APPROXIMATION STOCHASTIQUE DE L'ALGORITHME EM

Marc LAVIELLE
CNRS URA 743, Université Paris-Sud 91405 Orsay
UFR de Mathématiques et Informatique, Université Paris V

ERIC MOULINES
Télécom Paris/URA 820, 46, rue Barrault, 75634 Paris CEDEX 13

L'algorithme EM est une procédure très souvent utilisée pour calculer l'estimateur du maximum de vraisemblance dans des modèles à données incomplètes. Dans certaines situations, cet algorithme ne peut pas être utilisé car le calcul explicite des lois conditionnelles (étape E) est impossible. Pour remédier à cette limitation, une approximation stochatique de l'algorithme EM (SAEM) est proposée. En utilisant des résultats récents sur les algorithmes stochastiques, la convergence de l'algorithme SAEM vers un maximum de la vraisemblance incomplète est démontrée pour une très grande classe de vraisemblances complètes.

The Expectation-Maximization (EM) algorithm is a very popular tool for computing maximum likelihood estimates in incomplete data models. In certain situations however, this approach is not applicable, because the expectation step cannot be performed in closed-form. To deal with these problems, a stochastic approximation procedure is used, leading to the Stochastic Approximation EM algorithm (SAEM). Exploiting recent results on stochastic approximation algorithm, the convergence of the SAEM algorithm to a maxima of the incomplete-likelihood is demonstrated for a general class of complete-data likelihood functions.

1 Introduction

Dans de très nombreuses situations, en statistique (mélange de populations, modèles linéaires, censures), en traitement du signal (déconvolution, séparation de sources) ou en traitement de l'image (segmentation, classification), 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 considérons ici que le phénomène qui fait intervenir ces deux séries peut se modéliser en faisant intervenir un paramètre θ qui prend ses valeurs dans un certain ensemble Θ, ouvert de $\mathbb{R}^p$. Nous nous intéressons ici au problème de l'identification de ce modèle. Des versions stochastiques de l'algorithme EM [5] ont été proposées par différents auteurs dans différents contextes [2, 3, 8, 10, 11, 12]. Des simulations et des applications sur des données réelles donnent souvent de très bons résultats, mais peu de résultats théoriques généraux existent sur ces algorithmes.

Nous proposons ici un nouvel algorithme pour lequel des résultats précis de convergence sont démontrés pour une famille de modèles très généraux [9].

Soit y les données observées. Elles sont la réalisation d'une variable aléatoire $Y \in \mathbb{R}^q$, distribuée suivant une loi ayant une densité $g(y;\theta)$. L'objectif est de calculer l'estimateur du maximum de vraisemblance de θ, c'est-à-dire de calculer $\hat{\theta}$ qui maximise $g(y;\theta)$ pour y donné. On posera $l(y;\theta) \stackrel{\text{def}}{=} \log g(y;\theta)$.

On suppose que l'on peut considérer Y comme la projection d'une variable $X \in \mathbb{R}^l$, $l \geq q$, qui est liée plus naturellement au paramètre θ que les observations. En général, $X = (Y, Z)$ où Z représente les données manquantes. On notera $f(x;\theta)$ la vraisemblance complète. Dans de très nombreux problèmes, la maximisation de $g(y;\theta)$ est compliquée, contrairement à celle de $f(x;\theta)$ (mélange de populations, régression avec censure, convolution, séparation de sources ...).

L'algorithme EM maximise $g(y;\theta)$ en θ en maximisant itérativement les espérances conditionnelles de $\log f(X;\theta)$ [5]. 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 $\log f(X;\theta)$ conditionnellement à $Y = y$ et avec la valeur courante θ_k :

$$Q(\theta|\theta_k) = E^y_{\theta_k} \log f(X;\theta) \tag{1}$$

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

Sous quelques hypothèses, on montre que la séquence $\{\theta_k\}$ converge vers un point $\theta^\star$, où $\theta^\star$ est un point stationnaire de g: $\mathcal{D}_\theta g(y;\theta^\star) = 0$. Ce point peut être un maximum local ou un point-selle, suivant le point initial choisi.

(Dans certaines situations, une vraisemblance pénalisée peut être utilisée en définissant $Q_\lambda(\theta|\theta_k) = Q(\theta|\theta_k) - \lambda J(\theta)$ où $J(\theta)$ est une fonction à définir.)

L'algorithme SAEM: Lorsque l'étape E ne peut être réalisée, nous proposons de la substituer par une simulation de type Monte-Carlo suivie d'une approximation stochastique. Conditionnellement aux observations y, les données complètes suivent une loi ayant une densité $k(x|y;\theta_k) = f(x;\theta)/g(y;\theta)$. A l'itération $k+1$, les étapes de l'algorithme sont alors les suivantes :

- *Simulation*: on génère $m(k)$ réalisations indépendantes $x_k(j)$ $(j = 1, \cdots, m(k))$ des données complètes sous la loi *a posteriori* $k(x|y;\theta_k)$.
- *Approximation Stochastique* : on réactualise l'approximation de $Q(\theta|\theta_k)$:

$$\hat{Q}(\theta|\theta_k) = \hat{Q}(\theta|\theta_{k-1}) + \\ + \gamma_k\Big(\frac{1}{m(k)} \sum_{j=1}^{m(k)} \log f(x_k;\theta) - \hat{Q}(\theta|\theta_{k-1})\Big) \tag{2}$$

où (γ_k) est une séquence positive de pas décroissants.

- *Maximisation*: on calcule $\theta_{k+1} \in \Theta$ qui maximise $\hat{Q}(\theta|\theta_k)$.

Sous les hypothèses suivantes, des résultats de convergence de la suite $\{\theta_k\}$ peuvent être obtenus :

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

$$\forall\theta \in \Theta \quad f(x;\theta) = f(y,z;\theta) == \\ \exp\Big\{-\psi(y;\theta) + \Big\langle \tilde{S}(y,z), \phi(y;\theta)\Big\rangle + r(y,z)\Big\} \tag{3}$$

où $\tilde{S}$ est une fonction mesurable à valeur dans un convexe fermé $\mathcal{S} \subset \mathbb{R}^m$, $\psi : \mathcal{Y} \times \Theta \to \mathbb{R}^m$ et $\phi : \mathcal{Y} \times \Theta \to \mathbb{R}$, sont des fonctions deux fois continument différentiables en θ pour tout $y \in \mathcal{Y}$.

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

(H3) La séquence $\{\gamma_k\}$ est telle que : $0 \leq \gamma_k \leq 1$, $\sum_{k=0}^\infty \gamma_k = \infty$ et $\sum_{k=0}^\infty \gamma_k^2 < \infty$. Le nombre de simulations $m(k)$ tend vers une constante: $\lim_{k\to\infty} m(k) = m$.

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

$$\forall\theta \in \Theta, \quad \forall s \in \dot{\mathcal{S}}, \quad L(s;\tilde{\theta}(s)) \geq L(s;\theta) \tag{4}$$

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

$$L(s;\theta) = -\psi(y;\theta) + \langle s, \phi(y;\theta)\rangle . \tag{5}$$

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

(On pose $L(s;\theta) = -\psi(y;\theta) + \langle s, \phi(y;\theta)\rangle - \lambda J(\theta)$ pour la maximisation d'une vraisemblance pénalisée.)

Sous **(H1)-(H4)**, l'étape E de l'algorithme consiste à une approximation de la statistique exhaustive du modèle complet :

$$S_{k+1} = S_k + \gamma_k \left(\frac{1}{m(k)} \sum_{j=1}^{m(k)} \tilde{S}(X_k(j)) - S_k \right) \tag{6}$$

et l'étape M au calcul de l'estimateur du maximum de vraisemblance de θ :

$$\theta_{k+1} = \tilde{\theta}(S_{k+1}). \tag{7}$$

Finalement, puisque $X = (y, Z)$, l'étape de simulation consiste à simuler les données manquantes Z. A l'étape $k+1$, on a donc $X_k(j) = (y, Z_k(j))$. On fait l'hypothèse suivante :

(H5) Les variables aléatoires $Z_k(j)$ ($k \geq 0$, $1 \leq j \leq m(k)$) sont indépendantes ; pour tout $k \geq 0$, les variables $Z_k(j)$ ($1 \leq j \leq m(k)$) sont distribuées sous la loi $k(z|y;\theta_k)$.

En utilisant les résultats sur la convergence des algorithmes stochastiques, [4, 6, 7, 9], nous avons le théorème suivant :

Théorème 1 *Si les hypothèses* **(H1)-(H5)** *sont vérifiées et si la séquence* $\{S_k\}$ *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 (pénalisée)* g.

Sous quelques hypothèses supplémentaires, et en utilisant les résultats de [1], on montre que l'algorithme ne peut converger que vers les maxima (éventuellement locaux) de la vraisemblance :

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

(H7) La plus petite valeur propre de la matrice $E^y_\theta(\tilde{S}(X) - E^y_\theta \tilde{S}(X))(\tilde{S}(X) - E^y_\theta \tilde{S}(X))^t$ est minorée par une constante strictement positive pour tout θ appartenant à un compact $\mathcal{K} \subset \Theta$.

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

Théorème 2 *Si les hypothèses* **(H1) – (H8)** *sont vérifiées et si la séquence* $\{S_k\}$ *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 [4, 9]. En particulier, si l'on peut montrer qu'il existe une constante $A \geq 0$ telle que la condition $\|S_{k-1}\| > A$ implique $E^y_{\tilde{\theta}(S_{k-1})}(\|S_k\|) \leq \|S_{k-1}\|$ avec probabilité 1, alors, $\underline{\lim}\, \|S_k\| < \infty$ et la suite $\{S_k\}$ est $\mathcal{S}$-compacte.

EXEMPLE

Séparation de sources discrètes: Estimation des angles d'arrivée

$$Y = ZA + \varepsilon$$

$$Z : n \times m \; ; \; Y : n \times p$$

$$A = A(\alpha) : m \times p \quad \text{(matrice de mélange)}$$

$$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_m) \quad \text{angles d'arrivée}$$

$$\varepsilon \rightsquigarrow \mathcal{N}(0, \Gamma) \quad , \quad \Gamma_{kl} = \rho^{|k-l|}\sigma^2$$

Nombre de sources : $m = 2$
Nombre de récepteurs : $p = 4$
$\alpha = (10^\circ, 15^\circ)$
$n = 500$
Z : QAM4 : $Z_{ij} \in \{-1, +1, -i, +i\}$
Rapport Signal/Bruit = 10 dB

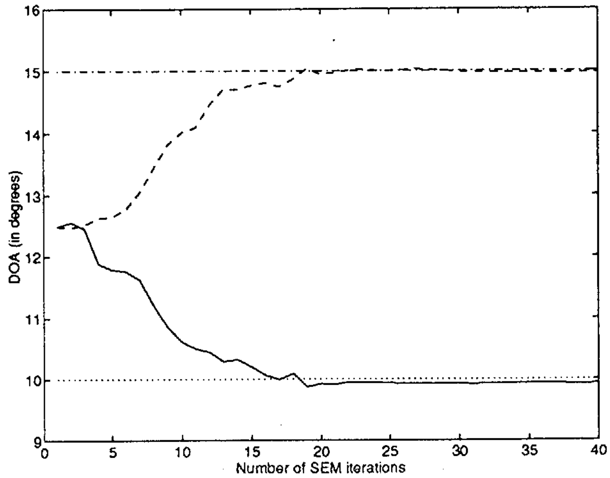


Figure 1 : Une trajectoire du SAEM

	MUSIC ordre 2	MUSIC ordre 4	SAEM
$\rho = 0$	9.78	10.00	10.000
	0.35	0.25	0.036
$\rho = 0.5$	8.45	9.99	10.000
	0.30	0.26	0.033
$\rho = 0.99$	7.98	10.00	10.001
	0.14	0.33	0.009

Estimation de $\theta_1 = 10^\circ$ avec MUSIC et avec SAEM

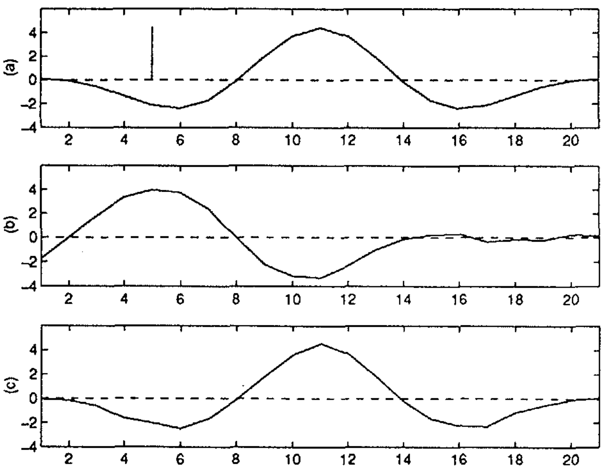
Une version recuit simulé de SAEM : Le théorème précédent nous assure que l'algorithme converge vers un maximum de la vraisemblance, mais pas nécessairement vers le maximum global. Une version *recuit simulé* de l'algorithme SAEM peut alors être envisagée.

L'algorithme de recuit simulé "classique" consisterait ici à rajouter dans (2) un bruit blanc gaussien dont la variance décroit lentement. Cette stratégie ne convient pas ici, puisque la nouvelle valeur de la statistique s_{k+1} n'appartient plus nécessairement au convexe $\mathcal{S}$ (ce qui peut conduire à estimer des variances négatives ...).

Considérons comme exemple le modèle de convolution $y = f * z + \sigma_\varepsilon \varepsilon$ 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 $y = f * z + (\sigma_\varepsilon + T_k)\varepsilon$ où la séquence T_k décroit 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.

On propose dans la figure suivante 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 signal/bruit vaut 10dB. 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 estime un filtre déphasé, alors que la version "recuit" donne une excellente estimation.



(a) Filtre original et initialisation (b) Estimation par SAEM sans recuit (c) Estimation par SAEM avec recuit

Références

[1] Brandiere, O., and Duflo, M. Les algorithmes stochastiques contournent-ils les pièges ? To appear in the Annales de l'Institut H. Poincaré, 1995.

[2] Celeux, G., and Diebolt, J. A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastics reports 41* (1992), 119–134.

[3] Chauveau, D. A stochastic algorithm for mixtures with censored data. *Preprint Université Marne la Vallée* (1994).

[4] Delyon, B. A deterministic approach to stochastic approximation. Submitted to IEEE Trans. On Autom. Control, 1995.

[5] Dempster, A., Laird, N., and Rubin, D. Maximum-likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B 39* (1977), 1–38.

[6] Duflo, M. Algorithmes stochastiques. Université de Marne-la-Vallée, 1994.

[7] Fort, J., and Pagès, G. Sur la convergence presque-sûre d'algorithmes stochastiques : le théorème de Kushner-Clark revisité. Preprint SAMOS Université Paris I, 1994.

[8] Lavielle, M. A stochastic procedure for parametric and non-parametric estimation in the case of incomplete data. *Signal Processing* (1995).

[9] Lavielle, M., and Moulines, E. On a stochastic approximation version of the EM algorithm. Tech. rep., Publication Université Paris-Sud, 1995.

[10] Tanner, M. *Tools for statistical inference: methods for exploration of posterior distributions and likelihood functions.* Springer-Verlag: Springer series in statistics, 1993.

[11] Wei, G., and Tanner, M. A Monte-Carlo implementation of the EM algorithm and the Poor's Man's data augmentation algorithm. *J. Amer. Stat. Assoc. 85* (1990), 699–704.

[12] Younes, L. Parametric inference for imperfectly observed Gibbsian fields. *Prob. Theory Rel. Fields 82* (1989), 625–645.