{S}_k} s(r)^2 - N_k\bar{s}^2 \end{cases}$$

où $\mathcal{S}_k$ est la région de l'image j appartenant à la classe k :

$$\begin{cases} \mathcal{S}_k = \{r \in \mathcal{S} \mid Z(r) = k\} \\ N_k = |\mathcal{S}_k| \end{cases}$$

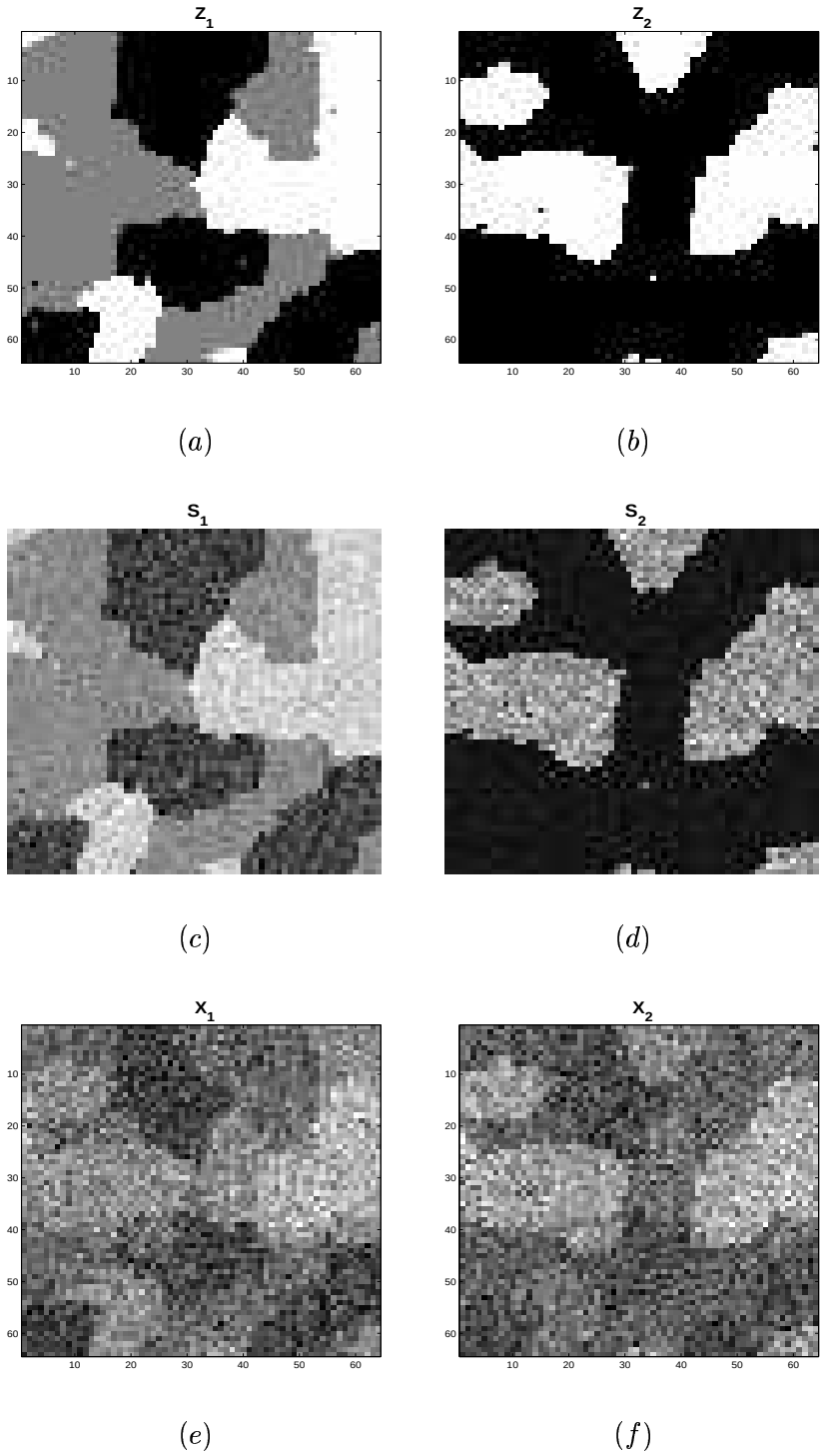


FIG. IV.5: (a) Classification $\boldsymbol{Z}^1$ de la source 1, (b) Classification $\boldsymbol{Z}^2$ de la source 2, (c) Source originale $\boldsymbol{S}^1$, (d) Source originale $\boldsymbol{S}^2$, (e) Image observée $\boldsymbol{X}^1$, (f) Image observée $\boldsymbol{X}^2$

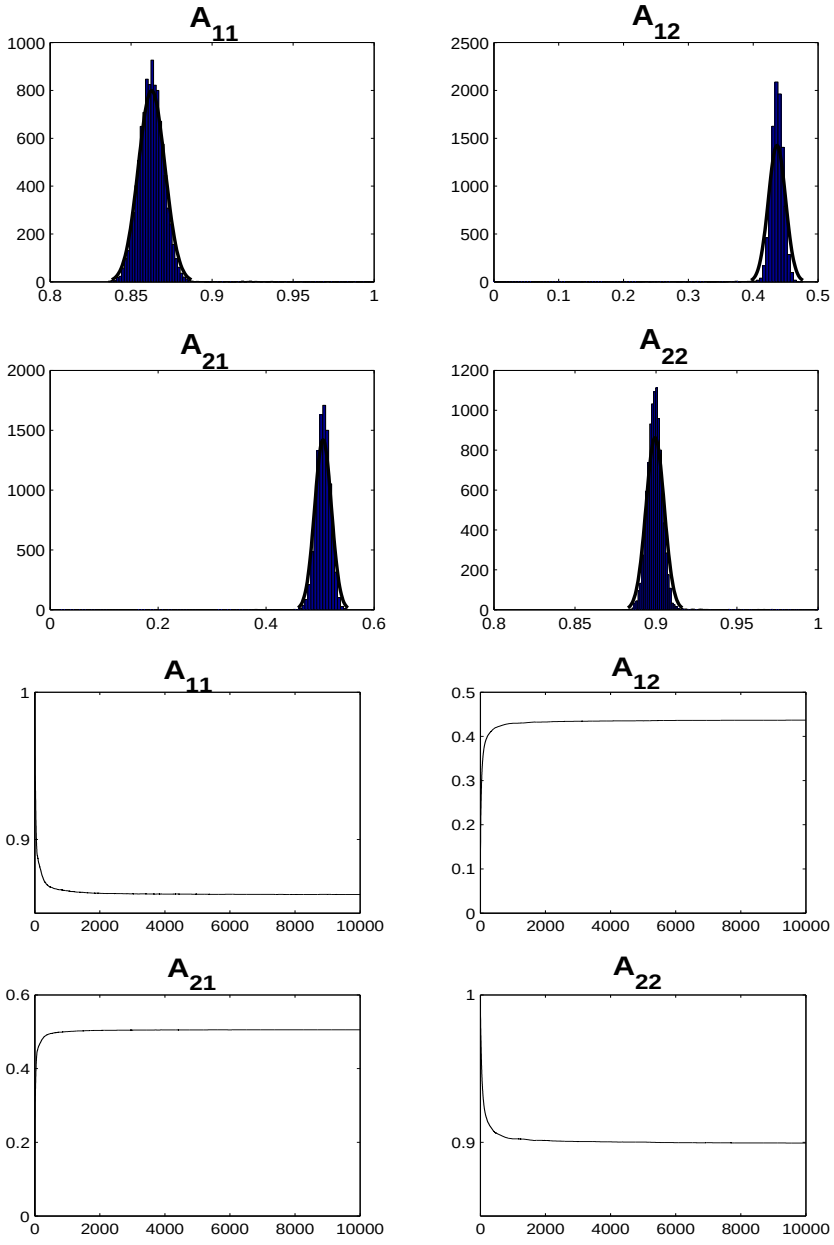


FIG. IV.6: Histogrammes et sommes empiriques des cœfficients de mélange a_{ij}. On note la convergence après 2000 itérations.

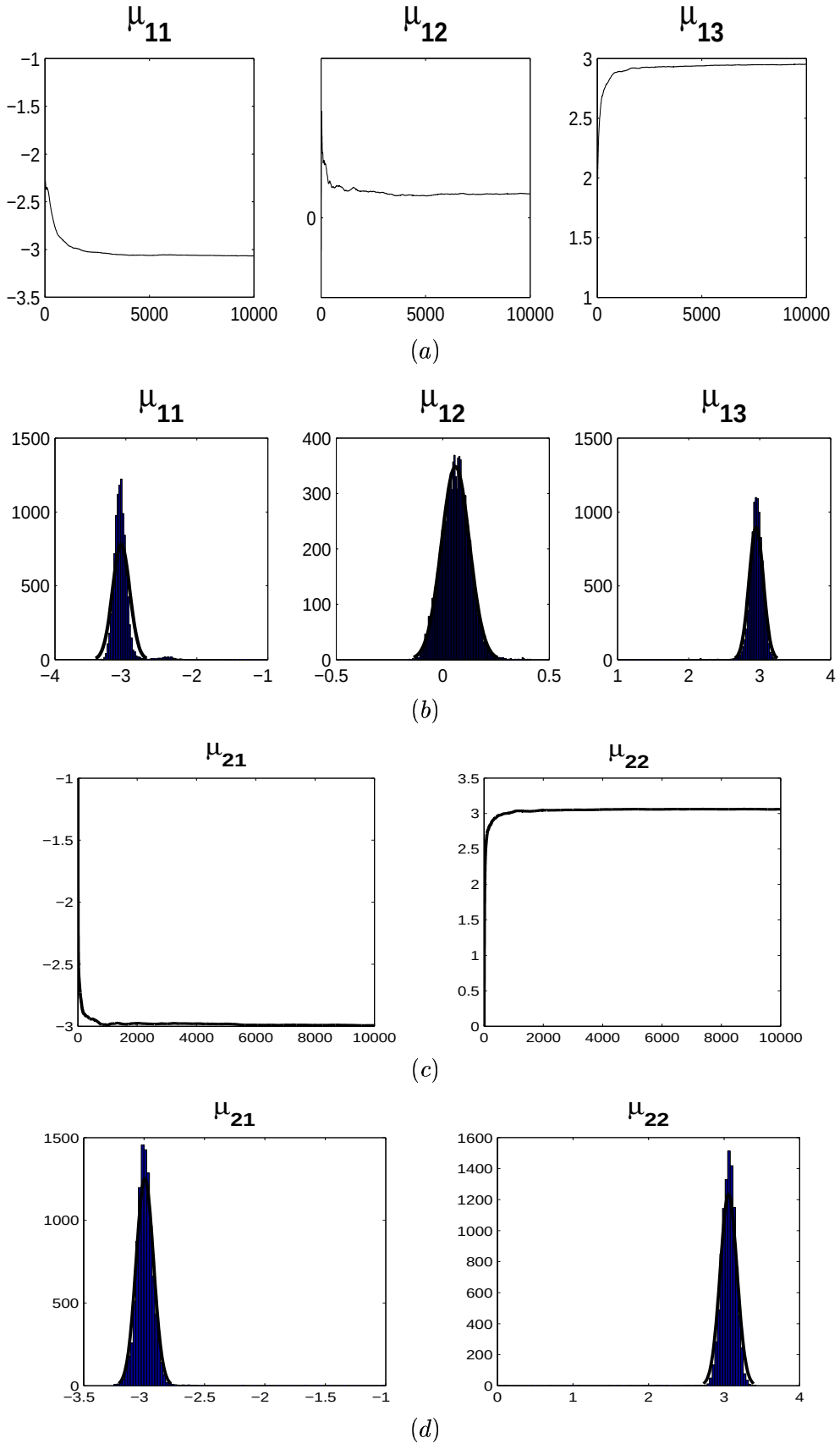
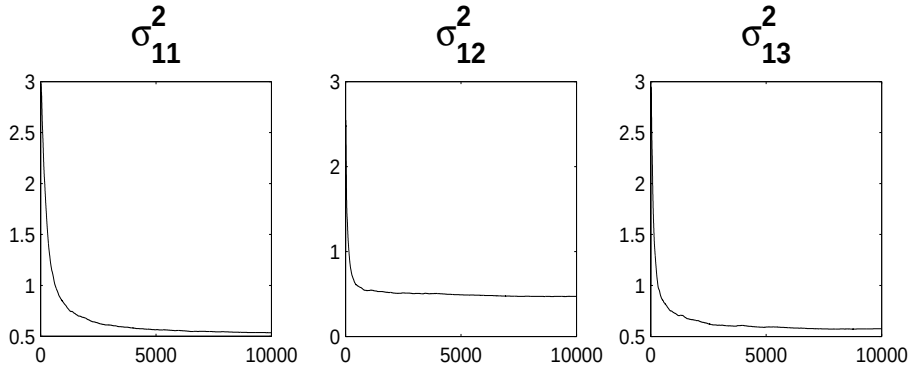
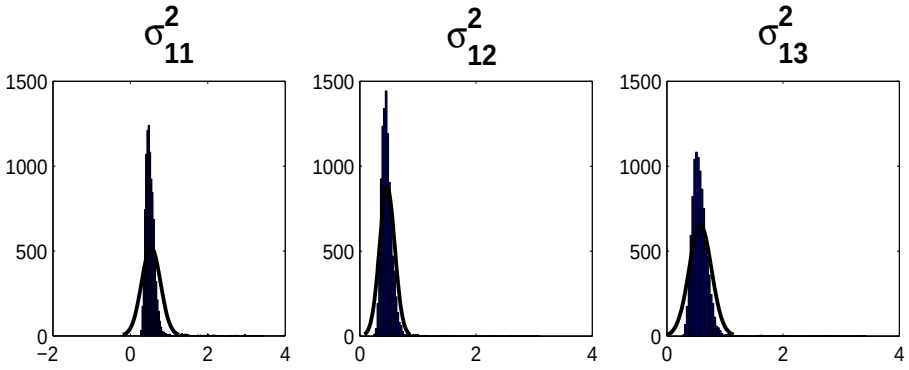


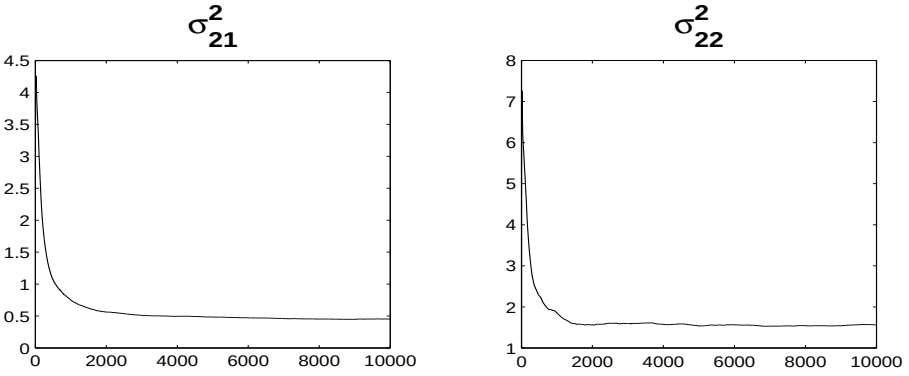
FIG. IV.7: (a)- Convergence des sommes empiriques des moyennes m_{ij} de la source 1 (b)- Histogrammes des moyennes de la source 1 (c)- Convergence des sommes empiriques des moyennes m_{ij} de la source 2 (d)-Histogrammes des moyennes de la source 2



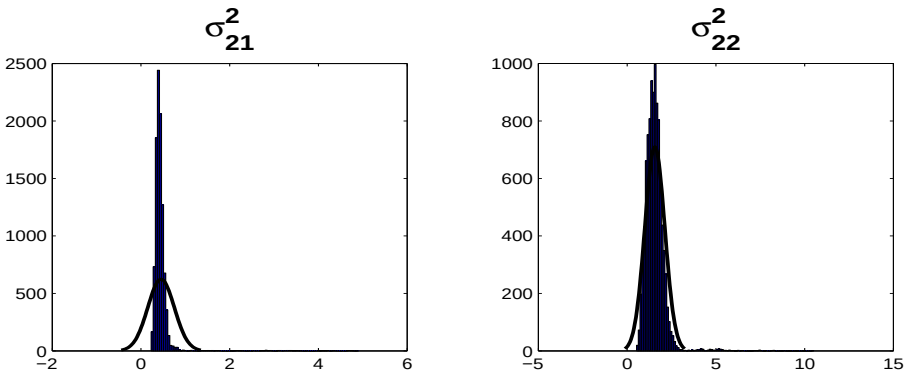
(a)



(b)



(c)



(d)

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

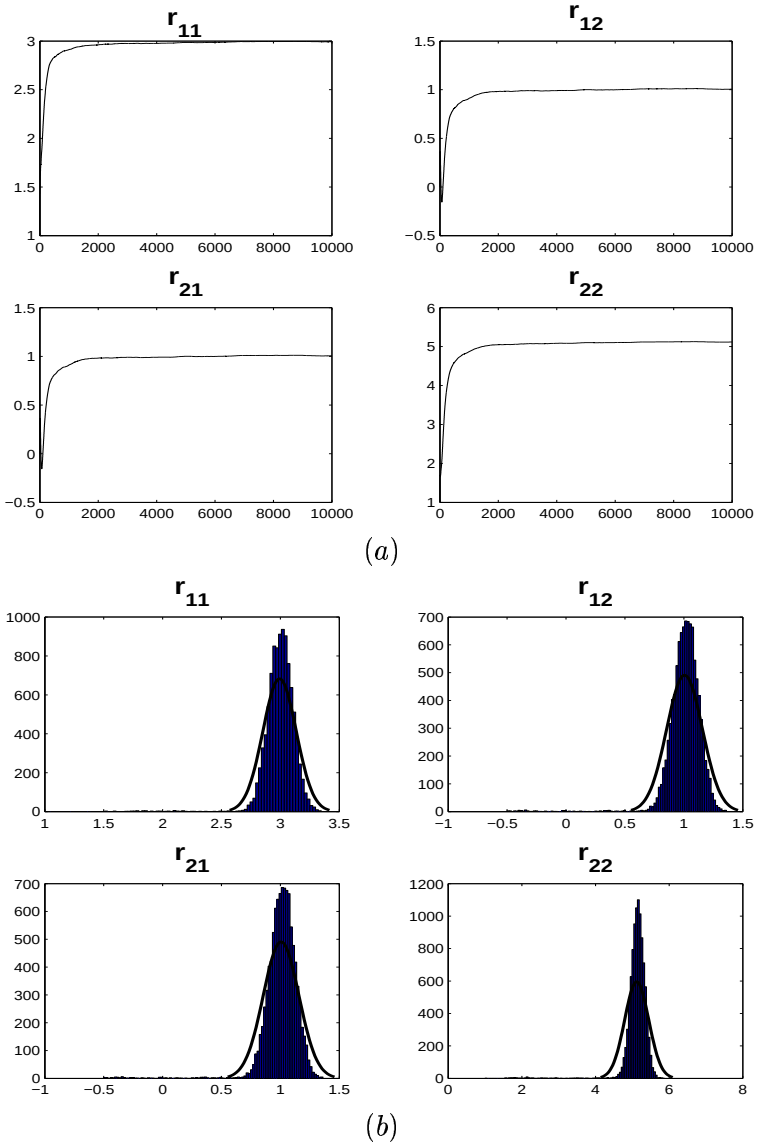


FIG. IV.9: (a)- Convergence de la somme empirique de la chaîne des variances du bruit, (b) histogrammes des variances du bruit

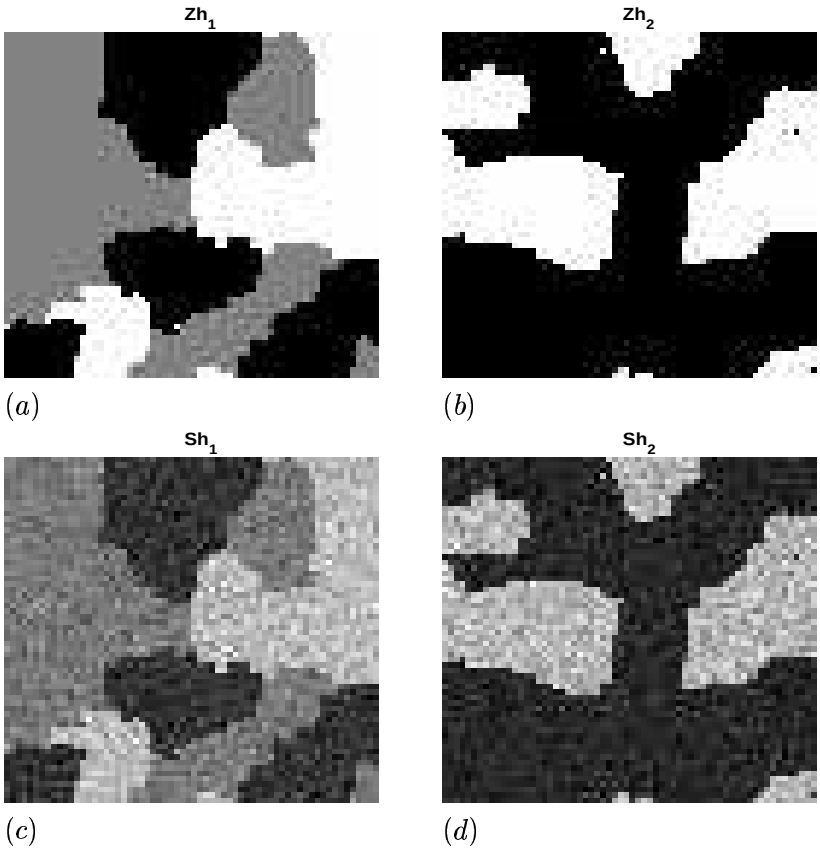


FIG. IV.10: (a)- Estimation de la classification de la source 1, (b)- Estimation de la classification de la source 2, (c)- Reconstruction de la source 1, (c)- Reconstruction de la source 2.

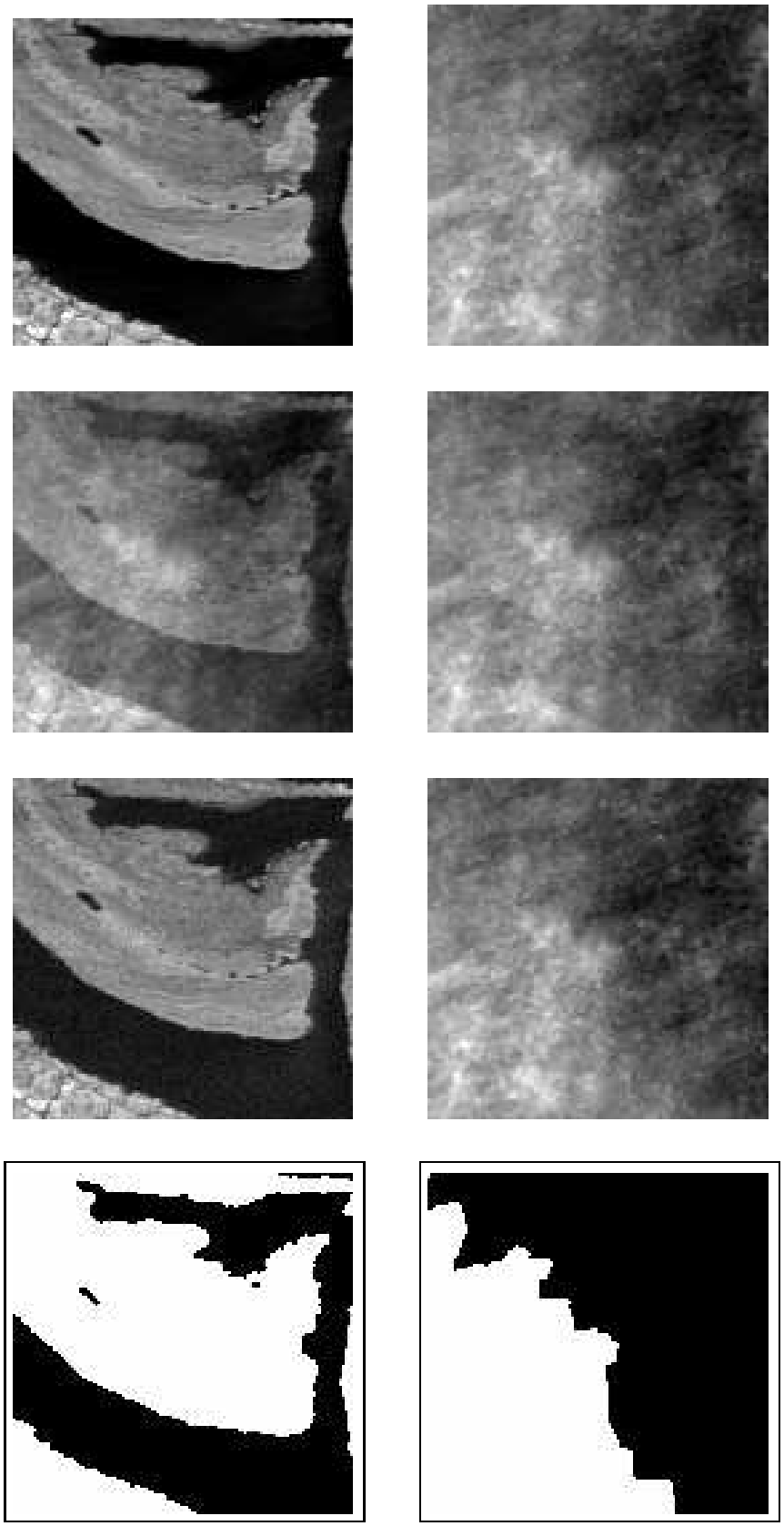


FIG. IV.11: Du haut vers le bas : sources originales, sources mélangées, sources estimées et sources ségmentées.