Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Statistics & Probability Letters 76 (2006) 2001–2006
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

Almost sure convergence of Titterington's recursive estimator for mixture models

Shaojun Wang[a,*,1], Yunxin Zhao[b]

[a]Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada T6G 2E8
[b]Department of Computer Science, University of Missouri at Columbia, Columbia, MO 65211, USA

Received 28 June 2005; accepted 23 May 2006
Available online 10 July 2006

Abstract

Titterington proposed a recursive parameter estimation algorithm for finite mixture models. However, due to the well known problem of singularities and multiple maximum, minimum and saddle points that are possible on the likelihood surfaces, convergence analysis has seldom been made in the past years. In this paper, under mild conditions, we show the global convergence of Titterington's recursive estimator and its MAP variant for mixture models of full regular exponential family.

Keywords: Recursive estimation; Incomplete data; Mixture model; Regular exponential family; Almost sure convergence; Stochastic approximation

1. Introduction

Recursive estimation algorithms have a wide range of applications in on-line system identification (cf. Weinstein et al., 1990), adaptive filtering, pattern recognition (cf. Fu, 1968), adaptive learning (cf. Wang and Zhao, 2001), sequential change detection (cf. Benveniste et al., 1990), and have attracted significant research interests due to the advantages of computational efficiency, reduced storage requirements, tracking time-varying parameters, as well as minimal processing delay. On the other hand, finite mixture models have provided a powerful framework to the statistical modeling of a wide variety of random phenomena. Fields in which mixture models have been successfully applied include astronomy, biology, genetics, medicine, economics, engineering, and marketing, among many others in the biological, physical, and social sciences (cf. McLachlan and Krishnan, 1997). As the result, the study of recursive estimation algorithms for mixture models is of a significant value.

*Corresponding author. Tel.: +1 780 492 0365; fax: +1 780 492 1071.
E-mail address: swang@cs.ualberta.ca (S. Wang).

[1]Work was done when the author was at Beckman Institute and Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.

0167-7152/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2006.05.017

In early 80s, Titterington (1984) proposed a recursive estimation algorithm for the general incomplete data problem which includes the finite mixture models as a special case. However, due to the well-known singularities and multiple maximum, minimum and saddle points that are possible on the likelihood surfaces for this problem, convergence analysis has seldom been made in the past years. Convergence analysis of Titterington's recursive estimator can be formulated in the framework of stochastic approximation. Results that are based on stochastic approximation often rely on restricted conditions (cf. Benveniste et al., 1990; Fabian, 1978; Nevelson and Khasminskii, 1976) on regression function. One example is the assumption of the unique root of regression function as in the Robbins–Monro (1951) procedure. Kushner and Clark (1978) gave more relaxed convergence conditions. The estimated parameter sequence however, needs to be bounded and returns infinitely often to a compact domain of attraction of one stationary point of the continuous vector field, which is difficult to verify in practice. The difficulty is due to the fact that when more than one stationary point exist (cf. Lazarev, 1992), this condition is stronger than boundedness and one needs to show that the parameter estimate sequence does not go infinitely often from one domain of attraction to another. Recently, Delyon (1996) gave a more general condition on the vector field under which the boundedness of parameter estimate will guarantee the convergence. We will use his result to show the global convergence of Titterington's recursive estimator and its MAP variant for mixture models of full regular exponential family which is useful for on-line recursive Bayesian learning.

The paper is organized as follows. In Section 2, we first derive Titterington's recursive estimator in a slightly different way as in (Titterington, 1984), and we then derive its MAP variant. In Section 3, we give the proof of global convergence properties for both recursive estimators and we draw conclusion in Section 4.

2. Titterington's recursive estimator

Assume that i.i.d. n-dimensional vector observations $y_1, \ldots, y_k$ are received sequentially where each observation has the underlying pdf $p(y|\lambda)$ with $\lambda \in \Theta \subset \mathscr{R}^d$, and for each observed data y_k, there is a missing or latent data x_k. Our objective is to derived a recursive estimation formula for λ in the general form of the recursive estimator (Benveniste et al., 1990):

$$\lambda^{(k+1)} = \lambda^{(k)} + \varepsilon_k h(y_{k+1}, \lambda^{(k)}), \tag{1}$$

where ε_k is a sequence of small positive gains which can be either fixed as a constant or decrease with the index k.

Denote $y^k = \{y_1, \ldots, y_k\}$, and $x^k = \{x_1, \ldots, x_k\}$. As in EM algorithm (cf. Dempster et al., 1977), define the auxiliary function $Q_{y^{k+1}}(\lambda, \lambda^{(k)}) = E[\log p(x^{k+1}, y^{k+1}|\lambda)|y^{k+1}; \lambda^{(k)}]$. It follows that maximizing $Q_{y^{k+1}}(\lambda, \lambda^{(k)})$ leads to improvements in $p(y^{k+1}|\lambda)$ (cf. Dempster et al., 1977; McLachlan and Krishnan, 1997). By maximizing the second-order Taylor series expansion of $(1/(k+1))Q_{y^{k+1}}(\lambda, \lambda^{(k)})$ with respect to λ and denoting the maximizing point by $\lambda^{(k+1)}$, we have

$$\lambda^{(k+1)} = \lambda^{(k)} + \left[-\frac{1}{k+1} \frac{\partial^2 Q_{y^{k+1}}(\lambda, \lambda^{(k)})}{\partial\lambda\partial\lambda^{\mathrm{T}}} \right]^{-1} \frac{1}{k+1} \frac{\partial Q_{y^{k+1}}(\lambda, \lambda^{(k)})}{\partial\lambda} \Bigg|_{\lambda=\lambda^{(k)}}.$$

If we replace $-(1/(k+1))(\partial^2 Q_{y^{k+1}}(\lambda, \lambda^{(k)})/\partial\lambda\partial\lambda^{\mathrm{T}})$ by its expectation, i.e., the complete data Fisher information matrix $I_{CF}(\lambda^{(k)}) = E[-\partial^2 \log p(x, y|\lambda)/\partial\lambda\partial\lambda^{\mathrm{T}}]|_{\lambda=\lambda^{(k)}}$, then we have

$$\lambda^{(k+1)} = \lambda^{(k)} + \frac{1}{k+1}[I_{CF}(\lambda^{(k)})]^{-1} \frac{\partial Q_{y^{k+1}}(\lambda, \lambda^{(k)})}{\partial\lambda} \Bigg|_{\lambda=\lambda^{(k)}}, \tag{2}$$

which corresponds to one step of the batch scoring EM algorithm (McLachlan and Krishnan, 1997), initialized at $\lambda^{(k)}$.

The batch algorithm of Eq. (2) is next converted into a recursive estimation algorithm. Let $\ell_{y_{k+1}}(\lambda, \lambda^{(k)}) = Q_{y^{k+1}}(\lambda, \lambda^{(k)}) - Q_{y^k}(\lambda, \lambda^{(k)}) = Q_{y_{k+1}}(\lambda, \lambda^{(k)})$. Then, $\partial Q_{y^{k+1}}(\lambda, \lambda^{(k)})/\partial\lambda = (\partial Q_{y^k}(\lambda, \lambda^{(k)})/\partial\lambda) + (\partial \ell_{y_{k+1}}(\lambda, \lambda^{(k)})/\partial\lambda)$. Assuming that $\lambda^{(k)}$ maximizes $Q_{y^k}(\lambda, \lambda^{(k)})$ so that $\partial Q_{y^k}(\lambda, \lambda^{(k)})/\partial\lambda|_{\lambda=\lambda^{(k)}} = 0$, then we have

$$\left.\frac{\partial Q_{y^{k+1}}(\lambda,\lambda^{(k)})}{\partial\lambda}\right|_{\lambda=\lambda^{(k)}}=\left.\frac{\partial \ell_{y_{k+1}}(\lambda,\lambda^{(k)})}{\partial\lambda}\right|_{\lambda=\lambda^{(k)}}.$$

Furthermore, the relation holds

$$\left.\frac{\partial}{\partial\lambda}Q_{y_{k+1}}(\lambda,\lambda^{(k)})\right|_{\lambda=\lambda^{(k)}}=\left.\frac{\partial}{\partial\lambda}E[\log p(x_{k+1},y_{k+1}|\lambda)|y_{k+1},\lambda^{(k)}]\right|_{\lambda=\lambda^{(k)}}=\left.\frac{\partial\log p(y_{k+1}|\lambda)}{\partial\lambda}\right|_{\lambda=\lambda^{(k)}}.$$

As the result, we obtain a recursive estimation formula as

$$\lambda^{(k+1)}=\lambda^{(k)}+\frac{1}{k+1}[I_{CF}(\lambda^{(k)})]^{-1}\left.\frac{\partial\log p(y_{k+1}|\lambda)}{\partial\lambda}\right|_{\lambda=\lambda^{(k)}}. \tag{3}$$

We call Eq. (3) as the EM type MLE recursive estimator and it is exactly Eq. (9) in Titterington (1984).

Now we consider the MAP variant of Eq. (3). Instead of assuming λ as fixed, we assume λ is generated by a prior pdf $p(\lambda|\phi)$ with a hyperparameter ϕ. Our objective is to derive a recursive formula for Bayesian learning of λ under the criterion of maximum a posteriori estimation.

Applying Bayes theorem, we obtain a recursive expression for the a posteriori pdf of λ, given y^{k+1}, as

$$p(\lambda|y^{k+1})=\frac{p(y_{k+1}|\lambda)p(\lambda|y^{k})}{\int p(y_{k+1}|\lambda)p(\lambda|y^{k})\,\mathrm{d}\lambda}. \tag{4}$$

Successive computation of Eq. (4) for $k=1,2,\ldots,$ introduces an ever-expanding combination of the previously obtained posterior pdfs and thus quickly leads to a combinatorial explosion of product terms. To overcome this difficulty, on-line quasi-Bayes learning (cf. Huo and Lee, 1997; Smith and Makov, 1978) first approximates the successive posterior distributions by the "closest" tractable distribution within a given class $\mathscr{P}$, under the criterion that both distributions have the same mode, and the EM algorithm is next applied to estimate the hyperparameters ϕ of the approximate posterior distribution and model parameters λ are incrementally updated. Empirical evidence showed that the quasi-Bayes algorithm in general converges to a good solution and it has a similar behavior with the batch MAP algorithm (cf. Huo and Lee, 1997).

Here we propose a new approach for recursive Bayesian learning. Define the auxiliary function of log posterior likelihood as $R_{y^{k+1}}(\lambda,\lambda^{(k)})=Q_{y^{k+1}}(\lambda,\lambda^{(k)})+\log p(\lambda|\phi)$. It follows that maximizing $R_{y^{k+1}}(\lambda,\lambda^{(k)})$ leads to improvements in $p(\lambda|y^{k+1})$ (cf. Dempster et al., 1977; McLachlan and Krishnan, 1997). Maximizing the second-order Taylor series expansion of $(1/(k+1))R_{y^{k+1}}(\lambda,\lambda^{(k)})$ with respect to λ and denoting the maximizing point by $\lambda^{(k+1)}$, we have

$$\lambda^{(k+1)}=\lambda^{(k)}+\left[-\frac{1}{k+1}\frac{\partial^2 R_{y^{k+1}}(\lambda,\lambda^{(k)})}{\partial\lambda\partial\lambda^{\mathrm{T}}}\right]^{-1}\frac{1}{k+1}\left.\frac{\partial R_{y^{k+1}}(\lambda,\lambda^{(k)})}{\partial\lambda}\right|_{\lambda=\lambda^{(k)}}.$$

If we replace $-(1/(k+1))(\partial^2 R_{y^{k+1}}(\lambda,\lambda^{(k)})/\partial\lambda\partial\lambda^{\mathrm{T}})$ by its expectation, then we have

$$\lambda^{(k+1)}=\lambda^{(k)}+\frac{1}{k+1}\left[I_{CF}(\lambda^{(k)})+\frac{1}{k+1}I_{\mathrm{p}}(\lambda^{(k)}|\phi)\right]^{-1}\left.\frac{\partial R_{y^{k+1}}(\lambda,\lambda^{(k)})}{\partial\lambda}\right|_{\lambda=\lambda^{(k)}}, \tag{5}$$

where $I_{\mathrm{p}}(\lambda|\phi)$ is the prior information matrix, i.e., negative Hessian matrix of log $p(\lambda|\phi)$.

To convert Eq. (5) into a recursive estimator, we define $\ell_{y_{k+1}}(\lambda,\lambda^{(k)})=R_{y^{k+1}}(\lambda,\lambda^{(k)})-R_{y^{k}}(\lambda,\lambda^{(k)})=Q_{y_{k+1}}(\lambda,\lambda^{(k)})$ and by following the same argument as for the EM type MLE recursive estimator, we have

$$\lambda^{(k+1)}=\lambda^{(k)}+\frac{1}{k+1}\left[I_{CF}(\lambda^{(k)})+\frac{1}{k+1}I_{\mathrm{p}}(\lambda^{(k)}|\phi)\right]^{-1}\left.\frac{\partial\log p(y_{k+1}|\lambda)}{\partial\lambda}\right|_{\lambda=\lambda^{(k)}}. \tag{6}$$

We refer to Eq. (6) as the EM type MAP recursive estimator and $I_{CF}(\lambda^{(k)})+(1/(k+1))I_{\mathrm{p}}(\lambda^{(k)}|\phi)$ as the complete-data Bayesian Fisher information matrix. It is easily seen that as k becomes large, the effect of prior information diminishes.

3. Convergence results

Global convergence properties of Eqs. (3) and (6) are unknown. However, if we assume the i.i.d. observations are generated by a mixture of a full regular exponential family, then under mild conditions to be stated below, we obtain global convergence results. Before showing the result, we first give the definition of a full a regular exponential family and a lemma which is proved in (Delyon, 1996).

Definition (*Brown, 1987*). Let ν be a σ-finite measure on the Borel subsets of $\mathscr{R}^n$. Let $\mathcal{N} = \{\theta\colon \int e^{\theta \cdot y} \nu(dy) < \infty\}$. Let $\varphi(\theta) = \log(\int e^{\theta \cdot y} \nu(dy))$ and define $p_\theta(y) = \exp(\theta \cdot y - \varphi(\theta)), \theta \in \mathcal{N}$. Let $\Theta \subseteq \mathcal{N}$. The family of probability densities $\{p_\theta : \theta \in \Theta\}$ is called a n-dimensional standard exponential family. $\mathcal{N}$ is called the natural parameter space. The family is called full if $\Theta = \mathcal{N}$. It is called regular if $\mathcal{N}$ is open, i.e. if $\mathcal{N} = \mathcal{N}^0$ where $\mathcal{N}^0$ denotes the interior of $\mathcal{N}$. Common pdfs such as normal, exponential, Poisson et al. belong to this family.

Lemma (*Delyon, 1996*). *Rewrite Eq.* (1) *as*

$$\lambda^{(k+1)} = \lambda^{(k)} + \varepsilon_k f(\lambda^{(k)}) + \varepsilon_k \mathrm{e}_k, \tag{7}$$

where $\lambda^{(0)}$ is given, $f(\lambda) = E_{\lambda^0}[h(y, \lambda)]$ is a vector field on $\Theta \subset \mathscr{R}^d$, $\mathrm{e}_k = h(y_k, \lambda^{(k)}) - f(\lambda^{(k)})$ is a perturbation, and ε_k is a nonnegative scalar gain sequence. Assume f is a continuous vector field defined on an open set $\Theta \subset \mathscr{R}^d$, such that $\Lambda = \{\lambda\colon f(\lambda) = 0\}$ is a compact subset of Θ, and there exists a nonnegative C^1 function V such that

(1) *$V(\lambda^0, \lambda)$ tends to ∞ if $\lambda \to \partial\Theta$ or $\|\lambda\| \to \infty$.*
(2) *$\langle \nabla_\lambda V(\lambda^0, \lambda), f(\lambda) \rangle < 0$ for $\lambda \notin \Lambda$.*

Further assume that the algorithm is A-stable, that is $\lambda^{(k)}$ remains in a compact subset of Θ, $\lim_{m\to\infty} \sum_{k=1}^m \varepsilon_k \mathrm{e}_k$ exists, and $\lim_{k\to\infty} |\varepsilon_k| = 0$. Then the distance of $\lambda^{(k)}$ to the set Λ, $d(\lambda^{(k)}, \Lambda)$, converges to 0 a.s., and in particular, if Λ is a finite set, $\lambda^{(k)}$ converges to some point of Λ.

Theorem 1 (*Global convergence property of the recursive MLE estimator*). *For an i.i.d. sequence, $y_1, y_2, \ldots, y_k, \ldots$, we assume the underlying pdf to be a mixture of a full regular exponential family $p(y|\lambda) = \sum_{g=1}^{\mathcal{G}} \omega_g p(y|\theta_g)$, with $\lambda = (\theta_1^{\mathrm{T}}, \ldots, \theta_{\mathcal{G}}^{\mathrm{T}})^{\mathrm{T}}$, and we use λ^0 to denote the true parameter. Assume that $[I_{CF}(\lambda^{(k)})]^{-1} < \infty$, $\partial \log p(y|\lambda)/\partial\lambda < \infty$ and $\lambda^{(k)}$ does not tend to infinity. Then the sequence of recursive estimator* (3) *converges a.s. to the set of parameters*

$$\Lambda = \left\{\lambda : \frac{\partial}{\partial\lambda} E_{\lambda^0}[\log p(y|\lambda)] = 0\right\} \tag{8}$$

even if this set contains several or nonisolated stationary points, and they may be stable, unstable or saddle.

Proof. Define the Kullback–Leibler measure or relative entropy as the potential (Lyapunov) function:

$$V(\lambda^0, \lambda) = E_{\lambda^0}\left[\log \frac{p(y|\lambda^0)}{p(y|\lambda)}\right].$$

It is well known that $V(\lambda^0, \lambda) \geqslant 0$ with equality if and only if $\lambda = \lambda^0$ (Cover and Thomas, 1991). It is easy to verify that $V(\lambda^0, \lambda) \to \infty$ if $\lambda \to \partial\Theta$ or $\|\lambda\| \to \infty$, since $p(y|\lambda)$ is a mixture of a full regular exponential family.

Taking derivative of $V(\lambda^0, \lambda)$ with respect to λ, we have

$$\nabla_\lambda V(\lambda^0, \lambda) = -E_{\lambda^0}\left[\frac{\partial \log p(y|\lambda)}{\partial\lambda}\right].$$

By Eq. (3), we have the gradient field

$$f(\lambda) = I_{CF}^{-1}(\lambda) E_{\lambda^0}\left[\frac{\partial \log p(y|\lambda)}{\partial\lambda}\right]$$

and thus

$$\langle \nabla_\lambda V(\lambda^0,\lambda), f(\lambda)\rangle = -E_{\lambda^0}\left[\frac{\partial \log p(y|\lambda)}{\partial \lambda}\right]^{\mathrm{T}} I_{CF}^{-1}(\lambda) E_{\lambda^0}\left[\frac{\partial \log p(y|\lambda)}{\partial \lambda}\right] < 0 \quad \text{for } \lambda \notin \Lambda.$$

The inequality is due to the fact that the complete-data pdf belongs to a full regular exponential family, and thus $I_{CF}(\lambda)$ is positive definite.

Let $M_k = \sum_{i=1}^{k} \varepsilon_i \mathrm{e}_i$, where in this case $\mathrm{e}_i = [I_{CF}(\lambda^{(i)})]^{-1}[(\partial \log p(y_i|\lambda^{(i)})/\partial\lambda^{(i)}) - E_{\lambda^0}(\partial \log p(y_i|\lambda^{(i)})/\partial\lambda^{(i)})]$ and $\varepsilon_k = 1/(k+1)$. By assumption, $[I_{CF}(\lambda^{(k)})]^{-1} < \infty$ and $\partial \log p(y|\lambda)/\partial\lambda < \infty$, and thus $\mathrm{e}_i < \infty$. Define $\mathscr{F}^k := \sigma(y^1, \ldots, y^k)$. Then for $k \geqslant 1$, we have $E[M_k|\mathscr{F}^{k-1}] = E[M_{k-1}|\mathscr{F}^{k-1}] + E[\varepsilon_k \mathrm{e}_k|\mathscr{F}^{k-1}] = M_{k-1} + \varepsilon_k E[\mathrm{e}_k] = M_{k-1}$, and thus M_k is a martingale. Moreover, since $E[(M_i - M_{i-1})^2] = E[(\varepsilon_i \mathrm{e}_i)^2] < C\varepsilon_i^2$, and $\sum_{k=1}^{\infty}(\varepsilon_k)^2 < \infty$, by orthogonality of martingale increment, $E[M_k^2] = \sum_{i=1}^{k} E[(M_i - M_{i-1})^2] = \sum_{i=1}^{k} E[(\varepsilon_i \mathrm{e}_i)^2] < \infty$, and thus M_k is bounded in $\mathscr{L}^2$, i.e., e_k is an L_2-bounded Martingale increment. As the result, $\lim M_k$ exists a.s. (Durrett, 1996), and the Λ-stability is satisfied. Thus $\lambda^{(k)}$ converges a.s. to Λ as defined by (8). □

Remark. By the strong law of large number, we have $(1/(k+1))\sum_{i=1}^{k+1} \log p(y_i|\lambda) \to E_{\lambda^0}[\log p(y|\lambda)] = H(p(y|\lambda^0)) - V(\lambda^0, \lambda)$ a.s., where $H(p(y|\lambda^0)) = E_{\lambda^0}[\log p(y|\lambda^0)]$ is the entropy of the source. On the other hand, $(1/(k+1))Q_{y^{k+1}}(\lambda, \lambda') = (1/(k+1))\sum_{i=1}^{k+1} Q_{y_i}(\lambda, \lambda') \to E_{\lambda^0}[Q_y(\lambda, \lambda')] = E_{\lambda^0}[\log p(y|\lambda)] + E_{\lambda^0}[E_{\lambda'} \log p(x|y, \lambda)]$ a.s. Since for each iteration, we always let $\lambda = \lambda'$, which enforces $E_{\lambda^0}[E_{\lambda'} \log p(x|y, \lambda)] = 0$. So maximizing the normalized auxiliary function $(1/(k+1))Q_{y^{k+1}}(\lambda, \lambda')|_{\lambda=\lambda'}$ is asymptotically equivalent to maximizing the normalized likelihood function $(1/(k+1)) \log p(y^{k+1}|\lambda)$. This gives an explanation why Eq. (3) converges to the set given by Eq. (8).

Theorem 2 (*Global convergence property of the recursive MAP estimator*). *For an i.i.d. sequence, $y_1, y_2, \ldots, y_k, \ldots$, we assume that the underlying pdf to be a mixture of a full regular exponential family $p(y|\lambda) = \sum_{g=1}^{\mathscr{G}} \omega_g p(y|\theta_g)$, $\lambda = (\theta_1^{\mathrm{T}}, \ldots, \theta_{\mathscr{G}}^{\mathrm{T}})^{\mathrm{T}}$ and we use λ^0 to denote the true parameter. Assume that $[I_{CF}(\lambda^{(k)})]^{-1} < \infty$, $[I_{\mathrm{p}}(\lambda^{(k)}|\phi)]^{-1} < \infty$, $(\partial \log p(y|\lambda)/\partial\lambda) < \infty$ and $\lambda^{(k)}$ does not tend to infinity. Then the sequence of MAP recursive estimator* (6) *converges a.s. to the same set of parameters of the EM type MLE recursive estimator, which is given by*

$$\Lambda = \left\{\lambda : \frac{\partial}{\partial\lambda} E_{\lambda^0}[\log p(y|\lambda)] = 0\right\} \tag{9}$$

even if this set contains several or nonisolated stationary points, and they may be stable, unstable or saddle.

Proof. By assumption of $[I_{\mathrm{p}}(\lambda^{(k)}|\phi)]^{-1} < \infty$, we have $(1/(k+1))I_{\mathrm{p}}(\lambda^{(k)}) \to 0$ as $k \to \infty$, i.e., the effect of prior information diminishes. The rest proof is the same as in Theorem 1.

Similar to the remarks made in Theorem 1, since the effect of prior information diminishes as k goes to ∞, $(1/(k+1)) \log p(\lambda|y^{k+1}) \to E_{\lambda^0}[\log\, p(y|\lambda)]$ a.s. and $(1/(k+1))R_{y^{k+1}}(\lambda, \lambda')|_{\lambda=\lambda'} \to E_{\lambda^0}[\log p(y|\lambda)]$ a.s. This gives an explanation why Eq. (6) converges to the set given by Eq. (9). □

4. Conclusion

In this paper, under mild conditions, we show the global convergence of Titterington's recursive estimator and its MAP variant for mixture models of a full regular exponential family. For certain exponential family such as multinomial distribution, stochastic constraints need to be satisfied among the model parameters. In such a case, unlike batch EM algorithm, projection to the constrained parameter space is needed. We do not provide convergence result for this type of exponential distributions at this time. For mixture models with Markov regime, usually called hidden Markov models, observation sequences are correlated due to the underlying Markov chain, and $(\partial \log p(y_{k+1}|y^k, \lambda))/\partial\lambda$ has to be used in Eqs. (3) and (6) instead of $(\partial \log p(y_{k+1}|\lambda))/\partial\lambda$. As the result, all past data need to be stored, which excludes the algorithm as an on-line one. To overcome this difficulty, approximations are normally made but convergence properties remain unknown.

References

Benveniste, A., Metivier, M., Priouret, P., 1990. Adaptive Algorithms and Stochastic Approximations. Springer, Berlin.

Brown, L., 1987. Fundamentals of Statistical Exponential Families. Institute of Mathematical Statistics.

Cover, T., Thomas, J., 1991. Elements of Information Theory. Wiley, New York.

Delyon, B., 1996. General results on the convergence of stochastic algorithms. IEEE Trans. Automat. Control 41 (9), 1245–1255.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood estimation from incomplete data via the EM algorithm. J. R. Statist. Soc. Ser. B 39, 1–38.

Durrett, R., 1996. Probability: Theory and Examples. Wadsworth Publishing Company, Belmont, CA.

Fabian, V., 1978. On asymptotically efficient recursive estimation. Ann. Statist. 6 (4), 854–866.

Fu, K., 1968. Sequential Methods in Pattern Recognition and Machine Learning. Academic Press, New York.

Huo, Q., Lee, C., 1997. Online adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate. IEEE Trans. Speech Audio Process. 5 (2), 161–172.

Kushner, H., Clark, D.D., 1978. Stochastic Approximation Methods for Constrained and Unconstrained Systems. Springer, Berlin.

Lazarev, V., 1992. Convergence of stochastic approximation procedures in the case of a regression equation with several roots. Probl. Inform. Transmission 66–78.

McLachlan, G., Krishnan, T., 1997. The EM Algorithm and Extensions. Wiley, New York.

Nevelson, B., Khasminskii, R., 1976. Stochastic Approximation and Recursive Estimation. American Mathematical Society, Providence, RI.

Robbins, H., Monro, S., 1951. A stochastic approximation method. Ann. Math. Statist. 22, 400–407.

Smith, A., Makov, U., 1978. A quasi-Bayes sequential procedure for mixtures. J. R. Statist. Soc. Ser. B 40 (1), 106–112.

Titterington, D., 1984. Recursive parameter estimation using incomplete data. J. R. Statist. Soc. B 46, 257–267.

Wang, S., Zhao, Y., 2001. Online Bayesian tree-structured transformation of HMMs with optimal model selection for speaker adaptation. IEEE Trans. Speech Audio Process. 9 (3), 663–677.

Weinstein, E., Feder, M., Oppenheim, A., 1990. Sequential algorithms for parameter estimation based on the Kullback–Leibler information measure. IEEE Trans. Acoustics Speech Signal Process. 38 (9), 1652–1654.