Available online at www.sciencedirect.com
ScienceDirect
C. R. Acad. Sci. Paris, Ser. I 346 (2008) 457–460
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Statistique

Sur la convergence de l'estimation conditionnelle itérative

Wojciech Pieczynski

Institut TELECOM ; TELECOM et Management SudParis ; Dépt. CITI ; CNRS UMR 5157, 9, rue Charles-Fourier, 91000 Evry, France

Reçu le 5 avril 2007 ; accepté après révision le 25 février 2008

Disponible sur Internet le 28 mars 2008

Présenté par Paul Deheuvels

Résumé

L'estimation conditionnelle itérative (ECI) est une méthode d'estimation itérative des paramètres dans le cas des données incomplètes. Sa mise en œuvre demande des hypothèses relativement faibles et peut être effectuée dans des situations relativement complexes, comme les modèles de Markov triplets. L'objet de cette Note est d'énoncer un théorème général de convergence de l'ECI, et de montrer son applicabilité dans le problème de l'estimation des proportions dans un mélange de lois multi-variées. ***Pour citer cet article : W. Pieczynski, C. R. Acad. Sci. Paris, Ser. I 346 (2008).***
© 2008 Académie des sciences. Publié par Elsevier Masson SAS. Tous droits réservés.

Abstract

On convergence of the iterative conditional estimation. The iterative conditional estimation (ICE) is an iterative estimation method of the parameters in the case of incomplete data. Its use asks for relatively weak hypotheses and it can be performed in relatively complex situations, as in triplet Markov models. The aim of this Note is to express a general theorem of convergence of ICE, and to show its applicability in the problem of the estimation of the proportions in a mixture of multivariate distributions. ***To cite this article: W. Pieczynski, C. R. Acad. Sci. Paris, Ser. I 346 (2008).***
© 2008 Académie des sciences. Publié par Elsevier Masson SAS. Tous droits réservés.

1. Introduction

Soit $Z = (X, Y)$ un couple de variables aléatoires, avec X observable et Y cachée, dont la loi $p(z|\theta)$ dépend d'un paramètre $\theta \in R^s$, avec $s \in \mathbb{N}^*$. L' « estimation conditionnelle itérative » (ECI) est une méthode itérative d'estimation de θ à partir de Y pouvant être mise en œuvre dès que : (i) l'on dispose d'un estimateur $\hat{\theta}(X, Y)$ de θ à partir des données complètes ; et (ii) l'on dispose, pour tout θ, d'une méthode de simulation de X selon la loi conditionnelle $p(x|y, \theta)$. L'ECI définit une suite $\theta^0, \theta^1, \ldots, \theta^q, \ldots$ par la procédure itérative suivante. L'initialisation θ^0étant donnée, θ^{q+1} est défini à partir de θ^q et Y par

$$\tilde{\theta}^{q+1}(Y, \theta^q) = E_{\theta^q}\big[\hat{\theta}(X, Y)|Y\big). \tag{1}$$

Proposée dans [9], l'ECI a été appliquée avec succès dans divers problèmes de traitements non supervisés des signaux ou des images ; citons [1,4–6] parmi les références récentes. Cependant, l'ECI souffrait du manque de résultats

Adresse e-mail : Wojciech.Pieczynski@int-edu.eu.

1631-073X/$ – see front matter © 2008 Académie des sciences. Publié par Elsevier Masson SAS. Tous droits réservés.
doi:10.1016/j.crma.2008.02.023

théoriques concernant son comportement asymptotique. L'objet de cette Note est de montrer un théorème général concernant ce comportement et d'illustrer son intérêt en considérant l'estimation de la loi a priori d'un mélange fini de densités multi-variées.

2. Convergence de l'ECI

Soient $X^n = (X_1, X_2, \ldots, X_n)$, $Y^n = (Y_1, Y_2, \ldots, Y_n)$ deux suites aléatoires, chaque $Z_i = (X_i, Y_i)$ étant à valeurs dans $\mathrm{X} \times \mathrm{Y}$. Les variables $Z_i = (X_i, Y_i)$ sont i.i.d. et leur loi $p(z_i|\theta)$ dépend d'un paramètre $\theta \in \Theta$, avec $\Theta \subset \mathbb{R}^s$, et $s \in \mathbb{N}^*$. On suppose l'existence d'un estimateur sans biais $\hat{\theta} : \mathrm{X} \times \mathrm{Y} \to \Theta$ et d'une matrice M tels que

$$\forall \theta, \theta^q \in \Theta, \quad \mathrm{Cov}_\theta\big[\tilde{\theta}^{q+1}(Y_1, \theta^q)\big] \leqslant M, \tag{2}$$

avec $\mathrm{Cov}_\theta[\tilde{\theta}^{q+1}(Y_1, \theta^q)]$ matrice de variance–covariance. Posons

$$\hat{\theta}_n(X^n, Y^n) = \frac{\hat{\theta}(X_1, Y_1) + \cdots + \hat{\theta}(X_n, Y_n)}{n}, \tag{3}$$

et appliquons (1) à l'estimateur $\hat{\theta}_n(X^n, Y^n)$:

$$\tilde{\theta}_n^{q+1}(Y^n) = \frac{E_{\theta^q}[\hat{\theta}(X_1, Y_1)|Y_1)] + \cdots + E_{\theta^q}[\hat{\theta}(X_n, Y_n)|Y_n)]}{n}. \tag{4}$$

En partant de $\theta^0 \in \Theta$, nous avons donc une famille de variables $\tilde{\theta}_n^q(Y^n)$ indicée par deux naturels (n, q), où n est la taille de l'échantillon et q est le nombre des itérations effectuées à partir de Y^n (précisons que dans notre démarche le même $\theta^0 \in \Theta$ sert à toutes les initialisations, quel que soit n).

Notre objectif est d'étudier le comportement de $\tilde{\theta}_n^q(Y^n)$ lorsque n et q tend tendent vers l'infini.

Définition. Pour $\theta^0 \in \Theta$ et $f : \mathbb{N} \to \mathbb{N}$ croissante, la suite $\tilde{\theta}_1^{f(1)}(Y^1), \tilde{\theta}_2^{f(2)}(Y^2), \ldots, \tilde{\theta}_n^{f(n)}(Y^n), \ldots$ sera dite « estimateur ECI relatif à (θ^0, f) ».

Pour tous θ, θ' dans Θ, posons $\varphi(\theta, \theta') = E_\theta[E_{\theta'}[\hat{\theta}(X_1, Y_1)|Y_1]]$, et supposons que φ est uniformément continue :

$$\forall \varepsilon > 0, \exists \alpha > 0, \forall \theta^1, \theta^2, \theta^3, \theta^4 \in \Theta, \quad \big\|(\theta^1, \theta^2) - (\theta^3, \theta^4)\big\| < \alpha \Rightarrow \big\|\varphi(\theta^1, \theta^2) - \varphi(\theta^3, \theta^4)\big\| < \varepsilon. \tag{5}$$

Nous pouvons énoncer le résultat suivant :

Théorème. *Pour tous $\theta^0, \theta \in \Theta$, soit $(\theta_\theta^q)_{q \in \mathbb{N}}$ la suite définie récursivement par θ^0 et $\theta_\theta^{q+1} = \varphi(\theta, \theta_\theta^q)$. Supposons que la suite des fonctions $\theta \to \theta_\theta^q$ converge uniformément vers l'identité* :

$$\forall \varepsilon > 0, \exists n_0 \in N \text{ tel que } \forall n \geqslant n_0, \ \forall \theta \in \Theta, \quad \|\theta_\theta^n - \theta\| < \varepsilon. \tag{6}$$

Alors il existe $f : \mathbb{N} \to \mathbb{N}$ croissante telle que l'estimateur ECI relatif à (θ^0, f) converges en probabilité uniformément sur Θ :

$$\forall \varepsilon > 0, \ \exists n_0 \in N \text{ tel que } \forall n \geqslant n_0, \ \forall \theta \in \Theta, \quad P_\theta\big[\big\|\tilde{\theta}_n^{f(n)}(Y^n) - \theta\big\| \geqslant \varepsilon\big] \leqslant \varepsilon. \tag{7}$$

Démonstration. On note $B(\theta, \rho)$ la boule ouverte de centre θ et de rayon $\rho > 0$. Soient $\theta^0, \theta \in \Theta$ et $(\theta_\theta^q)_{q \in N}$ la suite déterministe correspondante. Montrons la propriété (P) suivante, qui permettra de construire une suite $f : \mathbb{N} \to \mathbb{N}$ telle que $(\tilde{\theta}_n^{f(n)})_{n \in N}$ vérifie (7).

(P) Pour tout naturel $m \geqslant 1$, il existe deux naturels $k(m)$ et $n(m)$ tels que si $n \geqslant n(m)$, alors $P_\theta[\|\tilde{\theta}_n^{k(m)} - \theta\| \leqslant \frac{1}{m}] \geqslant 1 - 1/m$ pour tout $\theta \in \Theta$.

Soit $m \geqslant 1$. Selon (6), il existe un naturel $k(m)$, tel que $\|\theta_\theta^q - \theta\| < 1/2m$ pour $q \geqslant k(m)$ et tout $\theta \in \Theta$. Il en résulte que si $\theta' \in B(\theta_\theta^{k(m)}, 1/2m)$, alors $\|\theta' - \theta\| \leqslant 1/m$. Considérons les boules $B(\theta_\theta^1, \rho_1), B(\theta_\theta^2, \rho_2), \ldots, B(\theta_\theta^{k(m)}, \rho_{k(m)})$ dont les rayons, indépendants de θ en vertu de (5), sont définis par les récursions rétrogrades $\rho_{k(m)} = 1/2m$, $\rho_{k(m)-1}$ tel

que $\theta' \in B(\theta_\theta^{k(m)-1}, \rho_{k(m)-1})$ implique $\varphi(\theta,\theta') \in B(\theta_\theta^{k(m)}, \rho_{k(m)}/2), \ldots, \rho_{q-1}$ tel que $\theta' \in B(\theta_\theta^{q-1}, \rho_{q-1})$ implique $\varphi(\theta,\theta') \in B(\theta_\theta^{q}, \rho_q/2), \ldots, \rho_1$ tel que $\theta' \in B(\theta_\theta^{1}, \rho_1)$ implique $\varphi(\theta,\theta') \in B(\theta_\theta^{2}, \rho_2/2)$. Il est alors possible, en vertu de l'inégalité de Bienaymé–Tchebychev et de (2), de déterminer des naturels $n_1, n_2, \ldots, n_{k(m)}$ tels que pour tout $q = 1, \ldots, k(m)$ et tout $\theta \in \Theta$, lorsque $\theta' \in B(\theta^{q-1}, \rho_{q-1})$ la probabilité P_θ pour que

$$E_{\theta'}\big[\hat{\theta}_{n_q}|Y^{n_q}\big] = E_{\theta'}\big[\hat{\theta}_{n_q}(X^{n_q}, Y^{n_q})|Y^{n_q}\big]$$

soit dans $B(\theta^q, \rho_q)$ est supérieure ou égale à $1 - 1/m2^{q+1}$. En effet, d'une part l'événement $E_{\theta'}[\hat{\theta}_n|Y^n] \in B(\theta_\theta^q, \rho_q)$ contient l'événement $\|E_{\theta'}[\hat{\theta}_n|Y^n] - \varphi(\theta,\theta')\| < \rho_q/2$ et, d'autre part, la probabilité de ce dernier est supérieure ou égale, en vertu de l'inégalité de Bienaymé–Tchebychev et de (2), à $1 - 2M/n\rho_q$. Il est donc possible de prendre n_q tel que $1 - 2M/n\rho_q \geqslant 1 - 1/m2^{q+1}$ pour $n \geqslant n_q$. Considérons $n(m)$ le plus grand des $n_1, n_2, \ldots, n_{k(m)}$ et introduisons, pour $q = 1, \ldots, k(m)$, les variables $U_1^{n(m)}, \ldots, U_{k(m)}^{n(m)}$ prenant leurs valeurs dans $\{0, 1\}$ et définies par $U_q^{n(m)} = 1$ si $\tilde{\theta}_{n(m)}^q(Y^{n(m)}) \in B(\theta_\theta^q, \rho_q)$, et $U_q^{n(m)} = 0$ sinon. On peut alors affirmer, en reformulant ce qui précède, que

$$P_\theta\big[U_1^{n(m)} = 1\big] \geqslant 1 - 1/m2^2 \quad \text{et} \quad P_\theta\big[U_q^{n(m)} = 1 | U_{q-1}^{n(m)} = 1\big] \geqslant 1 - 1/m2^{q+1}$$

pour tout $q = 2, \ldots, k(m)$. La suite $U_1^{n(m)}, \ldots, U_{k(m)}^{n(m)}$ étant markovienne, cela implique

$$\begin{aligned} P_\theta\big[U_1^{n(m)} = 1, U_2^{n(m)} = 1, \ldots, U_{k(m)}^{n(m)} = 1\big] &\geqslant \left(1 - \frac{1}{m2}\right)\left(1 - \frac{1}{m2^2}\right)\cdots\left(1 - \frac{1}{m2^{k(m)+1}}\right) \\ &\geqslant 1 - \left(\frac{1}{m2} + \frac{1}{m2^2} + \cdots + \frac{1}{m2^{k(m)+1}}\right) \geqslant 1 - \frac{1}{m}. \end{aligned}$$

Finalement, on a $P_\theta[U_{k(m)}^{n(m)} = 1] \geqslant 1 - 1/m$, et cette inégalité est encore vraie en remplaçant $n(m)$ par tout $n \geqslant n(m)$, ce qui établit la propriété (P).

Considérons la suite $f : \mathbb{N} \to \mathbb{N}$ construite de la façon suivante. Pour $m = 1, 2, \ldots$ considérons deux suites, qui peuvent être supposées strictement croissantes, $n(1), n(2), \ldots$ et $k(1), k(2), \ldots$ données par la propriété (P). On pose alors : pour $1 \leqslant n \leqslant n(1)$, $f(n) = k(1)$, pour $n(1) < n \leqslant n(2)$, $f(n) = k(2), \ldots$, pour $n(m) < n \leqslant n(m+1)$, $f(n) = k(m), \ldots$. Montrons la convergence en probabilité de $\tilde{\theta}_n^{f(n)}$. Soit $\varepsilon > 0$. En vertu de (P), il existe $k(\varepsilon)$ et $n(\varepsilon)$ tels que si $n \geqslant n(\varepsilon)$, alors $P_\theta[\|\tilde{\theta}_n^{k(\varepsilon)} - \theta\| \leqslant \varepsilon] \geqslant 1 - \varepsilon$ pour tout θ(on prend pour $k(\varepsilon)$ le plus petit k tel que $\frac{1}{k} < \varepsilon$). Sachant que

$$P_\theta\left[\|\tilde{\theta}_n^{k(m+m')} - \theta\| \leqslant \frac{1}{m+m'}\right] \geqslant 1 - \frac{1}{m+m'}$$

pour tout $n(m+m') < n \leqslant n(m+m'+1)$, nous pouvons affirmer que si $n \geqslant n(\varepsilon)$ et $n(m+m') < n \leqslant n(m+m'+1)$, alors

$$P_\theta\left[\|\tilde{\theta}_n^{f(n)} - \theta\| \leqslant \frac{1}{m+m'}\right] \geqslant 1 - \frac{1}{m+m'},$$

pour tout θ. Il en résulte que pour tout $n \geqslant n(\varepsilon)$ on a $P_\theta[\|\tilde{\theta}_n^{f(n)} - \theta\| \leqslant \varepsilon] \geqslant 1 - \varepsilon$ pour tout θ, ce qui achève la démonstration. □

Exemple. Considérons un mélange $f^\theta = \theta_1 f_1 + \cdots + \theta_s f_s$ de densités de probabilités connues $f_1, \ldots, f_s$ sur $\mathrm{Y} = R^r$, avec $r \in \mathbb{N}^*$. Ainsi les variables $X_1, \ldots, X_n, \ldots$ sont à valeurs dans $\mathrm{X} = \{\omega_1, \ldots, \omega_s\}$. On souhaite estimer $\theta = (\theta_1, \ldots, \theta_{s-1})$, sachant que $\theta_s = 1 - (\theta_1 + \cdots + \theta_{s-1})$. On suppose qu'il existe $a > 0$ tel que chaque θ_i est dans $[a, 1-a]$. En prenant pour l'estimateur à partir des données complètes $\hat{\theta}_i(X_1, Y_1) = 1_{[X_1 = \omega_i]}$, montrons que l'hypothèse (6) du théorème est vérifiée. Pour simplifier, on omettra, dans les notations, la dépendance des θ^q de θ. En notant $f^{\theta^q} = \theta_1^q f_1 + \cdots + \theta_s^q f_s$, nous avons $\theta^{q+1} = \theta^q + A(\theta^q)(\theta - \theta^q)$, avec $A(\theta^q) = [a_{ij}(\theta^q)]$ matrice donnée par

$$a_{ij}(\theta^q) = \int_{R^r} \big(\theta_i^q f_i(f_j - f_s)\big)/f^{\theta^q}\,\mathrm{d}y.$$

Il en résulte que $\theta - \theta^{q+1} = [\mathrm{Id} - A(\theta^q)](\theta - \theta^q)$, d'où

$$\theta - \theta^{q+1} = \big[\mathrm{Id} - A(\theta^q)\big]\big[\mathrm{Id} - A\big(\theta^{q-1}\big)\big] \cdots \big[\mathrm{Id} - A(\theta^0)\big](\theta - \theta^0).$$

Montrons que $|a_{ij}(\theta^q)| \leqslant 1 - \varepsilon_{ij}$, avec $\varepsilon_{ij} > 0$. En posant $p_1(\omega_i|y) = (\theta_i^q f_i(y))/f^{\theta^q}$, on constate que $p^{(1)}(\omega_i) = \int_{R^r} p_1(\omega_i|y) f_j(y)\,\mathrm{d}y$ et $p^{(2)}(\omega_i) = \int_{R^r} p_1(\omega_i|y) f_s(y)\,\mathrm{d}y$ sont des probabilités non nulles, et que $a_{ij}(\theta_i^q) = p^{(1)}(\omega_i) - p^{(2)}(\omega_i)$, ce qui implique $|a_{ij}(\theta^q)| \leqslant 1$. Par ailleurs, la fonction $\theta_i^q \to a_{ij}(\theta_i^q)$ est continue et strictement inférieure à 1 sur le compact $[0, 1]$; elle est donc majorée par $1 - \varepsilon_{ij}$, avec $\varepsilon_{ij} > 0$. En prenant $\varepsilon = \inf_{i,j} \varepsilon_{ij}$ et $\|A(\theta^q)\| = \sup_{i,j} |a_{ij}(\theta^q)|$, on a $\|A(\theta^q)\| \leqslant 1 - \varepsilon$. Pour finir, on montre que les fonctions $\theta_i^q \to a_{ii}(\theta_i^q)$ sont strictement positives sur $[a, 1 - a]$ (elles sont concaves sur $[0, 1]$ et $a_{ii}(0) = a_{ii}(1) = 0$), ce qui implique $\|\mathrm{Id} - A(\theta^q)\| \leqslant 1 - \varepsilon$. Finalement $\|\theta - \theta^{q+1}\| \leqslant (1 - \varepsilon)^{q+1}\|\theta - \theta^0\|$, d'où $\theta^{q+1} \xrightarrow[q\to+\infty]{} \theta$. Les fonctions $\theta \to \theta_\theta^q$ étant continues et définies sur un compact, on a la convergence uniforme, ce qui montre l'hypothèse (6) du théorème. Notons que (2) est également vérifiée grâce à la compacité de $\Theta \times \Theta$ et la continuité de l'application $(\theta, \theta^q) \to \mathrm{Cov}_\theta[\tilde{\theta}^{q+1}(Y_1, \theta^q)]$.

Remarque. La méthode ECI est à rapprocher avec la méthode itérative « Expectation-Maximization » (EM, [7]), utilisant le principe $\theta^{q+1} = \arg\max_\theta E_{\theta^q}[\mathrm{Log}(p_\theta(X, Y))|Y = y]$ et dont l'intérêt est fondé sur l'optimalité asymptotique du Maximum de Vraisemblance. Pouvant donner d'excellents résultats, EM fonctionne sous des conditions plus restrictives que ECI ; en effet, l'existence du maximum de vraisemblance est requise. Par ailleurs, les calculs des itérations se heurtent souvent à des difficultés, qui ne peuvent être surmontées qu'au prix des approximations, généralement stochastiques, dont la convergence est difficile à étudier [2,7,10]. Notons que dans certains cas les deux méthodes donnent des estimateurs identiques [3]. Par ailleurs, dans le cas de mélanges Gaussiens les deux méthodes peuvent donner des résultats très proches [8]. Le résultat principal de la théorie de l'algorithme EM est la croissance, avec q, de la vraisemblance $p_{\theta^q}(y)$. Par ailleurs, la suite $(\theta^q)_{q\in N}$ converge, sous certaines conditions, vers une valeur des paramètres assurant un maximum local de la vraisemblance [2,7].

Références

[1] D. Benboudjema, W. Pieczynski, Unsupervised statistical segmentation of non stationary images using triplet Markov fields, IEEE Trans. Pattern Analysis and Machine Intelligence 29 (8) (2007) 1367–1378.

[2] O. Cappé, E. Moulines, T. Ryden, Inference in Hidden Markov Models, Series in Statistics, Springer, 2005.

[3] J.-P. Delmas, An equivalence of the EM and ICE algorithm for exponential family, IEEE Trans. Signal Processing 45 (10) (1997) 2613–2615.

[4] S. Derrode, G. Mercier, Multiscale oil slick segmentation from SAR image using a vector HMC model, Pattern Recognition 40 (3) (2007) 1135–1147.

[5] F. Destrempes, M. Mignotte, J.-F. Anger, Localization of shapes using statistical models and stochastic optimization, IEEE Trans. Pattern Analysis and Machine Intelligence 29 (9) (2007) 1603–1615.

[6] P. Lanchantin, J. Lapuyade-Lahorgue, W. Pieczynski, Unsupervised segmentation of triplet Markov chains hidden with long-memory noise, Signal Processing 88 (5) (2008) 1134–1151.

[7] G.J. McLachlan, T. Krishnan, The EM Algorithm and Extension, Wiley, 1997.

[8] A. Peng, W. Pieczynski, Adaptive mixture estimation and unsupervised local Bayesian image segmentation, Graphical Models and Image Processing 57 (5) (1995) 389–399.

[9] W. Pieczynski, Statistical image segmentation, Machine Graphics and Vision 1 (1/2) (1992) 261–268.

[10] D. Pierre-Loti-Viaud, Random perturbations of recursive sequences with an application to an epidemic model, J. Appl. Probab. 32 (1995) 559–578.