Almost Sure Convergence of Titterington's Recursive Estimator for Mixture Models

Shaojun Wang
Beckman Institute and Department of
Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801

Yunxin Zhao
Department of Computer Engineering
and Computer Science
University of Missouri at Columbia
Columbia, MO 65211

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.

Key Words-Recursive estimation, Incomplete data, Mixture model, Regular exponential family, Almost sure convergence, Stochastic approximation

---

Contact: Shaojun Wang
Address: School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213
Phone: (412) 268-1329
Fax: (412) 268-5576
Email: swang@cs.cmu.edu

1 Introduction

Recursive estimation algorithms have a wide range of applications in on-line system identification [18], adaptive filtering, pattern recognition [8], adaptive learning [17], sequential change detection [1], 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 [12]. As the result, the study of recursive estimation algorithms for mixture models is of a significant value.

In early eighties, Titterington [16] 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 [1, 7, 13] on regression function. One example is the assumption of the unique root of regression function as in the Robbins-Monro procedure [14]. Kushner and Clark [10] 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 [11], 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 [4] 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 [16], 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 $y_1, \cdots, y_k$ are received sequentially where each observation has the underlying pdf $p(y|\lambda)$ with $\lambda \in \Theta \subset \mathcal{R}^d$, 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 [1]:

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

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 $y^k = \{y_1, \cdots, y_k\}$, and $x^k = \{x_1, \cdots, x_k\}$. As in EM algorithm [5], define the auxiliary function $Q_{y^{k+1}}(\lambda, \lambda^{(k)}) = E[logp(x^{k+1}, y^{k+1}|\lambda)|y^{k+1}; \lambda^{(k)}]$. It follows that maximizing $Q_{\underline{y}^{k+1}}(\lambda, \lambda^{(k)})$ leads to improvements in $p(\underline{y}^{k+1}|\lambda)$ [5, 12]. By maximizing the second-order Taylor series expansion of $\frac{1}{k+1}Q_{y^{k+1}}(\lambda, \lambda^{(k)})$ with respect to λ and denoting the maximizing point by $\lambda^{(k+1)}$, we have

$$\lambda^{(k+1)} = \lambda^{(k)} + [-\tfrac{1}{k+1}\tfrac{\partial^2 Q_{y^{k+1}}(\lambda,\lambda^{(k)})}{\partial\lambda\partial\lambda^T}]^{-1}\tfrac{1}{k+1}\tfrac{\partial Q_{y^{k+1}}(\lambda,\lambda^{(k)})}{\partial\lambda}|_{\lambda=\lambda^{(k)}}$$

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

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

which corresponds to one step of the batch scoring EM algorithm [12], initialized at $\lambda^{(k)}$.

The batch algorithm of Eqn. (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, $\frac{\partial Q_{y^{k+1}}(\lambda,\lambda^{(k)})}{\partial\lambda} = \frac{\partial Q_{y^k}(\lambda,\lambda^{(k)})}{\partial\lambda} + \frac{\partial \ell_{y_{k+1}}(\lambda,\lambda^{(k)})}{\partial\lambda}$. Assuming that $\lambda^{(k)}$ maximizes $Q_{y^k}(\lambda, \lambda^{(k)})$ so that $\frac{\partial Q_{y^k}(\lambda,\lambda^{(k)})}{\partial\lambda}|_{\lambda=\lambda^{(k)}} = 0$, then we have

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

Furthermore the relation holds

$$\frac{\partial}{\partial\lambda}Q_{y_{k+1}}(\lambda, \lambda^{(k)})|_{\lambda=\lambda^{(k)}} = \frac{\partial}{\partial\lambda}E[logp(x_{k+1}, y_{k+1}|\lambda)|y_{k+1}, \lambda^{(k)}]|_{\lambda=\lambda^{(k)}} = \frac{\partial logp(y_{k+1}|\lambda)}{\partial\lambda}|_{\lambda=\lambda^{(k)}}$$

As the result, we obtain a recursive estimation formula as:

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

We call Eqn. (3) as the EM type MLE recursive estimator and it is exactly Eqn. (9) in [16].

Now we consider the MAP variant of Eqn. (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

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

Successive computation of Eqn. (4) for $k = 1, 2, \cdots$ 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 [15, 9] first approximates the successive posterior distributions by the "closest" tractable distribution within a given class $\mathcal{P}$, 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 [9].

Here we propose a new approach for recursive Bayesian learning. Define the auxiliary function of log posterior likelihood as $R_{y^{k+1}}(\lambda, \lambda^{(k)}) = Q_{y^{k+1}}(\lambda, \lambda^{(k)}) + logp(\lambda|\phi)$. It follows that maximizing $R_{y^{k+1}}(\lambda, \lambda^{(k)})$ leads to improvements in $p(\lambda|y^{k+1})$ [5, 12]. Maximizing the second-order Taylor series expansion of $\frac{1}{k+1} R_{y^{k+1}}(\lambda, \lambda^{(k)})$ with respect to λ and denoting the maximizing point by $\lambda^{(k+1)}$, we have

$$\lambda^{(k+1)} = \lambda^{(k)} + [-\tfrac{1}{k+1}\tfrac{\partial^2 R_{y^{k+1}}(\lambda,\lambda^{(k)})}{\partial\lambda\partial\lambda^T}]^{-1}\tfrac{1}{k+1}\tfrac{\partial R_{y^{k+1}}(\lambda,\lambda^{(k)})}{\partial\lambda}|_{\lambda=\lambda^{(k)}}$$

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

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

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

To convert Eqn. (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:

$$\lambda^{(k+1)} = \lambda^{(k)} + \frac{1}{k+1}[I_{CF}(\lambda^{(k)}) + \frac{1}{k+1} I_p(\lambda^{(k)}|\phi)]^{-1} \frac{\partial logp(y_{k+1}|\lambda)}{\partial\lambda}|_{\lambda=\lambda^{(k)}} \tag{6}$$

We refer to Eqn. (6) as the EM type MAP recursive estimator and $I_{CF}(\lambda^{(k)}) + \frac{1}{k+1} I_p(\lambda^{(k)}|\phi)$ 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 Eqns. (3) and (6) are unknown. However, if we assume the i.i.d. observations are generated by a mixture of full regular exponential families, then under mild conditions to be stated below, we obtain global convergence results. Before showing the result, we first give the definition of full regular exponential family and a lemma which is proved in [4].

Definition [2]: Let ν be a σ-finite measure on the Borel subsets of $\mathcal{R}^n$. Let $\mathcal{N} = \{\theta : \int e^{\theta\cdot y}\nu(dy) < \infty\}$. 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 [4]: Rewrite Eqn. (1) as

$$\lambda^{(k+1)} = \lambda^{(k)} + \varepsilon_k f(\lambda^{(k)}) + \varepsilon_k e_k \tag{7}$$

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

1) $V(\lambda)$ tends to ∞ if $\lambda \to \partial\Theta$ or $||\lambda|| \to \infty$.

2) $< \nabla_\lambda V(\lambda^0, \lambda), f(\lambda) >< 0$ for $\lambda \notin \Lambda$.

Further assume that the algorithm is A-stable, that is $\lambda^{(k)}$ remains in a compact subset of Θ, $lim_{k\to\infty}\sum_{k=1}^{m}\varepsilon_k e_k$ 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, $y_1, y_2, \cdots, y_k, \cdots$, we assume the underlying pdf to be a mixture of full regular exponential family $p(y|\lambda) = \sum_{g=1}^{\mathcal{G}} \omega_g p(y|\theta_g)$, with $\lambda = (\theta_1^T, \cdots, \theta_{\mathcal{G}}^T)^T$, and we use λ^0 to denote the true parameter. Assume that $[I_{CF}(\lambda^{(k)})]^{-1} < \infty$, $\frac{\partial logp(y|\lambda)}{\partial\lambda} < \infty$ and $\lambda^{(k)}$ does not tend to infinity. Then the sequence of recursive estimator (3) converges a.s. to the set of parameters

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

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:

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

It is well known that $V(\lambda^0, \lambda) \geq 0$ with equality if and only if $\lambda = \lambda^0$ [3]. It is easy to verify that $V(\lambda) \to \infty$ if $\lambda \to \partial\Theta$ or $||\lambda|| \to \infty$, since $p(y|\lambda)$ is a mixture of full regular exponential family.

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

$$\nabla_\lambda V(\lambda^0, \lambda) = -E_{\lambda^0}[\frac{\partial logp(y|\lambda)}{\partial\lambda}]$$

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

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

and thus

$$< \nabla_\lambda V(\lambda^0, \lambda), f(\lambda) >= -E_{\lambda^0}[\frac{\partial logp(y|\lambda)}{\partial\lambda}]^T I_{CF}^{-1}(\lambda)E_{\lambda^0}[\frac{\partial logp(y|\lambda)}{\partial\lambda}] < 0 \text{ for } \lambda \notin \Lambda$$

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 $M_k = \sum_{i=1}^{k}\varepsilon_i e_i$, where in this case $e_i = [I_{CF}(\lambda^{(i)})]^{-1}[\frac{\partial logp(y_i|\lambda^{(i)})}{\partial\lambda^{(i)}} - E_{\lambda^0}(\frac{\partial logp(y_i|\lambda^{(i)})}{\partial\lambda^{(i)}})]$ and $\varepsilon_k = \frac{1}{k+1}$. By assumption, $[I_{CF}(\lambda^{(k)})]^{-1} < \infty$ and $\frac{\partial logp(y|\lambda)}{\partial\lambda} < \infty$, and thus $e_i < \infty$. Define $\mathcal{F}^k := \sigma(y^1, \cdots, y^k)$. Then for $k \geq 1$, we have $E[M_k|\mathcal{F}^{k-1}] = E[M_{k-1}|\mathcal{F}^{k-1}] + E[\varepsilon_k e_k|\mathcal{F}^{k-1}] = M_{k-1} + \varepsilon_k E[e_k] = M_{k-1}$, and thus M_k is a martingale. Moreover, since $E[(M_i - M_{i-1})^2] = E[(\varepsilon_i e_i)^2] < C\varepsilon_i^2$, and $\sum_{k=1}^{\infty}(\varepsilon_k)^2 < \infty$, by orthogonality of martingale increment, $E[M_k^2] = \sum_{i=1}^{k} E[(M_i - M_{i-1})^2] = \sum_{i=1}^{k} E[(\varepsilon_i e_i)^2] < \infty$, and thus M_k is bounded in $\mathcal{L}^2$, i.e., e_k is an L_2-bounded Martingale increment. As the result, $limM_k$ exists a.s. [6], and the A-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 $\frac{1}{k+1}\sum_{i=1}^{k+1} logp(y_i|\lambda) \to E_{\lambda^0}[logp(y|\lambda)] = H(p(y|\lambda^0)) - V(\lambda^0, \lambda)$ a.s., where $H(p(y|\lambda^0)) = E_{\lambda^0}[logp(y|\lambda^0)]$ is the entropy of the source. On the other hand, $\frac{1}{k+1}Q_{y^{k+1}}(\lambda, \lambda')$

$= \frac{1}{k+1}\sum_{i=1}^{k+1} Q_{y_i}(\lambda,\lambda') \rightarrow E_{\lambda^0}[Q_y(\lambda,\lambda')] = E_{\lambda^0}[logp(y|\lambda)] + E_{\lambda^0}[E_{\lambda'}logp(x|y,\lambda)]$ a.s.. Since for each iteration, we always let $\lambda = \lambda'$, which enforces $E_{\lambda^0}[E_{\lambda'}logp(x|y,\lambda)] = 0$. So maximizing the normalized auxiliary function $\frac{1}{k+1}Q_{y^{k+1}}(\lambda,\lambda')|_{\lambda=\lambda'}$ is asymptotically equivalent to maximizing the normalized likelihood function $\frac{1}{k+1}logp(y^{k+1}|\lambda)$. This gives an explanation that why Eqn. (3) converges to the set given by Eqn. (8).

Theorem 2 (Global Convergence Property of the recursive MAP estimator): For an i.i.d. sequence, $y_1, y_2, \cdots, y_k, \cdots$, we assume that the underlying pdf to be a mixture of full regular exponential family $p(y|\lambda) = \sum_{g=1}^{\mathcal{G}} \omega_g p(y|\theta_g), \lambda = (\theta_1^T, \cdots, \theta_{\mathcal{G}}^T)^T$ and we use λ^0 to denote the true parameter. Assume that $[I_{CF}(\lambda^{(k)})]^{-1} < \infty$, $[I_p(\lambda^{(k)}|\phi)]^{-1} < \infty$, $\frac{\partial logp(y|\lambda)}{\partial\lambda} < \infty$ 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

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

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

Proof: By assumption of $[I_p(\lambda^{(k)}|\phi)]^{-1} < \infty$, we have $\frac{1}{k+1}I_p(\lambda^{(k)}) \rightarrow 0$ as $k \rightarrow \infty$, 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 ∞, $\frac{1}{k+1}logp(\lambda|y^{k+1}) \rightarrow E_{\lambda^0}[logp(y|\lambda)]$ a.s. and $\frac{1}{k+1}R_{y^{k+1}}(\lambda,\lambda')|_{\lambda=\lambda'} \rightarrow E_{\lambda^0}[logp(y|\lambda)]$ a.s.. This gives an explanation that why Eqn. (6) converges to the set given by Eqn. (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 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 $\frac{\partial logp(y_{k+1}|y^k,\lambda)}{\partial\lambda}$ has to be used in Eqns. (3) and (6) instead of $\frac{\partial logp(y_{k+1}|\lambda)}{\partial\lambda}$. As the result, all past data need to be stored, which excludes the algorithm as an on-line one. To overcome this difficulty, aproximations are normally made but convergence properties remain unknown.

References

[1] A. Benveniste, M. Metivier and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, 1990

[2] L. Brown, *Fundamentals of Statistical Exponential Families*, Institute of Mathematical Statistics, 1987

[3] T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991

[4] B. Delyon, "General Results on the Convergence of Stochastic Algorithms," *IEEE Trans. on Automatic Control*, Vol. 41, No. 9, pp. 1245-1255, September 1996

[5] A. Dempster, N. Laird and D. Rubin, "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm," *Journal of Royal Statistical Society, Series B*, Vol. 39, pp 1-38, 1977

[6] R. Durrett, *Probability: Theory and Examples*, Wadsworth Publishing Company, 1996

[7] V. Fabian, "On Asymptotically Efficient Recursive Estimation," *Annals of Statistics*, Vol. 6, No. 4, July, pp. 854-866, 1978

[8] K. Fu, *Sequential Methods in Pattern Recognition and Machine Learning*, Academic Press, 1968

[9] Q. Huo and C. Lee, "On-Line Adaptive Learning of the Continuous Density Hidden Markov Model Based on Approximate Recursive Bayes Estimate," *IEEE Trans. on Speech and Audio Processing*, Vol. 5, No. 2, pp. 161-172, March 1997

[10] H. Kushner and D. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer Verlag, 1978

[11] V. Lazarev, "Convergence of Stochastic Approximation Procedures in the Case of a Regression Equation with Several Roots," *Problems in Information Transmission*, pp. 66-78, July 1992

[12] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, John Wiley & Sons, New York, 1997

[13] B. Nevelson and R. Khasminskii, *Stochastic Approximation and Recursive Estimation*, American Mathematical Society, 1976

[14] H. Robbins and S. Monro, "A Stochastic Approximation Method," *Annals of Mathematical Statistics*, Vol. 22, pp. 400-407, 1951

[15] A. Smith and U. Makov, "A Quasi-Bayes Sequential Procedure for Mixtures," *Journal of the Royal Statistical Society*, Series B, Vol. 40, No. 1, pp. 106-112, 1978

[16] D. Titterington, "Recursive Parameter Estimation Using Incomplete Data," *Journal of Royal Statistical Society, B*, Vol 46, pp 257-267, 1984

[17] S. Wang and Y. Zhao, "On-Line Bayesian Tree-Structured Transformation of HMMs with Optimal Model Selection for Speaker Adaptation," submitted to *IEEE Trans. on Speech and Audio Processing*, 2000

[18] E. Weinstein, M. Feder and A. Oppenheim, "Sequential Algorithms for Parameter Estimation Based on the Kullback-Leibler Information Measure," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 38, No. 9, pp. 1652-1654, September 1990