REVUE DE STATISTIQUE APPLIQUÉE

M. LAVIELLE
E. MOULINES

Quelques exemples de problèmes inverses en statistique et en traitement du signal

Revue de statistique appliquée, tome 45, n° 4 (1997), p. 5-38
<http://www.numdam.org/item?id=RSA_1997__45_4_5_0>

© Société française de statistique, 1997, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

Rev. Statistique Appliquée, 1997, **XLV** (4), 5-38

QUELQUES EXEMPLES DE PROBLÈMES INVERSES EN STATISTIQUE ET EN TRAITEMENT DU SIGNAL

M. Lavielle*, E. Moulines**

* *UFR de Mathématiques et Informatique, Université Paris V and CNRS URA 743, Université Paris-Sud, lavielle@math-info.univ-paris5.fr*

** *Ecole Nationale Supérieure des Télécommunications, CNRS-URA 820, 46, rue Barrault, 75634 Paris CEDEX 13, moulines@sig.enst.fr*

RÉSUMÉ

Un problème inverse consiste à reconstruire une série de données non observées à partir d'une série de données observées. Lorsque le modèle considéré est paramétrique, l'algorithme EM, ou une approximation stochastique de EM (SAEM) peuvent être envisagés pour estimer les paramètres du modèle. On montre alors que l'algorithme SAEM converge, sous des hypothèses très générales, vers un maximum de la vraisemblance des observations. Une version recuit simulé de SAEM permet même d'éviter les maximas locaux. On peut ensuite reconstruire les données manquantes en maximisant la distribution *a posteriori* des données non observées (estimateur MAP). Des applications en déconvolution sont présentées. Il arrive enfin que dans certains problèmes, la reconstruction des données manquantes et l'estimation de certains paramètres soient effectuées simultanément. On définit alors une fonction globale à minimiser. Des applications en clustering et en détection de ruptures sont proposées. Quelques résultats asymptotiques pour l'estimateur des instants de ruptures peuvent être obtenus.

Mots-clés : *Problème inverse, algorithme EM, approximation stochastique, recuit simulé, déconvolution, détection de ruptures.*

ABSTRACT

An inverse problem consists in recovering a set of non-observed data from a set of observed data. When the model is parametric, EM algorithm, or a stochastic approximation version of EM (SAEM), can be proposed to estimate the parameters of the model. We show that SAEM converges towards a maximum of the observed likelihood. A simulated annealing version of SAEM avoids convergence towards a local maximum. The non observed data can be recovered by maximizing the posterior distribution (MAP estimate). Some applications in deconvolution are presented. In some contexts, the recovery of the non observed data and the estimation of the parameters of the model are performed simultaneously by minimizing an objective function. Some applications in clustering and change-points detection are given. Some asymptotical results can be obtained concerning the estimation of change-points.

Keywords : *Inverse problem, EM algorithm, stochastic approximation, simulated annealing, deconvolution, detection of change-points.*

1. Introduction

Dans de très nombreuses situations, en statistique, en traitement du signal, ou en traitement de l'image, on cherche à reconstruire une série de données non observées à partir d'une série de données observées. Ce type de problème est ce que l'on appelle un *problème inverse*.

Nous noterons $\underline{y}$ la série des observations et $\underline{z}$ celle des données à reconstruire. Nous considérons ici que le phénomène physique qui fait intervenir ces deux séries peut se modéliser en faisant intervenir certains paramètres, c'est-à-dire que le modèle mathématique sous-jacent est *paramétrique* :

$$\underline{y} = H(\underline{z}; \theta), \tag{1}$$

le paramètre θ prenant ses valeurs dans un certain ouvert Θ de $\mathbb{R}^p$.

Les problèmes qui se posent sont alors les suivants :

1) Choisir le modèle H	*Modélisation*
2) Estimer θ	*Identification*
3) Estimer $\underline{z}$	*Reconstruction*

Nous n'aborderons pas ici le problème du choix du modèle, c'est-à-dire que seule l'estimation de θ et de $\underline{z}$ nous intéressera. Pour tenter de résoudre ce double problème, nous adopterons une *approche probabiliste* : on considère que $\underline{y}$ et $\underline{z}$ sont la réalisation d'un couple de variables aléatoires (Y, Z) définies sur un certain espace de probabilité et à valeurs dans un espace $\mathcal{Y} \times \mathcal{Z}$. De plus, on suppose qu'un *bruit aléatoire* ε est présent; le modèle probabiliste a alors la forme générale :

$$Y = H(Z, \varepsilon; \theta). \tag{2}$$

Les exemples de problèmes inverses sont nombreux :

- En statistique, les problèmes de mélange de population (*clustering*) ou de détection de points aberrants (*outliers*) ont été particulièrement étudiés [4, 25, 27, 40]. On est confronté là à un problème de *classification* et Z est un processus de «label» : Z_i indique de quelle population provient l'observation Y_i. Les problèmes de classification ont dépassé le cadre de l'analyse de données, pour devenir un problème majeur en traitement de l'image (Z est souvent modélisé par un *champ de Markov* [3, 24]), comme en signal (Z est alors une *chaîne de Markov cachée* [36]).

- En traitement du signal, Z peut modéliser les signaux émis et Y les signaux reçus, ε modélise les erreurs de mesure, c'est alors un bruit additif. Citons, par exemple, les modèles de *convolution* $Y = f * Z + \varepsilon$, où f est un filtre de convolution [16, 18, 29, 30, 39], ou les modèles de *séparation de sources* $Y = AZ + \varepsilon$, où A est une matrice de mélange [9, 20, 26].

- En traitement du signal comme en traitement de l'image, les problèmes de segmentation sont aussi des problèmes inverses d'intérêt tout particulier. Le problème consiste à segmenter un signal en segments homogènes (*stationnaires*) [1, 8, 15, 31,

35], ou une image en zones homogènes [38]; Z modélise un processus de ruptures en signal et un processus de bords en image.

Plusieurs approches peuvent être envisagées pour résoudre ce type de problèmes. Dans un cadre purement bayésien, par exemple, le paramètre θ est considéré comme une variable aléatoire et le couple (θ, Z) est muni d'une distribution *a priori*. On cherche alors à estimer la distribution *a posteriori* de (θ, Z), c'est-à-dire conditionnellement aux observations. Pour estimer cette distribution *a posteriori*, on peut avoir recours à un algorithme de *Monte-Carlo par chaînes de Markov* (MCMC) de type Hastings-Metropolis ou échantilloneur de Gibbs [21, 23, 24, 37]. Signalons toutefois que les méthodes MCMC présentent souvent l'inconvénient, dans le cadre de problèmes de grandes dimensions, de ne pas être accompagnées de procédures satisfaisantes de contrôle de leur convergence vers la loi stationnaire. Nous ne pouvons bien sûr pas présenter de façon exhaustive toutes les méthodes de résolution de problème inverse. L'approche bayésienne n'est donc pas traitée ici, bien qu'elle puisse apporter des informations très utiles dans beaucoup d'applications.

Nous avons choisi de nous placer ici dans un contexte où les paramètres du modèle et les données manquantes sont estimés «ponctuellement», comme fonction des observations. Un statisticien aimerait bien sûr que ces estimateurs possèdent de bonnes propriétés statistiques. C'est pourquoi, par exemple, les paramètres du modèle sont estimés au sens du maximum de vraisemblance. Le problème que nous nous posons alors ici n'est pas l'étude des propriétés de cet estimateur, mais plutôt le moyen de le calculer.

Il arrive bien souvent que certains algorithmes utilisés dans la pratique donnent de très bons résultats, mais soient difficiles à justifier par la théorie. D'un autre côté, certains estimateurs présentant de bonnes propriétés théoriques, s'avèrent être inutilisables en pratique, la taille des observations étant trop petite, par exemple, pour pouvoir considérer que l'asymptotique est atteinte. La démarche adoptée ici consiste à essayer de mettre en œuvre des méthodes qui donnent de bons résultats sur des simulations comme sur des données réelles. Une fois ces méthodes jugées performantes, il peut alors être intéressant de les justifier grâce à des résultats théoriques.

Nous considérerons essentiellement trois types de situations dans cet article. La partie 2 est consacrée à l'estimation des paramètres par maximum de vraisemblance. Lorsque l'algorithme EM proposé par Demster *et al.* [14] ne peut être mis en œuvre, des versions stochastiques de EM s'avèrent très performantes, [11, 12, 32, 41]. Malheureusement, peu de résultats théoriques concernant ces algorithmes ont été établis à ce jour. Nous proposons ici une nouvelle approximation stochastique de EM pour lesquels des résultats précis de convergence sont obtenus. L'asymptotique ne porte pas ici sur la taille des observations, mais sur le nombre d'itérations. En utilisant des résultats sur les *algorithmes stochastiques* [17], nous montrons sous des hypothèses très générales la convergence presque-sûre de l'algorithme SAEM vers un maximum local de la vraisemblance. Une version *recuit simulé* de cet algorithme permet la convergence vers le maximum global.

Lorsque les paramètres du modèle sont fixés (soit parce qu'ils sont connus, soit parce qu'ils ont été estimés), nous nous intéressons dans la partie 3 à la reconstruction des données manquantes. L'estimateur MAP (Maximum a Posteriori) maximise

la distribution *a posteriori* des données manquantes. C'est-à-dire en minimisant une fonction qui tient compte à la fois de l'ajustement aux observations et d'une information *a priori* que l'on veut introduire. Dans ce type de situation tout-à-fait classique, on se trouve face à un problème essentiel : comment déterminer les poids relatifs donnés à ces différents critères? Aucune procédure automatique ne s'avère réellement satisfaisante et ces méthodes montrent là leurs limites : il faut souvent se contenter de «régler» ces poids afin que la solution obtenue corresponde au type de solution cherchée (remarquons à cette occasion que les hyper-paramètres du modèle sont aussi fixés arbitrairement dans les méthodes bayésiennes). Nous étudions plus particulièrement les modèles de déconvolution à titre d'exemple.

Finalement, la Section 4 traite des modèles où les paramètres et les données manquantes sont estimés simultanément. Là encore, les estimateurs sont obtenus en minimisant une fonction qui tient compte de différents critères et ne possèdent pas toujours de propriétés statistiques particulières. Dans un problème de *clustering*, par exemple, il n'y a aucune raison pour que la proportion d'observations bien classées augmente avec le nombre d'observations. Par contre, dans un cadre de détection de ruptures dans la moyenne d'un processus aléatoire, nous montrons que l'estimateur des moindres carrés pénalisés converge vers la véritable configuration de ruptures lorsque le nombre de ruptures est inconnu.

2. Identification du modèle

Nous cherchons dans un premier temps à estimer le paramètre $\theta \in \Theta$. Notons $\underline{y} = (y_1, \ldots, y_n)$ les valeurs observées (les y_i étant éventuellement vectoriels) et $\hat{\theta}^{MV}$ l'*estimateur du maximum de vraisemblance* de θ. Notons $g(\underline{y}; \theta)$ la vraisemblance des observations au point θ; $\hat{\theta}^{MV}$ est la valeur de θ la plus vraisemblable au vu des observations, c'est-à-dire :

$$\hat{\theta}^{MV} \stackrel{\text{def}}{=} \operatorname{Arg}\max_{\theta \in \Theta} g(\underline{y}; \theta). \tag{3}$$

On notera $f(\underline{y}, \underline{z}; \theta)$ la vraisemblance complète et $k(\underline{z}|\underline{y}; \theta)$ la vraisemblance conditionnelle de $\underline{z}$. Dans tous les problèmes que nous étudions, la maximisation de $g(\underline{y}; \theta)$ est compliquée, contrairement à celle de $f(\underline{y}, \underline{z}; \theta)$.

L'algorithme EM, [14], maximise $g(\underline{y}; \theta)$ en maximisant itérativement les espérances conditionnelles de $\log f(\underline{y}, Z; \theta)$. Chaque itération de l'algorithme EM est décomposée en deux étapes : à l'itération $k+1$, l'étape E consiste à évaluer l'espérance de $\log f(\underline{y}, Z; \theta)$ conditionnellement à $Y = \underline{y}$ et avec la valeur courante θ_k :

$$\begin{aligned} Q(\theta_k\ ,\ \theta) &\stackrel{\text{def}}{=} E\left\{\log f(\underline{y}, Z; \theta) | \underline{y}; \theta_k\right\} \\ &= E^{\underline{y}}_{\theta_k} \log f(\underline{y}, Z; \theta). \end{aligned} \tag{4}$$

L'étape M consiste à calculer $\theta_{k+1} \in \Theta$ qui maximise $Q(\theta_k\ ,\ \theta)$.

Sous certaines hypothèses, on montre que la suite $\{g(\underline{y};\theta_k)\}$ est croissante et que la suite $\{\theta_k\}$ converge vers un point $\theta^\star$ qui est un point stationnaire de g, c'est-à-dire tel que $\mathcal{D}_\theta g(\underline{y};\theta^\star) = 0$, où $\mathcal{D}_\theta g$ est le gradient de g. Ce point peut être un maximum local, un point-selle, ou même un minimum, suivant le point initial choisi [42].

2.1. L'algorithme SAEM

Très souvent, l'étape E ne peut pas être mise en œuvre car la loi *a posteriori* de Z est compliquée (déconvolution, séparation de sources, ...) et une version stochastique doit être utilisée. Ce type de procédure donne souvent de bons résultats dans la pratique. Ainsi, Wei et Tanner [41] proposent un Monte-Carlo EM (MCEM) dans lequel le calcul de l'espérance conditionnelle est approché par un Monte-Carlo. Celeux et Diebolt ont proposé l'algorithme SEM dans un cadre de mélanges, [11,12]. Lavielle a également utilisé SEM pour estimer de façon paramétrique et non-paramétrique la densité de données non observées, [32]. Younes a proposé un algorithme stochastique d'estimation pour des champs de Gibbs [43]. Les résultats théoriques concernant ces algorithmes sont malheureusement assez limités, ne s'appliquant que sous des conditions difficiles à vérifier, [5, 12].

Dans [33], nous proposons de substituer à l'étape E de l'algorithme EM une approximation stochastique et une étape de simulation de Monte-Carlo. Comme nous allons le voir, cela permet d'obtenir des résultats précis de convergence sous des hypothèses très générales.

Nous supposerons dans la suite que la vraisemblance complète est de type exponentielle courbe. Un peu plus précisément, nous ferons l'hypothèse suivante :

(H1) L'ensemble Θ est un ouvert de $\mathbb{R}^p$. Les données complètes (Y, Z) ont une densité de la forme :

$$\forall \theta \in \Theta \quad f(\underline{y},\underline{z};\theta) = \exp\left\{-\psi(\theta) + \langle \tilde{S}(\underline{y},\underline{z}), \phi(\theta))\rangle + r(\underline{y},\underline{z})\right\} \qquad (5)$$

où $\tilde{S}$ est une fonction mesurable à valeurs dans un convexe fermé $\mathcal{S} \subset \mathbb{R}^m$, $\psi : \Theta \to \mathbb{R}$ et $\phi : \Theta \to \mathbb{R}^m$, sont des fonctions deux fois continuement différentiables en θ.

Remarquons que sous **(H1)**, l'étape E de l'algorithme EM consiste à calculer

$$Q(\theta_k\,,\,\theta) = -\psi(\theta) + \langle E^{\underline{y}}_{\theta_k} \tilde{S}(\underline{y}, Z), \phi(\theta)\rangle. \qquad (6)$$

puisque le reste $E^{\underline{y}}_{\theta_k} r(\underline{y},\underline{z})$ n'est pas fonction de θ.

A l'itération $k+1$ de l'algorithme SAEM, les étapes sont les suivantes :

- *Simulation* : on génère $m(k)$ réalisations indépendantes $\underline{z}_k(j)$ $(j = 1, \cdots, m(k))$ des données manquantes sous la loi *a posteriori* $k(\underline{z}|\underline{y};\theta_k)$.

- *Approximation Stochastique* : on réactualise l'approximation de la statistique exhaustive :

$$\underline{s}_{k+1} = \underline{s}_k + \gamma_k \left(\frac{1}{m(k)} \sum_{j=1}^{m(k)} \tilde{S}(\underline{y}, \underline{z}_k(j)) - \underline{s}_k \right), \tag{7}$$

où $\tilde{S}(\underline{y}, \underline{z}_k(j)) \in \mathcal{S}$ est la statistique exhaustive du modèle complet calculée au point $(\underline{y}, \underline{z}_k(j))$ et où $\{\gamma_k\}$ est une séquence positive de pas décroissants tel que $\sum \gamma_k = \infty$ et $\sum \gamma_k^2 < \infty$.

- *Maximisation* : on calcule $\theta_{k+1} \in \Theta$ qui maximise la vraisemblance complète au point $\underline{s}_{k+1}$:

$$\theta_{k+1} = \operatorname{Arg} \max_{\theta \in \Theta} \left\{ -\psi(\theta) + \langle \underline{s}_{k+1}, \phi(\theta) \rangle \right\}. \tag{8}$$

Pour comprendre cet algorithme, il est intéressant d'écrire (7) sous la forme d'un algorithme de gradient stochastique [17,22]. Pour cela, posons

$$h(\underline{s}) = E^y_{\tilde{\theta}(\underline{s})} \tilde{S}(\underline{y}, Z) - \underline{s} = \int \tilde{S}(\underline{y}, \underline{z}) k(\underline{z} | \underline{y}; \tilde{\theta}(\underline{s})) d\mu_Z(\underline{z}) - \underline{s} \tag{9}$$

$$e_{k+1} = \frac{1}{m(k)} \sum_{j=1}^{m(k)} \tilde{S}(\underline{y}, Z_k(j)) - E^y_{\tilde{\theta}(S_k)} \tilde{S}(\underline{y}, Z).$$

On peut alors montrer que :

$$\begin{aligned} \underline{s}_{k+1} &= \underline{s}_k + \gamma_k h(\underline{s}_k) + \gamma_k e_{k+1} \\ &= \underline{s}_k - \gamma_k [\mathcal{D}_s^2 \log f(\underline{s}_k; \tilde{\theta}(\underline{s}_k))]^{-1} \mathcal{D}_s \log g(\underline{y}; \tilde{\theta}(\underline{s}_k)) + \gamma_k e_{k+1}. \end{aligned} \tag{10}$$

L'algorithme cherche alors à résoudre $h(\underline{s}) = 0$, c'est-à-dire $\mathcal{D}_s \log g(\underline{y}; \tilde{\theta}(\underline{s})) = 0$, où $\mathcal{D}_s \log g$ est le vecteur des dérivées partielles de $\log g$ par rapport à s. Cela signifie donc que l'appoximation stochastique de la statistique exhaustive du modèle complet se fait suivant un schéma de type gradient stochastique avec gain matriciel.

Si l'on étudie maintenant la convergence de la suite θ_k, on en déduit que :

$$\begin{aligned} \theta_{k+1} &= \tilde{\theta}(\underline{s}_{k+1}) \\ &= \theta_k - \gamma_k R(\theta_k) [\mathcal{D}_\theta^2 \log g(\underline{y}; \theta_k)]^{-1} \mathcal{D}_\theta \log g(\underline{y}; \theta_k) + \gamma_k \eta_{k+1} + \mathcal{O}(\gamma_k^2), \end{aligned} \tag{11}$$

où $\mathcal{D}_\theta^2 \log g$ est la matrice des dérivées secondes de $\log g$ et où

$$\begin{aligned} R(\theta_k) &= [\mathcal{D}_\theta^2 \log f(\underline{s}_k; \theta_k)]^{-1} \mathcal{D}_\theta^2 \log g(\underline{y}; \theta_k) \\ \eta_{k+1} &= \mathcal{D}_s \tilde{\theta}(\underline{s}_k) e_{k+1}. \end{aligned}$$

L'estimation de θ se fait donc aussi suivant un shéma de type gradient stochastique avec gain matriciel. De plus ce gain matriciel se décompose en un produit de deux matrices : la matrice inverse des dérivées secondes de $\log g$ (ce qui correspondrait à un algorithme de type Newton-Raphson, donc optimal en termes de vitesse de convergence) et la matrice $R(\theta_k)$, qui est le rapport des informations de Fisher entre le modèle incomplet et le modèle complet. Donc, plus cette matrice est proche de l'identité, et plus l'algorithme est proche d'un algorithme de Newton-Raphson.

En fait, des résultats précis de convergence de la suite $\{\theta_k\}$ peuvent être obtenus sous les hypothèses suivantes :

(H2) $\forall \theta \in \Theta \quad E_\theta^y \|\tilde{S}(\underline{y}, Z)^\ell\| < \infty\,, \quad \ell = 1, 2$. De plus, on peut différencier deux fois sous le signe somme la relation $g(\underline{y};\theta) = \displaystyle\int_{\mathcal{Z}} f(\underline{y}, \underline{z}; \theta)\mu_Z(d\underline{z})$

(H3) La suite $\{\gamma_k\}$ vérifie : $0 \leqslant \gamma_k \leqslant 1$, $\displaystyle\sum_{k=0}^{\infty} \gamma_k = \infty$ et $\displaystyle\sum_{k=0}^{\infty} \gamma_k^2 < \infty$. Le nombre de simulations $m(k)$ tend vers une constante : $\lim\limits_{k\to\infty} m(k) = m$.

(H4) Dans $\overset{\circ}{\mathcal{S}}$ (intérieur de $\mathcal{S}$), il existe une fonction $\tilde{\theta} : \overset{\circ}{\mathcal{S}} \to \Theta$, telle que :

$$\forall \theta \in \Theta, \quad \forall \underline{s} \in \overset{\circ}{\mathcal{S}}, \quad L(\underline{s}; \tilde{\theta}(\underline{s})) \geqslant L(\underline{s}; \theta) \tag{12}$$

où, pour $\underline{y}$ fixé, $L : \mathcal{S} \times \Theta \to \mathbb{R}$ est (à une constante près) la log-vraisemblance complète :

$$L(\underline{s}; \theta) = -\psi(\theta) + \langle \underline{s}, \phi(\theta) \rangle. \tag{13}$$

De plus, pour tout $\underline{s} \in \overset{\circ}{\mathcal{S}}$, la matrice des dérivées secondes de L, $\mathcal{D}_\theta^2 L(\underline{s}; \tilde{\theta}(\underline{s}))$, est définie négative. La fonction $\tilde{\theta}(\underline{s})$ est deux fois différentiable sur $\overset{\circ}{\mathcal{S}}$ et continue sur $\mathcal{S}$.

(H5) Les variables aléatoires $Z_k(j)$, $k \geqslant 0$, $1 \leqslant j \leqslant m(k)$, sont indépendantes et distribuées suivant la loi $k(\underline{z}|\underline{y}; \theta_k)$.

Notons que sous **(H1)-(H4)**, l'étape M de l'algorithme consiste à calculer l'estimateur du maximum de vraisemblance de θ : $\theta_{k+1} = \tilde{\theta}(S_{k+1})$.

En utilisant les résultats sur la convergence des algorithmes stochastiques, [13, 17, 19, 33], nous avons le théorème suivant :

Théorème 2.1 *Si les hypothèses* **(H1)-(H5)** *sont vérifiées et si la suite* $\{S_k\}$ *reste dans un compact* $\mathcal{K}$ *de* $\mathcal{S}$ *avec probabilité 1, alors la suite* $\{\theta_k\}$ *converge presque-sûrement vers un point stationnaire de la vraisemblance* g.

Sous quelques hypothèses supplémentaires, on montre, en utilisant les résultats de [7], que l'algorithme ne peut converger que vers les maxima (éventuellement locaux) de la vraisemblance :

(H6) Pour tout $\underline{s} \in \overset{\circ}{\mathcal{S}}$, il existe un voisinage ouvert non vide $\mathcal{W}(\underline{s}) \subset \overset{\circ}{\mathcal{S}}$ où $E^{\underline{y}}_{\tilde{\theta}(\underline{s})}(\tilde{S}(\underline{y}, Z))$ est deux fois continument différentiable.

(H7) Pour tout θ appartenant à un compact $\mathcal{K} \subset \Theta$, la plus petite valeur propre de la matrice $E^{\underline{y}}_{\theta}(\tilde{S}(\underline{y}, Z) - E^{\underline{y}}_{\theta}\tilde{S}(\underline{y}, Z))(\tilde{S}(\underline{y}, Z) - E^{\underline{y}}_{\theta}\tilde{S}(\underline{y}, Z))'$ est minorée par une constante strictement positive (ici, M' est la transposée de M).

(H8) Les solutions dans $\mathcal{S}$ de l'équation $\mathcal{D}_{\theta} g(\underline{y}; \tilde{\theta}(\underline{s})) = 0$ sont des points isolés.

Théorème 2.2 *Si les hypothèses* **(H1)** $-$ **(H8)** *sont vérifiées et si la suite* $\{S_k\}$ *reste dans un compact* $\mathcal{K}$ *de* $\mathcal{S}$ *avec probabilité 1, alors la suite* $\{\theta_k\}$ *converge presque-sûrement vers un maximum (éventuellement local) de la vraisemblance (pénalisée)* g.

Stabilisation : Dans de nombreuses applications, la suite $\{S_k\}$ n'est pas bornée naturellement, on peut toutefois donner quelques critères qui assurent la propriété de $\mathcal{S}$-compacité [13, 33].

Lemme 2.3 *On suppose que les conditions* **(H1)-(H5)** *sont vérifiées. On suppose de plus :* **(i)** *il existe* $R_0 \geqslant 0$, *tel que, pour tout* $R \geqslant R_0$, *l'ensemble* $\mathcal{S}_R = \{s \in \mathcal{S} : R \leqslant g(\underline{y}; \tilde{\theta}(\underline{s}))\}$ *est compact,* **(ii)** $\underline{\lim} g(\underline{y}, \tilde{\theta}(S_k)) > R_0$ *avec probabilité 1. Alors, la* $\{S_k\}$ *est* $\mathcal{S}$*-compacte avec probabilité* 1.

Bien souvent, **(i)** est vérifiée car la vraisemblance incomplète $g(\underline{y}; \tilde{\theta}(\underline{s}))\}$ tend vers 0 lorsque $\|\underline{s}\| \to \infty$. Dans ce cas, vérifier **(ii)** revient à montrer que $\overline{\lim}\|S_k\| < \infty$. On peut alors avoir recours au lemme suivant :

Lemme 2.4 *On suppose qu'il existe une constante* $A \geqslant 0$ *telle que la condition* $\|S_{k-1}\| > A$ *implique* $E^{\underline{y}}_{\tilde{\theta}(S_{k-1})}(\|S_k\|) \leqslant \|S_{k-1}\|$ *avec probabilité* 1. *Alors,* $\overline{\lim}\|S_k\| < \infty$.

2.2. Une version recuit simulé de l'algorithme SAEM

Le théorème précédent nous assure que l'algorithme converge vers un maximum de la vraisemblance, mais pas nécessairement vers le maximum global si cette vraisemblance n'est pas concave. Une version *recuit simulé* de l'algorithme SAEM peut alors être envisagée.

Rappelons que l'algorithme de *recuit simulé* est une procédure itérative d'optimisation [17, 24, 28]. La fonction à optimiser peut être définie sur un ensemble fini de très grande dimension, auquel cas, on est confronté à un problème combinatoire difficile. Si maintenant, cette fonction est définie sur $\mathbb{R}^n$ et comporte un grand nombre de maxima locaux, le recuit simulé permet aux itérations de sortir des bassins d'attaction des maxima locaux afin de converger vers le maxima global.

L'algorithme de recuit simulé «classique» consiste ici à rajouter dans (7) un bruit blanc dont la variance décroît lentement. En utilisant (10), cela revient à construire l'algorithme suivant :

$$\underline{s}_{k+1} = \underline{s}_k + \gamma_k h(\underline{s}_k) + \gamma_k e_{k+1} + c_k \xi_{k+1} \tag{14}$$

où $\{\xi_k\}$ est une suite de vecteurs aléatoires indépendants et de loi gaussienne centrée de covariance identité, indépendante de $\underline{s}_0$ [17, 22]. On peut alors obtenir des résultats de convergence pour ce type d'algorithme suivant les vitesses de décroissance des suites $\{\gamma_k\}$ et $\{c_k\}$.

Malheureusement, cette stratégie ne convient pas ici, puisque la nouvelle valeur de la statistique $\underline{s}_{k+1}$ n'appartient plus nécessairement au convexe $\mathcal{S}$ lorsque l'on rajoute un bruit gaussien isotrope (cela peut conduire à estimer des variances négatives ...).

Considérons l'exemple du modèle de convolution $\underline{y} = f * \underline{z} + \sigma_\varepsilon \varepsilon$, où ε est un bruit blanc gaussien de variance 1, et supposons que l'on cherche à estimer le filtre f et la variance σ_ε^2. Les différents maxima de la vraisemblance correspondent aux différentes phases du filtre f. En particulier, si Z est gaussien, on ne peut espérer estimer cette phase : la vraisemblance est invariante par changement de phase, pour des filtres ayant même fonction de transfert. Par contre, si Z n'est pas gaussien, on peut espérer retrouver la phase par maximum de vraisemblance.

L'algorithme consiste à modifier artificiellement la variance du bruit en définissant un nouveau modèle $\underline{y} = f * \underline{z} + (\sigma_\varepsilon + T_k)\varepsilon$, où la suite T_k décroît lentement vers 0. A chaque itération, on simule les données manquantes avec une «fausse» loi *a posteriori*; cette loi est dispersée au cours des premières itérations, afin d'éviter les maxima locaux de la vraisemblance. Lorsque T est proche de 0, on retrouve la bonne loi et on converge vers le maximum de la vraisemblance des observations.

On propose en Figure 1 les résultats d'une simulation. Dans cet exemple, les Z_i sont des v.a. i.i.d., dont la loi est un mélange de deux gaussiennes centrées, de variances différentes. La variance σ_ε^2 est telle que le rapport des variances du signal et du bruit vaut 10. Le filtre à estimer est de longueur 21; l'initialisation est un pic placé en 5. On dispose de $n = 1000$ observations. Sans recuit (en posant $T_k = 0$), l'algorithme converge vers le maximum le plus proche, c'est-à-dire vers un filtre dont la phase dépend du point initial, alors que la version «recuit» donne ici une excellente estimation, puisque la bonne phase est estimée.

3. Reconstruction des données manquantes

3.1. L'estimateur MAP (Maximum a Posteriori)

Lorsque le modèle a été identifié (soit parce que θ est connu, soit parce qu'il a été estimé, soit parce qu'il a été fixé à une valeur arbitraire), on est amené à reconstruire les données manquantes $\underline{z}$. Dans les problèmes que nous considérons ici, il n'est pas possible «d'inverser» H défini dans (1) pour calculer directement $\underline{z}$. En d'autres termes, un grand nombre de $\underline{z}$ sont candidats comme solution : le problème est un *problème mal posé*. Plusieurs approches sont alors envisageables, suivant le type

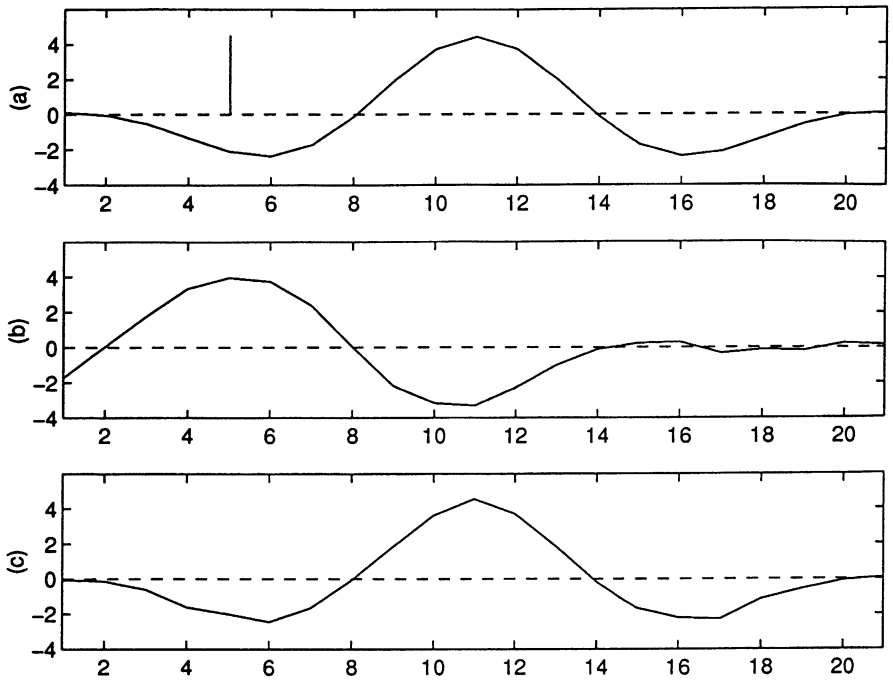


FIGURE 1
Estimation du filtre de convolution avec l'algorithme SAEM
(a) Filtre original et initialisation (pic en 5),
(b) Estimation par SAEM *sans recuit,*
(c) Estimation par SAEM *avec recuit*

d'application : on peut chercher à estimer les distributions marginales *a posteriori* $k(z_i|\underline{y};\theta)$ pour tout i. On peut aussi être conduit à obtenir une estimation ponctuelle de $\underline{z}$, c'est-à-dire à choisir un $\underline{z}$ particulier dans $\mathcal{Z}$. C'est ce type d'approche que nous privilègerons ici. Le principe pour reconstruire $\underline{z}$ est alors le suivant :

1) on définit des critères de bonne reconstruction,
2) on choisit $\underline{z} \in \mathcal{Z}$ qui est optimal pour ces critères.

Dans les méthodes bayésiennes de reconstruction, les critères doivent :

- être liés aux observations : on cherche une solution qui ajuste le mieux possible les observations,
- être liés à l'information *a priori* dont on dispose, ou au type de solution que l'on souhaite favoriser.

La combinaison de ces différents critères nous amène tout naturellement à considérer la distribution *a posteriori* des variables manquantes. Ainsi, choisir la valeur de $\underline{z}$ la plus vraisemblable au vu des observations et pour une valeur donnée

de θ nous amène à calculer l'estimateur du MAP (Maximum a Posteriori) :

$$\begin{aligned}\hat{\underline{z}} &= \text{Arg}\max_{\underline{z}\in\mathcal{Z}} k(\underline{z}|\underline{y};\theta) \\ &= \text{Arg}\max_{\underline{z}\in\mathcal{Z}} h(\underline{y}|\underline{z};\theta_1)\pi(\underline{z};\theta_2) \end{aligned} \tag{15}$$

où h est la vraisemblance conditionnelle de $\underline{y}$ et dépend d'un paramètre θ_1, tandis que π est la distribution *a priori* de z et dépend d'un paramètre θ_2 (on a décomposé ici le paramètre θ en $\theta = (\theta_1, \theta_2)$ où θ_1 est lié à h tandis que θ_2 est lié à π). Notons que cette distribution *a posteriori* ne permet pas seulement de reconstruire $\underline{z}$ en cherchant son mode, mais aussi d'évaluer la probabilité d'avoir n'importe quelle configuration $\underline{z}$ en calculant $k(\underline{z}|\underline{y};\theta)$.

L'ensemble des modèles que nous étudions sont des modèles exponentiels, tels que les densités h et π peuvent s'écrire sous la forme :

$$\begin{aligned} h(\underline{y}|\underline{z};\theta_1) &= C_1(\theta_1)\exp\{-U_1(\underline{z},\underline{y};\theta_1)\} \\ \pi(\underline{z};\theta_2) &= C_2(\theta_2)\exp\{-U_2(\underline{z};\theta_2)\}, \end{aligned} \tag{16}$$

où $C_1(\theta_1)$ et $C_2(\theta_2)$ sont des constantes de normalisation et U_1 et U_2 des fonctions appelées *énergies* ou *contrastes*. Donc, d'après (15) et (16), l'estimateur du MAP minimise l'énergie totale $U = U_1 + U_2$: pour $\underline{y}$ et θ fixés,

$$\begin{aligned}\hat{\underline{z}} &= \text{Arg}\min_{\underline{z}\in\mathcal{Z}} U(\underline{z},\underline{y};\theta) \\ &= \text{Arg}\min_{\underline{z}\in\mathcal{Z}}\{U_1(\underline{z},\underline{y};\theta_1) + U_2(\underline{z};\theta_2)\}. \end{aligned} \tag{17}$$

Dans le calcul du MAP, l'approche probabiliste sert essentiellement à construire une «bonne» fonction d'énergie à minimiser, c'est-à-dire permettant d'obtenir une solution «optimale» pour certains critères. D'un autre coté, les méthodes *déterministes* consistent généralement à construire directement, et «à la main», une fonction à minimiser qui tienne compte de l'ajustement aux observations et du type de solution que l'on souhaite obtenir. Ainsi, même si ces deux approches de reconstruction ont souvent été opposées, elles sont tout à fait similaires. Dans le premier cas, un modèle probabiliste permet de définir des distributions de probabilité *a priori* et *a posteriori* pour se ramener ensuite à un problème déterministe de minimisation d'énergie lorsqu'il s'agit de calculer le MAP. Dans le second cas, on construit une fonction à minimiser qui peut éventuellement s'interpréter en termes de distributions de probabilité.

La fonction à minimiser U est rarement convexe. On peut alors utiliser un algorithme de recuit simulé pour minimiser U. L'idée de cet algorithme stochastique consiste à définir une suite de lois *a posteriori* $\{k_n\}_{n>0}$ par

$$k_n(\underline{z}|\underline{y};\theta) = C_n\exp\left\{-\frac{U(\underline{z},\underline{y};\theta)}{T_n}\right\} \tag{18}$$

où $T_n \to 0$. On remarque alors que $k_n \to \mathcal{U}_0$ où $\mathcal{U}_0$ est la distribution uniforme sur l'ensemble des minima globaux de $U(\cdot, \underline{y}; \theta)$. On suppose maintenant que l'on sait simuler k_n, pour T_n constant, en utilisant une dynamique de type Metropolis (ou échantilloneur de Gibbs). L'algorithme de recuit consiste alors à faire décroître lentement à chaque itération de Metropolis (ou Gibbs) la température T_n vers 0. On peut alors, dans certains cas, obtenir des résultats théoriques précis de convergence, suivant le schéma de température utilisé [10, 24].

Notons qu'il existe également des algorithmes déterministes pour minimiser des fonctions non convexes, comme l'algorithme GNC, mais dont l'utilisation reste spécifique à telle où telle application et pour lesquels il n'existe pas de résultats théoriques généraux [6].

Signalons pour finir que d'autres estimateurs bayésiens existent : l'ICM (Iterative Conditionnal Mode) peut s'interpréter comme un algorithme de gradient : on cherche, par maximisations successives des distributions conditionnelles $k(z_i|\underline{y}, z_j\, j \neq i; \theta)$, à converger vers un maxima local de la distribution *a posteriori* $k(\underline{z}|\underline{y}; \theta)$. Le MPM (Marginal Posterior Mode) maximise les distributions marginales *a posteriori* $k(z_i|\underline{y}; \theta)$.

3.2. Exemple : la déconvolution

Dans le modèle de convolution

$$\underline{y} = f * \underline{z} + \sigma_\varepsilon \varepsilon \tag{19}$$

la série des signaux émis $\underline{z}$ est bien sûr inconnue; on cherche à la reconstruire à partir de la série des signaux reçus $\underline{y}$. f est un filtre à réponse impulsionnelle finie et ε est un bruit blanc gaussien de variance 1. On a donc

$$\theta_1 = (f\ ,\ \sigma_\varepsilon^2) \tag{20}$$

$$U_1(z, y; \theta_1) = \frac{1}{2\sigma_\varepsilon^2} \|\underline{y} - f * \underline{z}\|^2 \tag{21}$$

L'estimateur MAP de $\underline{z}$ est donc obtenu en minimisant la fonction d'énergie U définie par :

$$U(\underline{z}, \underline{y}; \theta) = \frac{1}{2\sigma_\varepsilon^2} \|\underline{y} - f * \underline{z}\|^2 + U_2(\underline{z}; \theta_2), \tag{22}$$

où $U_2(\underline{z}; \theta_2)$ dépend de la distribution *a priori* π des signaux émis. Voyons quelques exemples :

Déconvolution de signaux uniformément distribués : Supposons que les signaux émis sont indépendants et prennent leurs valeurs dans un ensemble E (cercle unitaire, alphabet fini, etc.) avec une distribution uniforme. Si n est le nombre de signaux émis, l'ensemble des configurations possibles est ici $\mathcal{Z} = E^{\otimes n}$, U_2 est une fonction constante et $\hat{\underline{z}}$ est défini par :

$$\hat{\underline{z}} = \operatorname{Arg} \min_{\underline{z} \in E^n} \|\underline{y} - f * \underline{z}\|^2. \tag{23}$$

(Pour ne pas alourdir les notations, la dépendance de $\hat{\underline{z}}$ vis-à-vis de θ et de $\underline{y}$ est implicite). Si cette modélisation *a priori* de $\underline{z}$ est raisonnable en communication numérique, ce n'est plus le cas en géophysique, en imagerie médicale ou en contrôle non-destructif, où $\underline{z}$ représente des coefficients de réflexion [30]. Ces coefficients de réflexion sont souvent modélisés par des variables Bernoulli-Gaussiennes.

Déconvolution de signaux Bernoulli-Gaussiens [30] : Supposons maintenant que la distribution des Z_i est la combinaison d'une distribution $\delta-$ Dirac en 0 et d'une distribution gaussienne centrée :

$$Z_i \rightsquigarrow p\mathcal{N}(0, \sigma_1^2) + (1-p)\delta_0. \tag{24}$$

où $0 \leqslant p \leqslant 1$. On montre que la fonction à minimiser définie dans (22) devient :

$$U(\underline{z}, \underline{y}; \theta) = \frac{1}{2\sigma_\varepsilon^2}\|\underline{y} - f * \underline{z}\|^2 + \frac{1}{2\sigma_1^2}\|\underline{z}\|^2 + K(p)n_0(\underline{z}), \tag{25}$$

où $n_0(\underline{z})$ est le nombre de termes nuls dans la série $\underline{z}$ et $K(p)$ une fonction décroissante de p :

$$K(p) = \log\left(\frac{1-p}{p}\right)^2 + \text{ cte.} \tag{26}$$

Les trois termes de cette fonction sont faciles à interpréter : on cherche à minimiser l'erreur quadratique totale, tout en contrôlant les amplitudes et le nombre des signaux non nuls.

Un exemple est proposé en Figure 2. Les signaux originaux (Figure 2-b) ont été convolués avec le filtre (Figure 2-a) et un bruit blanc gaussien a été rajouté avec un rapport Signal/Bruit de 10dB (cela signifie que la variance du signal est 10 fois celle du bruit). On peut constater au vu des résultats que l'estimateur MAP est ici très proche de la configuration originale.

Ce type de méthode peut être envisagée avec des distributions *a priori* plus complexes. On peut ainsi considérer que les signaux z_i sont corrélés. D'autre part, la déconvolution 1-D peut être étendue à la déconvolution 2-D. Cela se justifie en géophysique, par exemple.

Déconvolution 2-D de traces sismiques [29] : Les observations sont ici les traces sismiques qui donnent une image 2-D du sous-sol. En première approximation, on peut modéliser les traces sismiques comme la convolution d'une onde émise f avec la série des coefficients de réflexion du sous-sol. Un bruit de mesure additif ε est présent. La structure de couches du sous-sol et la continuité latérale des coefficients de réflexion nous conduisent à construire une distribution *a priori* de Z plus complexe que dans les exemples précédents. D'autre part, la distribution Bernoulli-Gaussienne n'est plus adaptée ici, puisqu'elle supposerait des couches parfaitement uniformes (sans aucune réflexion entre deux interfaces).

Nous allons en fait considérer deux types de réflecteurs : les réflecteurs principaux, qui indiquent les changements de couche, et les réflecteurs secondaires, qui indiquent les petites réflexions à l'intérieur des couches. On introduit un champ

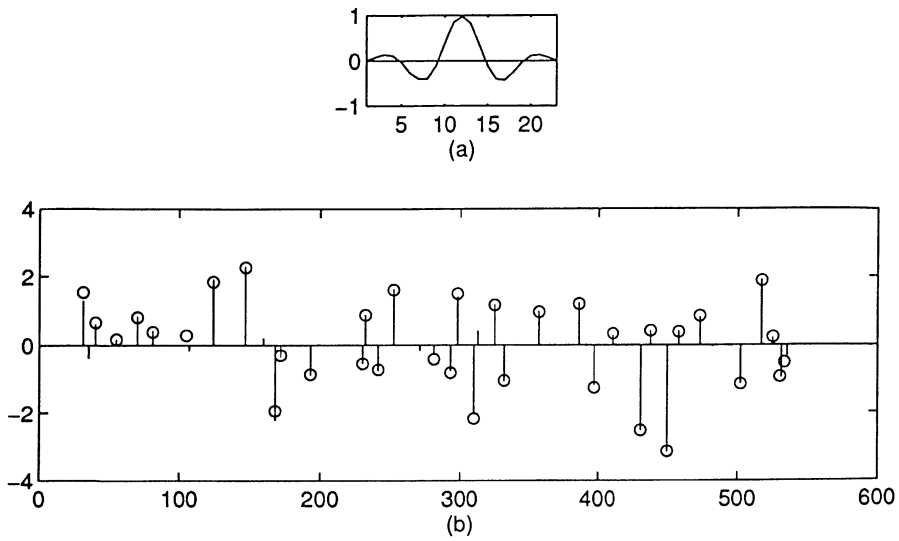


FIGURE 2
Déconvolution de signaux Bernoulli-Gaussiens
(a) : le filtre de convolution utilisé, (b) : les signaux originaux (\|) et estimés (o).

de Markov $Q = \{q_s\}_{s\in S}$ à valeurs dans $\{0,1\}$, et où $S \subset \mathbb{Z}^2$, pour modéliser la disposition des différents types de réflecteurs :

$$P\left(Q = \underline{q}\right) = C(\alpha,\beta)\exp\left\{-\alpha\sum_{s\in S} q_s - \beta \sum_{(s,t)\ voisins} \mathbb{1}_{q_s\neq q_t}\right\} \tag{27}$$

où $\beta > 0$.

Ici, $Q_s = 0$ (resp. $Q_s = 1$) si le réflecteur au site s est secondaire (resp. principal). Le problème est d'autant plus compliqué que le système de voisinages n'est pas homogène dans cet exemple, pour tenir compte des variations de pente dans les lignes de réflecteurs, les bifurcations, les failles, etc.

Si $Q_i = 0$, Z_i est une variable gaussienne de variance σ_0^2 alors que si $Q_i = 1$, Z_i est une variable gaussienne de variance σ_1^2 (avec $\sigma_1^2 > \sigma_0^2$). Toute cette étape de modélisation permet d'écrire la vraisemblance *a posteriori* du couple (Q, Z) et de calculer la fonction U à minimiser en fonction des différents paramètres :

$$\begin{aligned}\pi_{Q,Z}(\underline{q},\underline{z};\theta_2) &= \pi_Z(\underline{z}|\underline{q};\sigma_0^2;\sigma_1^2)\pi_{Q|Z}(\underline{q};\alpha,\beta)\\ &= \exp\left\{-\alpha\sum_{s\in S} q_s - \beta \sum_{(s,t)\ voisins} \mathbb{1}_{q_s\neq q_t} - \frac{1}{2\sigma_0^2}\sum_{q_s=0} z_s^2 - \frac{1}{2\sigma_1^2}\sum_{q_s=1} z_s^2\right\}. \end{aligned}\tag{28}$$

Donc

$$U(\underline{q},\underline{z},\underline{y};\theta) = \lambda_1\|\underline{y} - f * \underline{z}\|^2 + \lambda_2 \sum_{s\in S} q_s + \lambda_3 \sum_{(s,t)\ voisins} \mathbb{1}_{q_s\neq q_t} + \lambda_4 \sum_{q_s=0} z_s^2 + \lambda_5 \sum_{q_s=1} z_s^2. \quad (29)$$

où $\lambda_1 = 1/2\sigma_\varepsilon^2$, $\lambda_2 = \alpha + \log(\sigma_1/\sigma_0)$, $\lambda_3 = \beta$, $\lambda_4 = 1/2\sigma_0^2$, $\lambda_5 = 1/2\sigma_1^2$.

Chaque terme de (29) est directement relié à un critère de bonne reconstruction. Le premier terme contrôle l'ajustement aux observations, le second le nombre de réflecteurs principaux, le troisième la continuité latérale des réflecteurs principaux, les deux derniers contrôlent les amplitudes des coefficients de réflexion. Les coefficients λ_i indiquent les poids relatifs que l'on donne à ces critères.

Dans cet exemple, comme dans beaucoup de situations, l'ensemble des paramètres ne peut être estimé de façon satisfaisante. En effet, il est difficile ici de calculer l'estimateur de vraisemblance de θ, la forme de la vraisemblance incomplète étant particulièrement compliquée. Un algorithme de type SAEM peut être mis en œuvre pour estimer certains paramètres comme le filtre f, mais l'estimation de α, β, σ_0^2 et σ_1^2 est mauvaise (certaines hypothèses assurant la convergence de SAEM ne sont pas vérifiées, présence de très nombreux maximas locaux, problèmes d'identifiabilité, etc.). De plus, le choix de l'estimateur du maximum de vraisemblance n'est pas forcément le meilleur pour certains critères. Il arrive donc que certains paramètres ne soient pas estimés, mais fixés arbitrairement à des valeurs qui donnent une solution acceptable pour l'œil du spécialiste.

Posons $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$. Dans cet exemple, f est estimé, et donc fixé à sa valeur estimée. Tout cela nous conduit à paramétrer la fonction d'énergie U, non plus par θ, mais par $(f\ , \lambda)$. On obtient alors une solution $\hat{\underline{z}}(\lambda)$ qui dépend du λ choisi, c'est-à-dire des critères de reconstruction choisis.

Un exemple sur des données réelles est proposé en Figure 3. On voit dans les résultats que si $\lambda_3 = 0$ (ce qui correspond à déconvoluer chaque trace séparément), aucune continuité latérale n'est visible, contrairement aux deux autres reconstructions où $\lambda_3 = 1$. Notons que dans les trois exemples, l'ajustement aux observations est très bon, puisque la convolution $\hat{\underline{z}}(\lambda) * f$ donne des traces très proches des traces originales.

4. Estimation simultanée des paramètres et des données manquantes

Nous nous plaçons maintenant dans la situation où les paramètres du modèle (ou une partie des paramètres) sont estimés en même temps que les données manquantes. Comme dans l'exemple traitant de la déconvolution 2-D de traces sismiques, nous allons re-paramétrer le modèle en considérant $\phi \in \Phi$ qui contient les paramètres du modèle à estimer et $\lambda \in \Lambda$ qui contient les paramètres fixes, soit parce qu'ils sont connus, soit parce qu'ils ont été estimés au préalable, soit encore parce qu'ils sont fixés arbitrairement à des valeurs données. Nous allons construire une fonction U

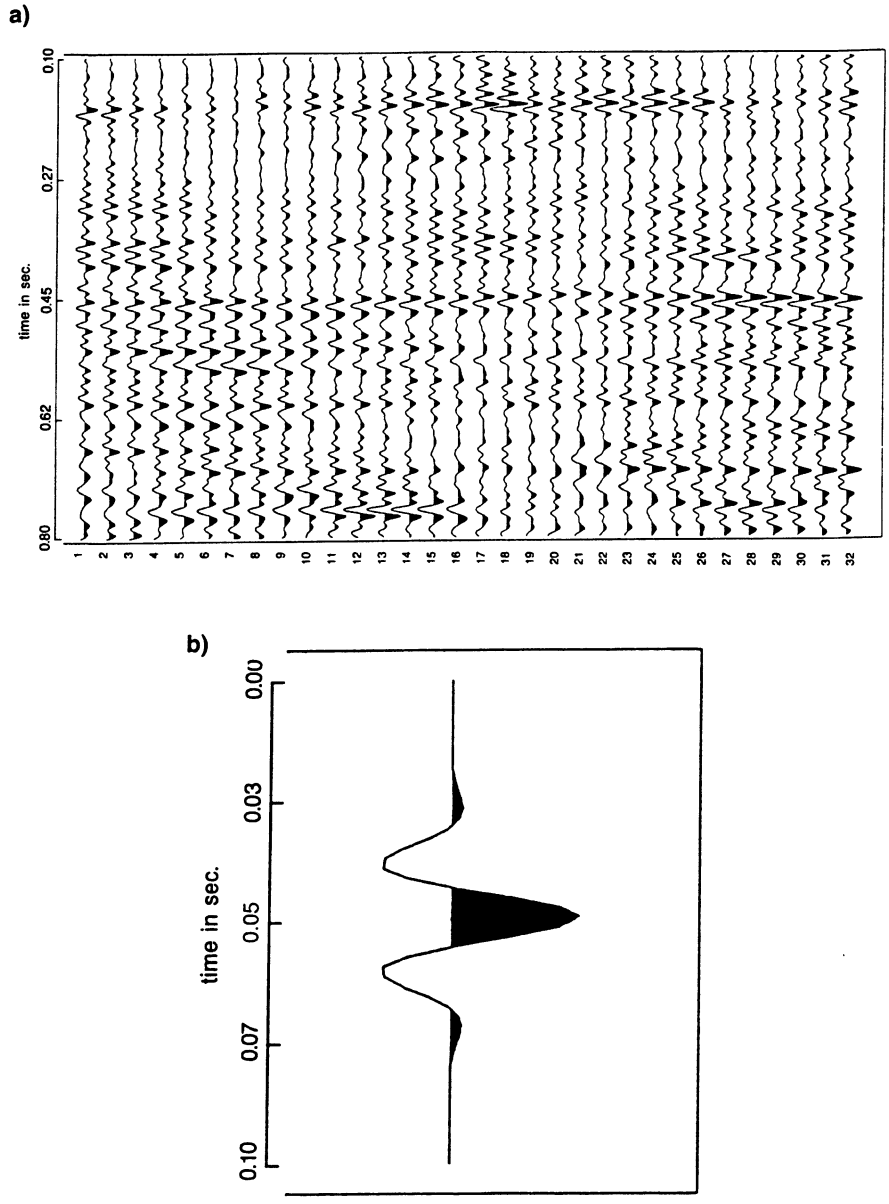
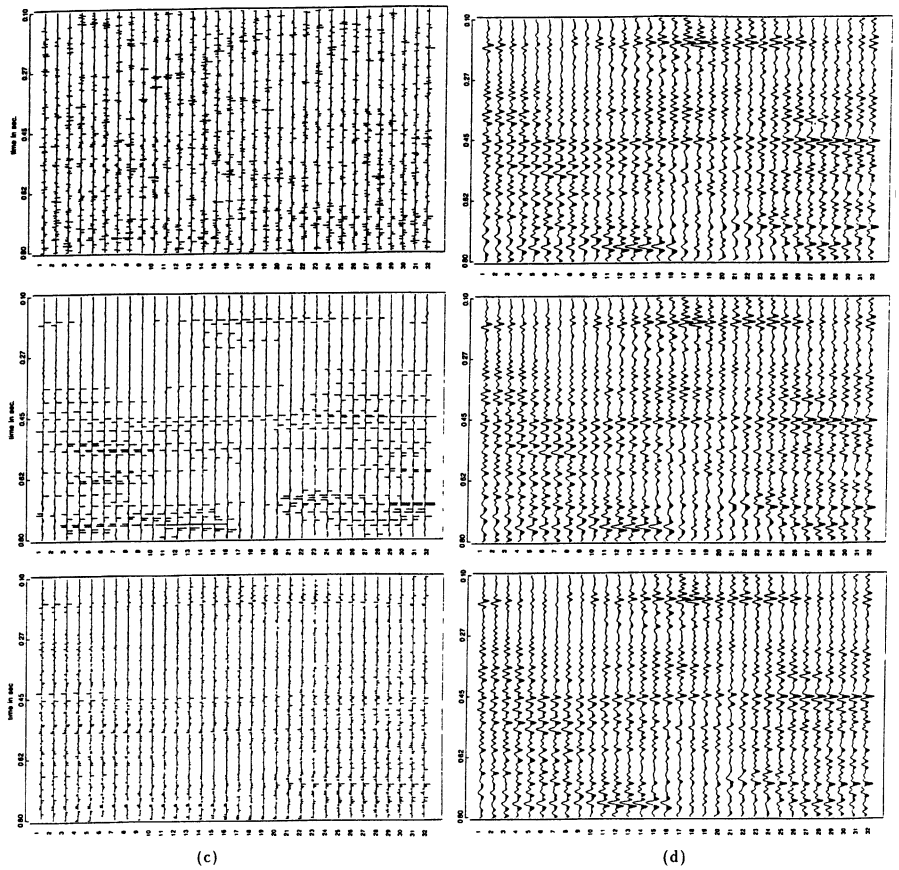


FIGURE 3
Déconvolution 2-D de traces sismiques :
(a) : les 32 traces sismiques originales; (b) : l'ondelette estimée.



(c) (d)

FIGURE 3

Déconvolution 2-D de traces sismiques :
(c) : 3 reconstructions de la série de réflecteurs obtenues avec différents paramètres $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$ *:* $\lambda = (10,\ 1,\ 0,\ 5,\ 0.25)$, $\lambda = (1,\ 1,\ 1,\ 5,\ 0.25)$ *et* $\lambda = (1,\ 0,\ 1,\ 5,\ 0.5)$ *(d) la convolution de ces trois séries estimées avec l'ondelette.*

à minimiser à la fois en $\underline{z}$ et en ϕ. Cette fonction U, définie sur $\mathcal{Z} \times \Phi \times \mathcal{Y} \times \Lambda$ et à valeurs dans $\mathbb{R}$, est telle que,

i) pour tout $\underline{z}$ et tout λ, l'estimateur $\hat{\phi}(\underline{z})$ est défini comme solution du problème de minimisation suivant :

$$U(\underline{z}, \hat{\phi}(\underline{z}), \underline{y}, \lambda) \leqslant U(\underline{z}, \phi, \underline{y}, \lambda) \quad \forall \phi \in \Phi \tag{30}$$

ii) pour tout λ, l'estimateur $\hat{\underline{z}}(\lambda)$ est défini comme solution du problème de minimisation suivant :

$$U(\hat{\underline{z}}(\lambda), \hat{\phi}(\hat{\underline{z}}(\lambda)), \underline{y}, \lambda) \leqslant U(\underline{z}, \hat{\phi}(\underline{z}), \underline{y}, \lambda) \quad \forall \underline{z} \in \mathcal{Z} \times \Phi \tag{31}$$

Nous exprimons ici explicitement la dépendance de $\hat{\underline{z}}$ vis-à-vis de λ puisque, comme nous le verrons au moyen d'exemples, la solution obtenue dépend de la valeur de λ choisie.

Comme précédemment, cette fonction U peut être définie à partir de la distribution *a posteriori* des données manquantes :

$$U(\underline{z}, \phi, \underline{y}, \lambda) = -\log k(\underline{z}|\underline{y}; \phi, \lambda). \tag{32}$$

La fonction U peut aussi être une fonction ayant de bonnes propriétés et facile à calculer. A titre d'exemple, nous proposons l'estimateur des moindres carrés pour détecter des ruptures dans la moyenne d'un processus aléatoire.

4.1. Clustering en présence de points aberrants

4.1.1. Le modèle

On considère ici que les observations sont n variables aléatoires indépendantes, issues d'un mélange de K populations, et contaminées par des points aberrants (*outliers*). On introduit donc une variable de classification Z qui prend ses valeurs dans $\{0, 1 \cdots K\}$ et telle que

$$Z_i \stackrel{\text{def}}{=} \begin{cases} k & \text{si l'observation } i \text{ appartient à la population } k,\ 1 \leqslant k \leqslant K, \\ 0 & \text{si l'observation } i \text{ est un point aberrant.} \end{cases} \tag{33}$$

D'autre part, nous allons supposer que Z est un n-échantillon de variables indépendantes et équidistribuées. On pose alors $p_k = \mathrm{P}(Z_i = k)$ pour tout $1 \leqslant i \leqslant n$ et $0 \leqslant k \leqslant K$. On a donc $\theta_2 = (p_0, p_1 \cdots, p_K)$. Ici, p_0 est la proportion *a priori* de points aberrants.

A titre d'exemple, nous allons faire les hypothèses suivantes sur Y et Z :

- Conditionnellement à Z, les Y_i sont indépendants.
- Conditionnellement à $Z_i > 0$, Y_i est un vecteur gaussien. Donc, pour tout i et pour tout $k > 0$,

$$h(y_i|Z_i = k; \theta_1) = (2\pi)^{-d/2} |\Gamma_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y - m_k)^t \Gamma_k^{-1} (y - m_k)\right\}. \tag{34}$$

  et $\theta_1 = (m_1, \Gamma_1, \cdots, m_K, \Gamma_K)$.
- Conditionnellement à $Z_i = 0$, Y_i est uniformément distribuée sur un compact O de $\mathbb{R}^d$. Il existe donc une constante $c_0 \in \mathbb{R}$ telle que,

$$h(y_i|Z_i = 0; \theta_1) \stackrel{\text{def}}{=} (2\pi c_0^2)^{-\frac{d}{2}} \mathbb{1}_{y_i \in O}. \tag{35}$$

- Si l' observation i n'est pas un point aberrant, Z_i est uniformément distribué sur $\{1, \cdots, K\}$,

$$p_1 = p_2 = \cdots = p_K = \frac{1 - p_0}{K}. \tag{36}$$

Par conséquent, la vraisemblance *a posteriori* de $\underline{z}$ s'écrit :

$$k(\underline{z}|\underline{y};\theta) = C \times \exp\left\{ -\frac{1}{2} \sum_{k=1}^{K} \left(\sum_{\{i|z_i=k\}} (y_i - m_k)^t \Gamma_k^{-1} (y_i - m_k) + n_k \log|\Gamma_k| \right) \right.$$
$$\left. - n_0 \left(\log\left(\frac{1-p_0}{Kp_0} \right) + d \log c_0 \right) \right\} \tag{37}$$

où $n_k = \sum_{i=1}^{n} \mathbb{1}_{z_i=k}$ est le nombre d'observations attribuées au cluster k et $n_0 = \sum_{i=1}^{n} \mathbb{1}_{z_i=0}$ le nombre de valeurs aberrantes.

i) Lorsque toutes les matrices de covariance sont égales à la matrice identité, à une constante inconnue près ($\Gamma_k = \sigma^2 I_d$), les paramètres à estimer sont $\phi = (m_1, \ldots, m_K)$. Pour $\underline{z}$ donné, $\hat{\phi}(\underline{z}) = (\hat{m}_1(\underline{z}), \ldots \hat{m}_K(\underline{z}))$ est l'estimateur des moindres carrés des moyennes :

$$\hat{m}_k(\underline{z}) = \operatorname{Arg\,min}_{m_k} \sum_{i=1}^{n} \|y_i - m_k\|^2 \mathbb{1}_{z_i=k}$$
$$= \frac{\sum_{i=1}^{n} y_i \mathbb{1}_{z_i=k}}{\sum_{i=1}^{n} \mathbb{1}_{z_i=k}}. \tag{38}$$

Posons

$$\lambda = \sigma^2 \left(2 \log\left(\frac{1-p_0}{Kp_0} \right) + d \log\left(\frac{c_0^2}{\sigma^2} \right) \right). \tag{39}$$

Pour λ donné, la configuration $\hat{\underline{z}}(\lambda)$ minimise $U(\underline{z}, \hat{\phi}(\underline{z}), \underline{y}, \lambda)$ défini par :

$$U(\underline{z}, \hat{\phi}(\underline{z}), \underline{y}, \lambda) = \sum_{k=1}^{K} \sum_{i=1}^{n} \|y_i - \hat{m}_k(\underline{z})\|^2 \mathbb{1}_{z_i=k} \; + \; \lambda \sum_{i=1}^{n} \mathbb{1}_{z_i=0}. \tag{40}$$

ii) Maintenant, si les matrices de covariance intra-groupes doivent être estimées, et en supposant que ces matrices soient toutes égales, les paramètres à estimer sont $\phi = (m_1, \dots m_K, \Gamma)$. Posons alors

$$\lambda = 2 \log \left(\frac{1 - p_0}{K p_0} \right) + d \log c_0^2 \tag{41}$$

et

$$U(\underline{z}, \phi, \underline{y}, \lambda) = \sum_{k=1}^{K} \sum_{i=1}^{n} \left((y_i - m_k)^t \Gamma^{-1} (y_i - m_k) + |\Gamma| \right) \mathbb{1}_{z_i = k} \; + \; \lambda \sum_{i=1}^{n} \mathbb{1}_{z_i = 0}. \tag{42}$$

Pour $\underline{z}$ donné, $\hat{\phi}(\underline{z})$ est l'estimateur empirique des moyennes et de la matrice d'autocovariance obtenue avec la configuration $\underline{z}$.

iii) Finalement, si l'on suppose que les matrices de covariance peuvent être différentes, les paramètres à estimer sont $\phi = (m_1, \Gamma_1, \dots m_K, \Gamma_k)$. Comme précédemment, pour $\underline{z}$ donné, $\hat{m}_k(\underline{z})$ et $\hat{\Gamma}_k(\underline{z})$ sont les estimateurs empiriques de la moyenne et de la matrice de covariance de Y_i dans le cluster k obtenue avec la configuration $\underline{z}$. Posons

$$\lambda = 2 \log \left(\frac{1 - p_0}{K p_0} \right) + d \log c_0^2 - d. \tag{43}$$

On montre alors que $\hat{\underline{z}}(\lambda)$ est obtenu en minimisant la fonction définie par

$$U(\underline{z}, \hat{\phi}(\underline{z}), \underline{y}, \lambda) = \sum_{k=1}^{K} \left(\log |\hat{\Gamma}_k(\underline{z})| \sum_{i=1}^{n} \mathbb{1}_{z_i = k} \right) \; + \; \lambda \sum_{i=1}^{n} \mathbb{1}_{z_i = 0}. \tag{44}$$

4.1.2. Exemple numérique

Nous avons simulé 57 variables aléatoires gaussiennes indépendantes dans $\mathbb{R}^2$ de la façon suivante :

Cluster 1 :

$$n_1 = 20 \; ; \;\; m_1 = (2, 2) \; ; \;\; \Gamma_1 = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$$

Cluster 2 :

$$n_2 = 30 \; ; \;\; m_2 = (-2, -2) \; ; \;\; \Gamma_2 = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

Valeurs aberrantes :

$$n_0 = 7$$

Les données originales sont représentées en Figure 4-a. La classification et la détection des valeurs aberrantes ont été effectuées dans un premier temps en minimisant la fonction définie dans (40).

La solution obtenue avec $\lambda = 5$ est représentée en Figure 4-b. Nous avons également représenté au moyen de cercles centrés sur les centres des clusters la règle de décision associée à cet algorithme : à l'intérieur d'un cercle, une observation est attribuée à un cluster, alors qu'elle est considérée comme valeur aberrante si elle est à l'extérieur.

Dans cet exemple, on détecte bien le bon nombre de valeurs aberrantes ($n_0 = 7$) avec $\lambda = 5$, mais 8 erreurs de classification sont faites. Ainsi, les trois points proches de $(0, 0)$ sont considérés comme suffisament loin des centres m_1 et m_2 pour être classés parmi les points aberrants si l'on suppose les deux composantes indépendantes ($\Gamma_1 = \Gamma_2 = \sigma^2 I_d$). En fait, une corrélation existe entre les deux composantes dans le cluster 2 (représenté par des $+$), et ce groupe de points est alors considéré comme plus proche du centre m_2 pour cette nouvelle métrique ($\Gamma_2(1, 2) > 0$). Signalons que la bonne configuration est obtenue en minimisant la fonction E_3 définie en (44) et en posant $\lambda = 4$ (Figure 4-c).

Bien sûr $\lambda = 4$ est une valeur particulière qui donne de bons résultats lorsque l'on minimise (44) dans cet exemple. Une plus petite valeur de λ conduit à sur-estimer le nombre de valeurs aberrantes et une plus grande valeur à le sous-estimer. La table 1 résume les résultats obtenus sur cet exemple avec différentes valeurs de λ. On peut voir ainsi que la bonne configuration est obtenue avec $3.5 \leqslant \lambda \leqslant 4.5$.

TABLEAU 1

Les classifications obtenues en minimisant la fonction d'énergie définie dans (44) avec différentes valeurs de λ. Ici, n_1 et n_2 représentent les nombres d'observations classées dans les clusters 1 et 2, n_0 est le nombre de valeurs aberrantes détectées et n_{err} le nombre total d'observations mal classées.

λ	n_1	n_2	n_0	n_{err}
2.0	13	20	24	17
2.5	17	28	12	5
3.0	19	29	9	2
3.5	20	30	7	0
4.5	20	30	7	0
5.0	21	30	6	1
10.0	22	30	5	2
15.0	24	32	1	6

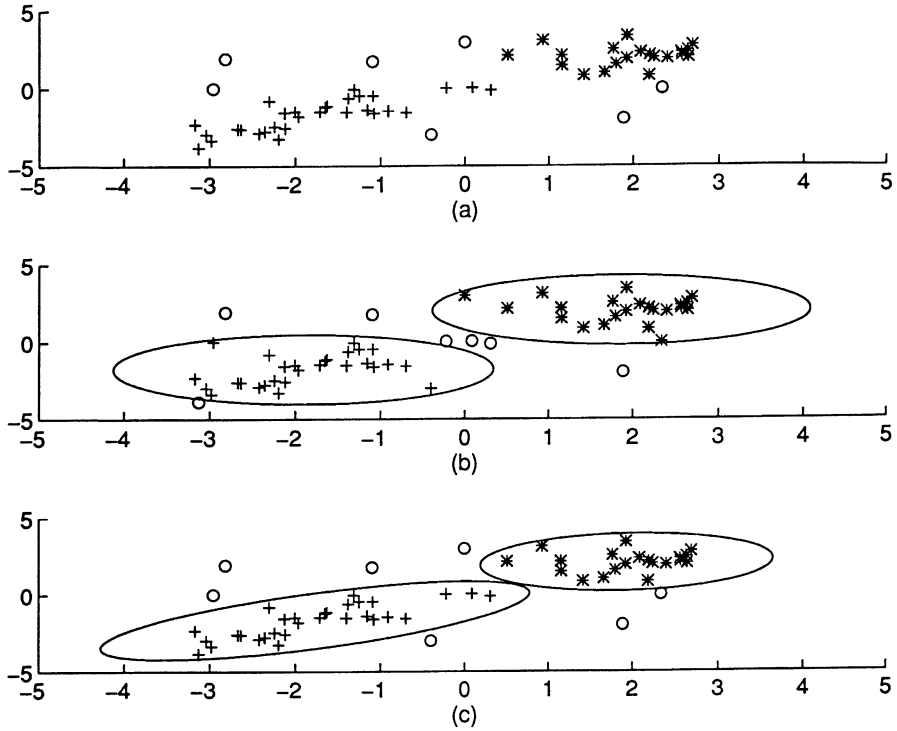


FIGURE 4

*Exemples de classification : (a) Les données originales; (b) La classification obtenue en minimisant (39) avec $\lambda = 5$; (c) La classification obtenue en minimisant (43) avec $\lambda = 4$; * : Cluster 1, + : Cluster 2, o : Valeur aberrante.*

4.2. Détection de ruptures

Segmenter un processus non stationnaire consiste à supposer qu'il est stationnaire par morceaux et à détecter les instants de rupture [1, 2, 8, 15, 31]. On ne cherche pas ici à détecter les ruptures séquentiellement, mais globalement. Un des intérêt majeur de la méthode proposée ici est sa capacité à calculer, par le choix d'un *a priori* judicieux, la meilleure solution possible, en accord avec le niveau de résolution choisi. Cet algorithme peut être utilisé dans de nombreux modèles paramétriques et non paramétriques. Il peut également être utilisé pour le lissage par fonctions splines lorsque le nombre et la position des nœuds sont inconnus. De plus, des résultats théoriques peuvent être obtenus sur la consistance et la vitesse de convergence de l'estimateur des instants de ruptures lorsque l'asymptotique porte sur la taille des segments.

4.2.1. Le modèle

Soit $Y = \{Y_t\}_{t\geqslant 0}$ un processus stationnaire par morceaux dont on observe une réalisation $\underline{y} = \{y_t\}_{1\leqslant t\leqslant n}$. Il existe donc des instants inconnus $\{z_k\}_{k\geqslant 1}$ que l'on cherche à déterminer tels que $(Y_{z_{k-1}+1}, \ldots, Y_{z_k})$ soit stationnaire.

Soit K le nombre de segments, Z prend ses valeurs dans l'ensemble

$$\mathcal{Z} = \bigcup_{1\leqslant K\leqslant n} \{\underline{z} = (z_1, \ldots, z_K) \;\; z_k \in \{1, \ldots, n\} \;\; 1 \leqslant z_1 < \ldots < z_K = n\}.$$

On définit la distribution *a priori* π de Z de la façon suivante : pour toute configuration $\underline{z} = (z_1, \ldots, z_K)$,

$$\pi(\underline{z}) = C(\alpha)e^{-\alpha K}\ , \;\; \alpha \in \mathbb{R}. \tag{45}$$

Le choix de cet *a priori* signifie en particulier que toutes les configurations ayant un nombre K de segments ont même probabilité.

Remarquons dans un premier temps que beaucoup de problèmes de détection de ruptures peuvent se ramener à un problème de détection de ruptures dans la moyenne d'un processus. Par exemple, détecter des ruptures dans la variance d'une suite de variables (centrées) $Y_1 \ldots Y_n$ revient à détecter des ruptures dans la moyenne de $Y_1^2 \ldots Y_n^2$, détecter des ruptures dans l'autocovariance revient à détecter des ruptures dans la moyenne de $Y_1Y_2, Y_2Y_3 \ldots Y_{n-1}Y_n$, etc. Nous avons donc étudié en détail ce cas particulier, et obtenu des résultats précis de convergence pour l'estimateur des moindres carrés proposé.

Mais avant d'établir ces résultats de convergence, voyons dans un premier temps la forme de l'estimateur utilisé dans certains modèles paramétriques et non paramétriques.

4.2.2. Rupture dans une tendance polynomiale

On considère le processus aléatoire réel :

$$Y_t = f(t) + \varepsilon_t, \quad 1 \leqslant t \leqslant n \tag{46}$$

où la fonction f est définie par morceaux :

$$f(t) = f_k(t), \quad z_{k-1} < t \leqslant z_k. \tag{47}$$

Dans le segment k, f_k est une fonction polynomiale de degré d :

$$f_k(t) = a_{k,d}t^d + a_{k,d-1}t^{d-1} + \cdots + a_{k,1}t + a_{k,0}, \quad t \geqslant 0, \;\; k \geqslant 1. \tag{48}$$

Si ε est une bruit blanc gaussien de variance σ_ε^2, la distribution *a posteriori* de Z est :

$$k(\underline{z}|\underline{y};\theta) = C(\theta)\exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{k=1}^{K}\sum_{i=z_{k-1}+1}^{z_k}(y_t - f_k(t))^2 + \alpha K\right\} \tag{49}$$

où $\theta = ((a_{k,j}), \sigma_\varepsilon^2, \alpha)$.

Comme dans l'exemple précédent, nous allons considérer ici que les paramètres à estimer sont les coefficients des polynomes $f_1 \dots f_K$. Posons donc $\phi = (a_{1,0}, a_{1,1}, \dots, a_{K,d})$ et $\lambda = 2\alpha\sigma_\varepsilon^2$. Pour λ donné, $(\hat{\underline{z}}(\lambda), \hat{\phi}(\hat{\underline{z}}(\lambda)))$ minimise la fonction d'énergie définie par :

$$U(\underline{z}, \phi, \underline{y}, \lambda) = \sum_{k=1}^{K}\sum_{i=z_{k-1}+1}^{z_k}(y_t - f_k(t))^2 + \lambda K. \tag{50}$$

Le premier terme de droite dans (50) est lié à l'ajustement aux observations, tandis que le second terme est lié au nombre de ruptures. Le paramètre λ contrôle le nombre de ruptures détectées, donc les probabilités de fausses alarmes et d'oublis.

Nous allons voir maintenant quelques exemples plus précisément.

Ruptures dans la moyenne d'un processus : On considère des polynômes de degré 0 : $f(t) = a_k$ pour tout $z_{k-1} < t \leqslant z_k$ et tout $k \geqslant 0$. Donc, $\hat{\underline{z}}$ est obtenu en minimisant :

$$U(\underline{z}, \hat{\phi}(\underline{z}), \underline{y}, \lambda) = \sum_{k=1}^{K}\sum_{t=z_{k-1}+1}^{z_k}(y_t - \hat{a}_k(\underline{z}))^2 + \lambda K, \tag{51}$$

où $\hat{a}_k(\underline{z})$ est la moyenne empirique de Y sur le segment k de la configuration $\underline{z}$:

$$\hat{a}_k(\underline{z}) = \frac{1}{z_k - z_{k-1}}\sum_{t=z_{k-1}+1}^{z_k} y_t. \tag{52}$$

Remarquons que minimiser la fonction U définie dans (51) revient à minimiser la fonction V définie par :

$$V(\underline{z}, \hat{\phi}(\underline{z}), \underline{y}, \lambda) = -\sum_{k=1}^{K}(z_k - z_{k-1})\hat{a}_k^2(\underline{z}) + \lambda K. \tag{53}$$

Le paramètre λ définit le niveau de résolution de la segmentation. On propose en Figure 5 différentes segmentations d'un même processus obtenues avec différentes valeurs de λ.

Lissage par fonctions splines (nombre et position des nœuds inconnus) : On pose maintenant $d > 0$. On suppose de plus que f est continue, ainsi que ses $d - 1$

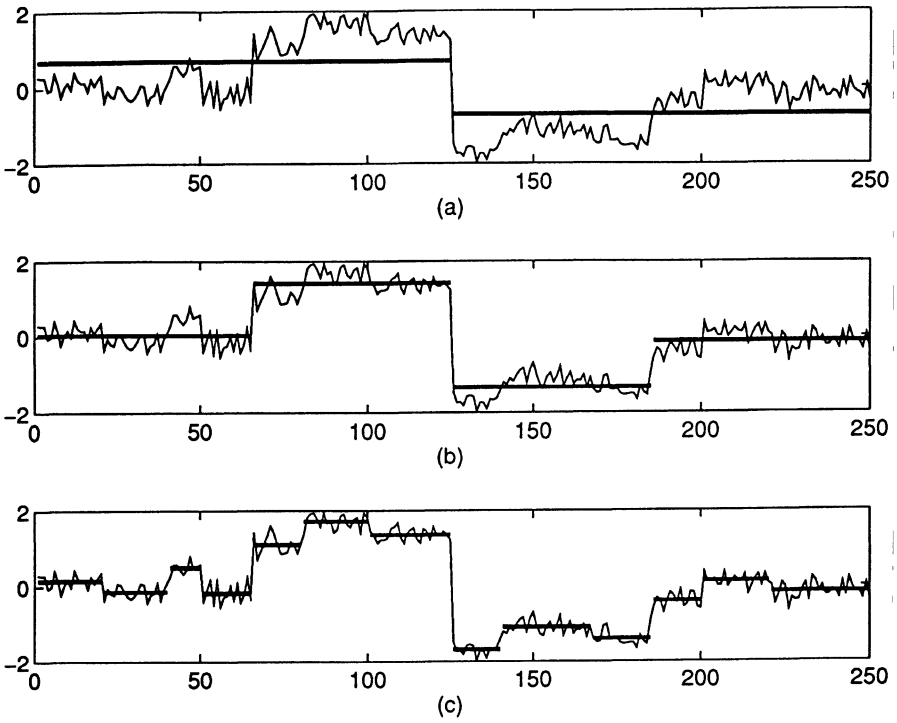


FIGURE 5
Détection de ruptures dans la moyenne d'un processus.
Exemples de segmentation avec différentes pénalisations :
(a) $\lambda = 100$*; (b)* $\lambda = 10$*; (c)* $\lambda = 1$.

premières dérivées. On est donc conduit à minimiser la fonction U définie dans (51) sous contraintes. Pour une configuration $\underline{z}$ donnée,

$$\hat{\phi}(\underline{z}) = \operatorname{Arg}\min_{\phi} \sum_{k=1}^{K} \sum_{t=z_{k-1}+1}^{z_k} (y_t - f_k(t))^2 \tag{54}$$

sous la contrainte :

$$\begin{aligned} f_k(z_k) &= f_{k+1}(z_k) \\ f_k^{(1)}(z_k) &= f_{k+1}^{(1)}(z_k) \\ &\vdots \\ f_k^{(d-1)}(z_k) &= f_{k+1}^{(d-1)}(z_k) \end{aligned}$$

Le problème d'optimisation sous contrainte est ici difficile. On peut, par exemple, utiliser un algorithme de recuit simulé avec contraintes [23].

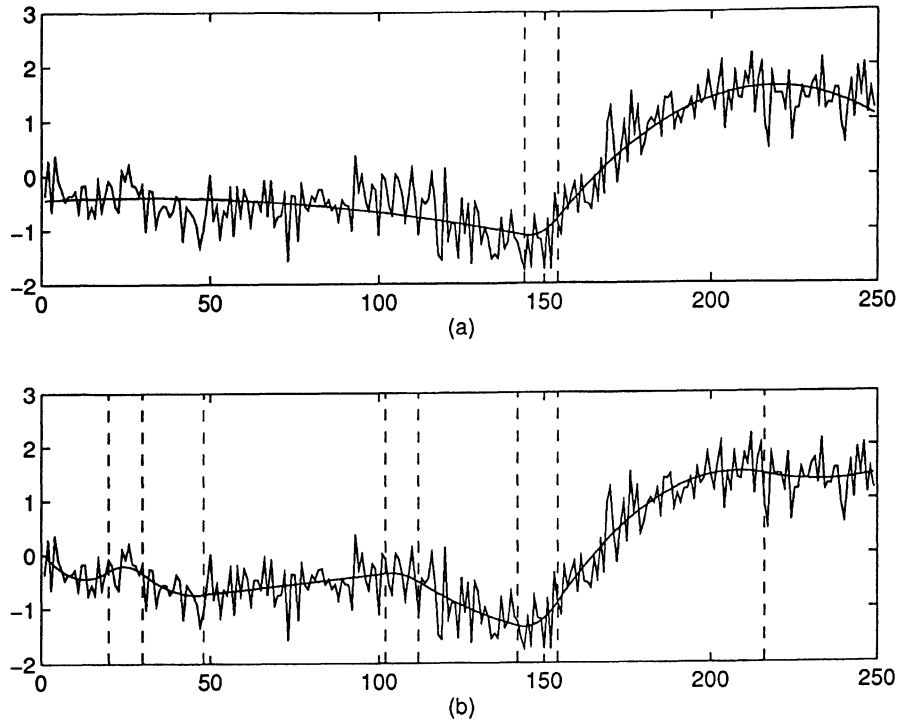


FIGURE 6

Détection de ruptures dans la dérivée seconde d'une tendance polynomiale de degré 2. Exemples de lissage avec différentes pénalisations : (a) $\lambda = 10$; (b) $\lambda = 1$. Les lignes verticales indiquent les instants de ruptures détectées.

La Figure 6 représente la trajectoire sur un axe d'un point de la colonne vertébrale durant une flexion. Cette trajectoire a été filmée et est donc bruitée. Nous avons utilisé ici des polynômes de degré 2 pour lisser cette série.

4.2.3. Ruptures dans des distributions non paramétriques

On se place ici dans le cas où la distribution du processus observé Y n'est pas paramétrique. L'idée consiste alors à construire à partir du processus Y un nouveau processus W, dont la distribution est paramétrique et dépend d'un paramètre θ, puis reconstruire Z en maximisant la distribution *a posteriori* $\mathrm{P}(Z = \underline{z}|W = \underline{w}; \theta)$.

Nous allons maintenant voir deux exemples de ce type de situation :

Ruptures dans la distribution marginale d'une suite de v.a. : On considère une suite Y de variables identiquement distribuées par morceaux. Dans cet exemple, la nouvelle statistique sera l'histogramme de Y : on définit la suite $\{x_m\}_{0 \leqslant m \leqslant M}$ tel que $x_0 < x_1 < \ldots < x_M$; on pose alors $W_t = m$ si $x_{m-1} < Y_t \leqslant x_m$.

On définit $p_{k,m} = \mathrm{P}(W_t = m)$ pour $z_{k-1} < t \leqslant z_k$ comme la probabilité que, dans le segment k, Y_t soit compris entre x_{m-1} et x_m. Les paramètres à estimer sont ici les probabilités $\phi = (p_{k,m})$, $1 \leqslant k \leqslant K, 1 \leqslant m \leqslant M$.

Lorsque Y est une suite de variables indépendantes, maximiser la distribution *a posteriori* $\mathrm{P}(Z = \underline{z}|W = \underline{w}; \theta)$ revient à minimiser la fonction

$$U(\underline{z}, \phi, \underline{w}, \lambda) = \sum_{k=1}^{n_z} \sum_{m=1}^{M} \left(\log p_{k,m} \sum_{t=z_{k-1}+1}^{z_k} \mathbb{1}_{w_t=m} \right) + \lambda K, \tag{55}$$

où ici $\lambda = \alpha$. Pour une configuration z donnée,

$$\hat{p}_{k,m}(\underline{z}) = \frac{1}{z_k - z_{k-1}} \sum_{t=z_{k-1}+1}^{z_k} \mathbb{1}_{w_t=m}, \tag{56}$$

et là encore, $\hat{\underline{z}}(\lambda)$ minimise $U(\underline{z}, \hat{\phi}(\underline{z}), \underline{w}, \lambda)$, où $\hat{\phi}(\underline{z}) = (\hat{p}_{k,m}(\underline{z}))$, $1 \leqslant l \leqslant K$, $1 \leqslant m \leqslant M$. Lorsque maintenant Y est une suite de variable dépendantes, la fonction U définie dans (55) n'est plus directement liée à la vraisemblance *a posteriori* de Z; c'est toutefois une fonction de contraste «raisonnable», qui donne de bons résultats dans la pratique.

La Figure 7 représente le rythme cardiaque d'un nouveau-né. On cherche à identifier des périodes de sommeil profond et des périodes de sommeil agité. Plutôt qu'une approche paramétrique, qui donne d'assez mauvais résultats ici, nous avons adopté une approche non paramétrique, en considérant que les ruptures affectent ici la distribution marginale des observations. Les ruptures détectées en minimisant (55) sont pratiquement les mêmes que celles détectées par les cliniciens en utilisant une information supplémentaire (mouvement oculaire, EEG, etc.)

Ruptures dans le spectre d'un processus : Supposons maintenant que les ruptures affectent le spectre du processus observé Y. Plutôt que de modéliser ce processus par un ARMA, et chercher à détecter des ruptures dans les paramètres de cet ARMA, nous allons chercher à détecter ces ruptures directement dans la distribution spectrale du processus, dans des bandes de fréquences $[x_{m-1}, x_m[$, $1 \leqslant m \leqslant M$, avec $0 \leqslant x_1 < x_2 \cdots x_M \leqslant \pi$. Nous allons donc construire une nouvelle statistique W construite à partir de la distribution spectrale empirique du processus.

On définit $W = \{W_{k,m}\}$, $1 \leqslant k \leqslant n_z$, $1 \leqslant m \leqslant M$, par

$$W_{km} = \int_{x_{m-1}}^{x_m} I_k(x)dx, \tag{57}$$

où $I_k(x)$ est le périodogramme de Y calculé sur le segment k.

Lorsque le processus Y est stationnaire entre deux instants de ruptures et sous des hypothèses assez générales, on peut montrer que les $W_{k,m}$ sont des estimateurs consistants de la distribution spectrale de Y :

$$W_{k,m} \underset{z_k - z_{k-1} \to \infty}{\longrightarrow} \int_{x_{m-1}}^{x_m} f_k(x)dx \tag{58}$$

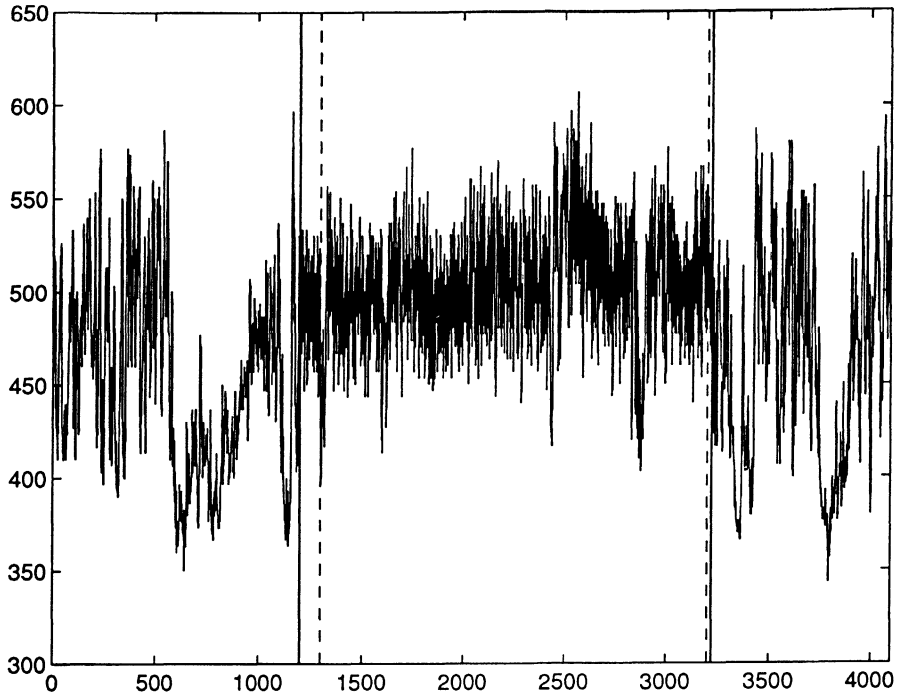


FIGURE 7

Détection de ruptures dans la distribution marginale d'une suite de variables. Application à la segmentation du rythme cardiaque :
--- les instants de rupture détectés par les cliniciens,
— les instants de rupture détectés par l'algorithme.

où f_k est la densité spectrale de Y dans le segment k. De plus, on peut montrer que les $W_{k,m}$ sont asymptotiquement gaussiens et indépendants (dans l'espace du temps et dans l'espace des fréquences) [31].

On se ramène donc à détecter des ruptures dans la moyenne de W, dont la loi limite est connue. On peut alors approximer la loi de W par sa loi limite et maximiser la distribution *a posteriori* de Z. Comme nous l'avons vu plus haut, on détecte des ruptures dans la moyenne d'une suite de variables aléatoires en minimisant la fonction V définie dans (53) avec $\hat{\phi}(\underline{z}) = W$. On cherche donc la configuration z qui minimise

$$V(\underline{z}, \hat{\phi}(\underline{z}), \underline{y}, \lambda) = -\sum_{k=1}^{K} \sum_{m=1}^{M} (z_k - z_{k-1}) W_{k,m}^2 + \lambda K. \tag{59}$$

Une application à l'analyse d'électro-encéphalogrammes est proposée Figure 8. Cinq bandes de fréquences connues, et qui correpondent à des activités électriques particulières du cerveau, ont été utilisées. Dans cet exemple, une période d'activité *alpha* est très bien détectée par l'algorithme vers $4s$ et une autre, plus courte, vers $1.5s$.

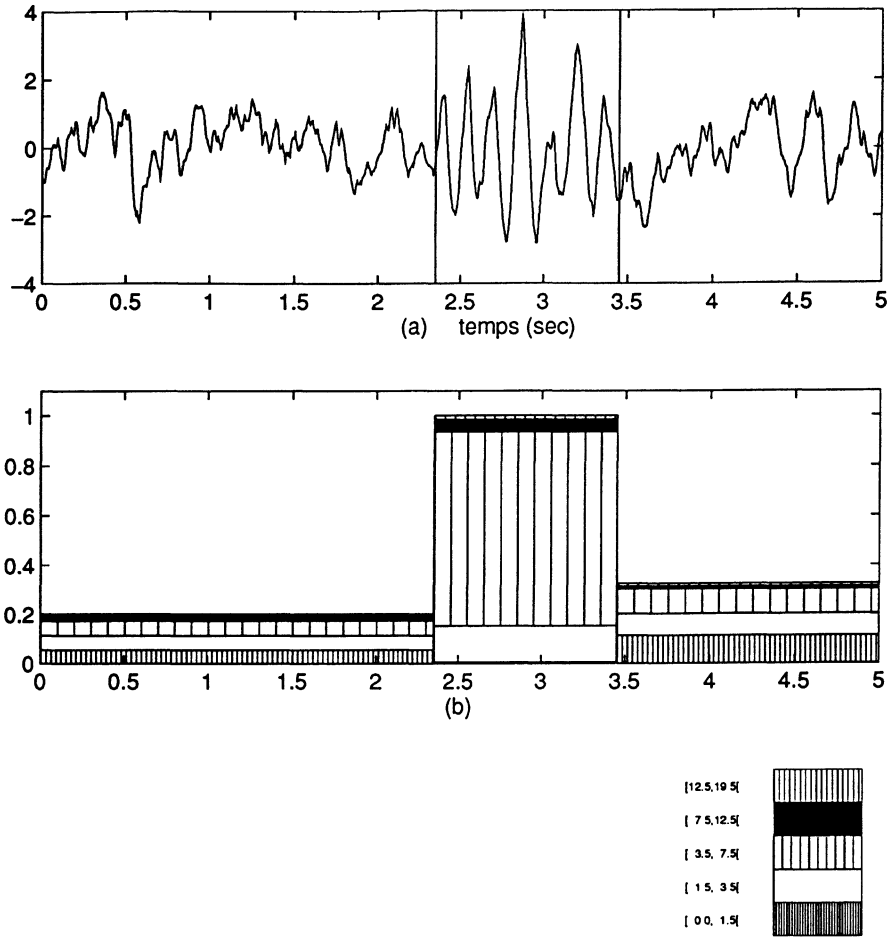


FIGURE 8

Détection de ruptures dans le spectre d'un processus. Application à la segmentation d'un EEG (a) l'EEG et les ruptures détectées; (b) estimation de la distribution spectrale dans les différents segments et dans les 5 bandes de fréquences (en Hz) ci-dessus.

4.2.4. Convergence de l'estimateur des moindres carrés

Revenons au problème de détection de ruptures dans la moyenne d'un processus aléatoire. On considère donc le modèle suivant :

$$Y_t = \mu(t) + \varepsilon_t \ ; \ t = 0, 1, 2, \ldots, n, \tag{60}$$

où μ est une fonction constante par morceaux et où $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ est un processus stochastique centré, stationnaire au second-ordre.

On suppose que $K-1$ ruptures affectent la fonction μ à des instants inconnus $\{z_k\}_{1\leqslant k\leqslant K}$ avec $z_0 = 0 < z_1 < \ldots < z_K = n$. Il y a donc K segments et la moyenne μ prend K valeurs : $\mu(t) = \mu_k$ si $z_{k-1}+1 \leqslant t \leqslant z_k$ pour $k \in 1, 2, \ldots, K$. On définit alors la configuration des instants de ruptures normalisés $\underline{\tau} = \{\tau_k\}_{1\leqslant k\leqslant K}$ où $\tau_k = z_k/n$. Le problème consiste à estimer le vecteur de moyennes $\underline{\mu} = \{\mu_k\}_{1\leqslant k\leqslant K}$ et à estimer la configuration des instants de ruptures normalisés $\underline{\tau}$ à partir de n observations $Y_1, Y_2, \ldots, Y_n$. Alors que nous avions proposé plus haut un estimateur des instants de ruptures absolus $\underline{z}$ (entre 1 et n), les résultats de convergence vont porter sur l'estimateur des instants de ruptures normalisés $\underline{\tau}$ (entre 0 et 1). La véritable configuration de ruptures est notée $\underline{\tau}^\star$ tandis que le véritable vecteur de paramètres est noté $\underline{\mu}^\star$.

Pour obtenir des résultats dans un cadre assez général, nous considérerons l'hypothèse suivante sur le processus ε : pour $1 \leqslant \phi < 2$,

• **H1**(ϕ) Il existe $C_1 < \infty$ tel que pour tout $n > 0$,

$$\mathrm{E}\sum_{i=1}^{n} \varepsilon_i^2 \leqslant C_1 n^\phi$$

L'hypothèse **H1** est vérifiée pour une grande famille de processus ε. Si ε est *α-mélangeant*, **H1** est vérifiée pour une certaine valeur de ϕ qui dépend de la décroissance des coefficients de mélange fort. (bien sûr pour $\phi = 1$ si ε est un bruit blanc). Si ε est un processus stationnaire à dépendance forte tel que $\mathrm{E}\varepsilon_s\varepsilon_t \approx C|s-t|^{d-1}$, où $0 \leqslant d < 1$, alors $\mathrm{E}\sum_{i=1}^{n} \varepsilon_i^2 = O(n^{d+1})$ et **H1** est vérifiée avec $\phi = d+1$.

D'autre part, nous serons amenés à supposer que la durée d'un segment est bornée inférieurement, tout comme les sauts de moyenne. Plus précisément :

• **H2**

i) Il existe $0 < \Delta_\tau^\star < 1$ tel que $\tau_k^\star - \tau_{k-1}^\star \geqslant \Delta_\tau^\star$, $1 \leqslant k \leqslant K$.

ii) Il existe $\Delta_\mu^\star > 0$ tel que $\mu_k^\star - \mu_{k-1}^\star \geqslant \Delta_\mu^\star$, $2 \leqslant k \leqslant K$.

Lorsque le nombre de ruptures est connu, égal à K, notons $\mathcal{A}_K$ l'ensemble des configurations possibles,

$$\mathcal{A}_K = \{\tau_0 = 0 < \tau_1 < \ldots < \tau_K = 1\}$$

Nous avons le résultat suivant :

Théorème 4.1. *Soit $\hat{\underline{\tau}}_n$, l'estimateur des moindres carrés des instants de rupture défini comme la solution du problème de minimisation suivant :*

$$U(\underline{\tau}, \underline{\mu}, \underline{y}) = \sum_{k=1}^{K} \sum_{t=[n\tau_{k-1}]+1}^{[n\tau_k]} (Y_t - \mu_k)^2 \tag{61}$$

$$U(\hat{\underline{\tau}}_n, \mu(\hat{\underline{\tau}}_n), \underline{y}) \leqslant U(\underline{\tau}, \underline{\mu}, \underline{y}) \quad \forall (\underline{\tau}, \underline{\mu}) \in \mathcal{A}_K \times \mathbb{R}^K.$$

Alors, sous (**H1**(ϕ),**H2**), *avec* $1 \leqslant \phi < 2$, $(\hat{\underline{\tau}}_n, \underline{\mu}(\hat{\underline{\tau}}_n))$ *converge en probabilité vers* $(\underline{\tau}^\star, \underline{\mu}^\star)$.

(Signalons que la convergence presque-sûre peut également être établie sous des hypothèses de mélange fort, [34]).

La preuve est classique, car l'estimateur est défini comme le minimum d'un processus de contraste. Il s'agit donc de s'assurer que la fonction de contraste associée au processus de contraste à un minimum global unique à la valeur vraie des paramètres, et d'autre part que le processus de contraste converge uniformément vers la fonction de contraste.

On peut en fait obtenir des résultats plus fins en montrant que la vitesse de convergence de l'estimateur des instants de rupture est $n^{2-\phi}$:

Lemme 4.2. *Supposons que les hypothèses* **H1**(ϕ), **H2** *soient vérifiées avec* $1 < \phi < 2$. *Alors* $\{n^{2-\phi}\|\hat{\underline{\tau}}_n - \underline{\tau}^\star\|_\infty\}$ *est uniformément tendue en* $\mathrm{P}_{\underline{\tau}^\star}$*-probabilité :*

$$\lim_{M\to\infty} \lim_{n\to\infty} \mathrm{P}_{\underline{\tau}^\star}(n^{2-\phi}\|\hat{\underline{\tau}}_n - \underline{\tau}^\star\|_\infty \geqslant M) = 0.$$

Supposons maintenant que le nombre de segments K est inconnu mais qu'un majorant $\overline{K}$ soit connu. Nous nous proposons d'estimer la configuration des ruptures $\underline{\tau}$ et le nombre de ruptures K au moyen de l'estimateur des moindres carrés pénalisés, défini de la façon suivante :

$$(\hat{\underline{\tau}}_n, \mu(\hat{\underline{\tau}}_n), \hat{K}_n) = \arg \min_{1\leqslant K\leqslant \overline{K}} \min_{\underline{\tau}\in\mathcal{A}_K} \left\{ \sum_{k=1}^{K} \sum_{t=[n\tau_{k-1}]+1}^{[n\tau_k]} (Y_t - \mu_k)^2 + \lambda_n K \right\}. \tag{62}$$

La suite $\{\lambda_n\}$ est une suite positive décroissant vers 0 avec une vitesse compatible avec l'obtention de résultat asymptotique de consistance et de vitesse. Le procédé est intuitivement clair : il s'agit de compenser le critère des moindres carrés de façon à pénaliser la sur-segmentation (le critère des moindres carrés non compensés décroissant avec le nombre de segments). La compensation ne doit pas toutefois être trop rapide, afin de ne pas sous-estimer le nombre de segments. La vitesse de la compensation qu'il faut appliquer au critère des moindres carrés dépend du type d'hypothèse dont on dispose :

• **H3** L'hypothèse **H1**(ϕ) est satisfaite pour $1 \leqslant \phi < 2$ et la suite $\{\lambda_n\}$ est telle que

$$\lambda_n \underset{n\to\infty}{\longrightarrow} 0 \text{ et } \frac{n^{\phi-2}}{\lambda_n} \underset{n\to\infty}{\longrightarrow} 0. \tag{63}$$

Nous avons alors le résultat suivant :

Théorème 4.3. *L'estimateur des moindres carrés pénalisés* $(\hat{\underline{\tau}}_n, \mu(\hat{\underline{\tau}}_n), \hat{K}_n)$ *converge en probabilité sous (***H2, H3***) vers* $(\underline{\tau}^\star, \underline{\mu}^\star, K^\star)$*, où* $K^\star$ *est le vrai nombre de segments dans* $\underline{\tau}^\star$.

Références

[1] BASSEVILLE M and NIKIFOROV N (1993), *The Detection of abrupt changes – Theory and applications*, Prentice-Hall : Information and System sciences series.

[2] BENVENISTE A, METIVIER M, and PRIOURET P (1987), *Algorithmes adaptatifs et approximations stochastiques*, Masson.

[3] BESAG J.E. (1986), «On the statistical analysis of dirty pictures», *J. R. Statist. Soc. B*, vol. 48, pp. 259–302.

[4] BINDER D.A. (1978), «Bayesian clustering analysis», *Biometrika*, vol. 68, pp. 275–286.

[5] BISCARAT J.C. (1994), «Almost sure convergence of a class of stochastic approximation algorithms», *Stoch. Proc. Appl.*, vol. 50, pp. 83–100.

[6] BLAKE A. and ZISSERMAN A. (1987), *Visual Reconstruction*. Cambridge, MA : MIT Press.

[7] BRANDIERE O. and DUFLO M. (1996), «Les algorithmes stochastiques contournent-ils les pièges ?», *Annales de l'Institut H. Poincaré*, vol. 32, pp. 395–427.

[8] BRODSKY B.E. and DARKHOVSKY B.S. (1993), *Nonparametric methods in change-point problems*. Kluwer Academic Publishers, the Netherlands.

[9] CARDOSO J.F. and MOULINES E. (1994), «How much more DOA information in higher-order statistics ?», in *Proc. Int. Conf. on Acoust. Speech and Sig. Proc.*, pp. 199–202.

[10] CATONI O. (1992), «Rough large deviation estimates for simulated annealing : application to exponential shedules», *The Annals of Proba.*, vol. 20, no. 3, pp. 1109–1146.

[11] CELEUX G. and DIEBOLT J. (1985), «The SEM algorithm : a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem», *Computational Statistics Quarterly*, vol. 2, pp. 73–82.

[12] CELEUX G. and DIEBOLT J. (1992), «A stochastic approximation type EM algorithm for the mixture problem», *Stoch. and Stoch. Reports*, vol. 41, pp. 119–134.

[13] DELYON B. (1996), «General results on stochastic approximation», *IEEE Trans. On Autom. Control*, 41, pp. 1245–1255.

[14] DEMPSTER A., LAIRD N., and RUBIN D. (1977), «Maximum-likelihood from incomplete data via the EM algorithm», *J. R. Statist. Soc. B*, vol. 39, pp. 1–38.

[15] DESHAYES J. and PICARD D. (1986), «Off-line statistical analysis of change point models using non-parametric and likelihood method», in *Detection of abrupt changes in signals and systems*, Lecture Notes in Control and Information Sciences, pp. 259–275, Edited by M. Basseville et A. Benveniste.

[16] DONOHO D. (1981), «On minimum entropy deconvolution», in *Applied time-series analysis II*, pp. 565–609, Academic Press.

[17] DUFLO M. (1996), *Algorithmes Stochastiques*. SMAI, Springer.

[18] GAMBOA F. and GASSIAT E. (1996), «Blind deconvolution of discrete linear systems», *The Annals of Stat.*, vol. 24, 1964–1981.

[19] FORT J.C. and PAGÈS G. (1996), «Convergence of stochastic algorithms : from Kushner-Clark theorem to the Lyapounov functional method», *Adv .appl. prob.*, vol. 28, pp. 1072–1094.

[20] GAETA M. and LACOUME J.L. (1990), «Estimateurs du maximum de vraisemblance étendus à la séparation de sources non gaussiennes», *Traitement du Signal*, vol. 7, no. 5, pp. 419–434.

[21] GELFAND A.E. and SMITH A.F.M. (1990), «Sampling based approach for calculating marginal densities», *J. Amer. Stat. Assoc.*, vol. 85, pp. 398–409.

[22] GELFAND S.B. and MITTER S.K. (1993), «Metropolis-type annealing algorithms for global optimization in $\mathbb{R}^{d*}$», *SIAM J. Control and optimization*, vol. 31, no. 1, pp. 111–131.

[23] GEMAN D. (1990), *Random Fields and Inverse Problems in Imaging*. Lecture Notes in Mathematics, Springer-Verlag.

[24] GEMAN S. and GEMAN D. (1984), «Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images», *IEEE Trans. on Pattern Anal. Machine Intell.*, vol. 6, pp. 721–741.

[25] HAWKINS D.M. (1980), «Identification of outliers», in *Monographs on Applied Probability and Statistics*, Chapman and Hall.

[26] JUTTEN C. and Hérault J. (1988), «Une solution neuromimétique au problème de séparation de sources», *Traitement du Signal*, vol. 5, no. 6, pp. 389–403.

[27] KAUFMAN L. and ROUSSEUW P.J. (1990), *Finding groups in data. An introduction to cluster analysis*. Wiley : New-York.

[28] KIRKPATRICK S., GELATT C.D., and VECCHI M.P. (1983), «Optimisation by simulated annealing», *Science*, vol. 220, no. 13 May, pp. 621–680.

[29] LAVIELLE M. (1991), «2-D Bayesian deconvolution», *Geophysics*, vol. 56, no. 12, pp. 2008–2018.

[30] LAVIELLE M. (1993), «Bayesian deconvolution of Bernoulli-Gaussian processes», *Signal Processing*, vol. 33, pp. 67–79.

[31] LAVIELLE M. (1993), «Detection of changes in the spectrum of a multidimensional process», *IEEE Tr. on Sig. Proc.*, vol. 41, no. 2, pp. 742–749.

[32] LAVIELLE M. (1995), «A stochastic procedure for parametric and non-parametric estimation in the case of incomplete data», *Signal Processing*, vol. 42, pp. 3–17.

[33] LAVIELLE M. and MOULINES E. (1995), «On a stochastic approximation version of the EM algorithm», tech. rep., Publication Université Paris-Sud.

[34] LAVIELLE M. and MOULINES E. (1997), «Détection de ruptures multiples dans la moyenne d'un processus aléatoire», *note C.R. Acad. Sci. Paris*, t. 324, I, pp. 239–243.

[35] PICARD D. (1985), «Testing and estimating change points in time series», *J. Applied Prob.*, vol. 17, pp. 841–867.

[36] QIAN W. and TITTERINGTON D.M. (1991), «Estimation of parameters in hidden Markov models», *Phil. Trans. Roy. Soc. London, A*, vol. 337, pp. 407–428.

[37] ROBERT C.P. (1996), *Méthodes de Monte Carlo par Chaînes de Markov*. Statistique mathématique et Probabilité, Economica.

[38] ROSENFELD A. and KAK A. (1982), *Digital picture processing*, vol. 2. London : Academic-Press.

[39] SHALVI O. and WEINSTEIN E. (1990), «New criteria for blind deconvolution of nonminimum phase systems (channels)», *IEEE Tr. on Info. Theory*, vol. 36, no. 2, pp. 312–321.

[40] TITTERINGTON D.M., SMITH A., and MAKOV U. (1985), *Statistical analysis of finite mixture distributions*. Wiley-New York.

[41] WEI G. and TANNER M. (1990), «A Monte-Carlo implementation of the EM algorithm and the Poor's Man's data augmentation algorithm», *J. Amer. Stat. Assoc.*, vol. 85, pp. 699–704.

[42] WU C. (1983), «On the convergence property of the EM algorithm», *The Annals of Stat.*, vol. 11, pp. 95–103.

[43] YOUNES L. (1989), «Parametric inference for imperfectly observed Gibbsian fields», *Prob. Theory Rel. Fields*, vol. 82, pp. 625–645.