Communications of the Korean Statistical Society
2011, Vol. 18, No. 3, 357–365

DOI: 10.5351/CKSS.2011.18.3.357

Estimating Parameters in Muitivariate Normal Mixtures

SungMahn Ahn[1,a], Sung Wook Baik[b]

[a]College of Business Administration, Kookmin University
[b]School of Computer Engineering, Sejong University

Abstract

This paper investigates a penalized likelihood method for estimating the parameter of normal mixtures in multivariate settings with full covariance matrices. The proposed model estimates the number of components through the addition of a penalty term to the usual likelihood function and the construction of a penalized likelihood function. We prove the consistency of the estimator and present the simulation results on the multi-dimensional nor-mal mixtures up to the 8-dimension.

Keywords: Multivariate normal mixtures, penalized likelihood, consistency of estimator.

1. Introduction

Multivariate normal mixture provides a well-defined model for high-dimensional data. In recent years, many researchers applied the model to various disciplines such as cluster analysis (Raftery and Dean, 1998), classification (Alexandridis *et al.*, 2004), and other areas.

However, several issues associated with the model have been pointed out and discussed over the past decades. First of all, testing for the number of components in a mixture is a difficult problem. Roeder and Wasserman (1997) used BIC, although Solka *et al.* (1995) used AIC to find the correct number of components. Another issue is the identifiability problem that states that the true distribution is represented by more than one parameter. Redner (1981) handled this problem by using the quotient topological space where equivalent solutions are mapped into a single point. In addition, another problem is related to estimating its parameters.

The maximum likelihood framework is among the most commonly used approaches. But the likelihood function is not bounded above and this usually happens on the boundary points of the parameter space. In order to avoid the problem, authors place constraints on the parameter space. Hathaway (1985), for example, proposed a constrained maximization of the likelihood function by putting a restriction on the variances. However, a more general approach is to use penalized maximum likelihood estimation(PMLE), where new likelihood function is formulated by adding a penalty term. Several authors including Ciuperca *et al.* (2003), Ingrassia (2004), and Chen and Tan (2009) used this approach and what they used as a constraint was mostly on variances or variance matrices in multivariate cases.

In this paper, we investigate a penalized likelihood method for estimating the parameters of normal mixtures in a multivariate setting. The proposed model estimates the number of components while avoiding the infinite likelihood problem. In order to estimate the number of components, we impose one assumption that the initial solution of the EM algorithm is overfitted in terms of the number of components.

[1] Corresponding author: Associate professor, College of Business Administration, Kookmin University, 861-1, Jeongneung-dong, Seongbuk-gu, Seoul 136-702, Korea. E-mail: sahn@kookmin.ac.kr

The paper is organized as follows. In Section 2, the proposed model is explained with some assumptions. Section 3 presents theorems on the consistency of the proposed model and parameter estimation algorithm along with simulation results follow in Section 4.

2. Penalized Likelihood

Let $x_1, x_2, \ldots, x_n$ be d-dimensional *i.i.d.* random variables. We assume that the distribution of x_1 is known except for some parameter, γ. The set of all parameter points Γ is called the parameter space and γ_0 will denote the true parameter. It is also assumed that there is a σ -finite measure μ such that for each $\gamma\varepsilon\Gamma$ the probability measure μ_γ is absolutely continuous with respect to μ. We let $g(x;\gamma)$ denote the density of μ_γ with respect to μ. In case of normal mixtures γ can be represented as

$$\gamma = \left(\pi_1, \ldots, \pi_g, \mu_1, \ldots, \mu_g, \Sigma_1, \ldots, \Sigma_g\right)^T,$$

where $\pi_j's$ are mixing proportion such that $\sum_{j=1}^{g} \pi_j = 1$, $\mu_j's$ are component means, and $\Sigma_j's$ are component covariance matrices. Normal mixture is defined as follows.

$$g(x;\gamma) = \sum_{j=1}^{g} \pi_j \varphi\left(x;\mu_j,\Sigma_j\right),$$

where $\varphi_j(x;\mu_j,\Sigma_j) = (2\pi)^{-d/2}|\Sigma_j|^{-1/2}\exp\{-1/2(x-\mu_j)^T\Sigma^{-1}(x-\mu_j)\}$.

The most commonly used method to estimate the parameters is the maximum likelihood framework, where the likelihood function is given by

$$g_n(x;\gamma) = \prod_{i=1}^{n} g(x_i;\gamma).$$

The main idea of our approach to estimating γ is such that we have an excessive number of components to start with and try to eliminate spurious components during the process of parameter estimation. We consider a penalized likelihood function defined as

$$f_n(x;\gamma) = g_n(x;\gamma)p_n(\gamma). \tag{2.1}$$

The penalized likelihood such as (2.1) was formulated, for example, in Ciuperca *et al.* (2003) and Chen and Tan (2009). Chen and Tan (2009) especially used a penalty that depends on the covariance matrices and the sample size. However, instead of covariance we chose the penalty to be a function of the mixing proportion and propose the following penalty function.

$$p_n(\gamma) = \left[\prod_{j=1}^{g} \pi_j^{(\alpha-1)}\right]^{\lambda n}, \tag{2.2}$$

where α is a positive number smaller than 1 and λ is a small positive number for controlling the overall effect of the penalization on the likelihood. By maximizing the penalty function, we encourage some of $\pi_j's$ to take on values close to zero, since (2.2) reaches the minimum when all the $\pi_j's$ have the same value. Since we assume that the current model is overfitted in terms of the number of components, it would be reasonable to express the assumption as follows.

$$\prod_{j=1}^{g} \pi_j^{(\alpha-1)} < \prod_{j=1}^{g} \pi_{0j}^{(\alpha-1)}, \tag{2.3}$$

where $\pi'_{0j}s$ are the true mixing proportions.

Now we summarize some assumptions made in the proposed model.

Assumption 1. *$\pi_j > \epsilon$ (ϵ is an arbitrary small number). We can eliminate a component if the corresponding π_j is smaller than ϵ.*

Assumption 2. *The inequality (2.3) is satisfied.*

All the assumptions in Redner (1981) and thus in Wald (1949) are also made here. One of them is restated below.

Assumption 3. *The parameter space Γ is a metric space with metric $\delta(\cdot,\cdot)$ and has the property that every closed and bounded subset of Γ is compact. This is from Redner (1981) and an example can be found in Hathaway (1985). As with Wald (1949), we define functions f and f^* as follows.*

$$f(x;\gamma,\rho) = \sup\{f_1(x;\gamma') : \delta(\gamma,\gamma') < \rho,\ \gamma' \in \Gamma\},$$

where $f_1(x;\gamma) = \sum_{j=1}^{g} \pi_j \varphi_j(x;\mu_j,\Sigma_j)[\prod_{j=1}^{g} \pi_j^{(\alpha-1)}]^{\lambda}$ and

$$f^*(x;\gamma,\rho) = \max(1, f(x;\gamma,\rho)).$$

3. Consistency of the Penalized Maximum Likelihood Estimator

We first give some lemmas that will be used in the main theorem.

Lemma 1. *$\int \log f^*(x;\gamma,\rho)du_{\gamma_0}$ is finite.*

Proof: According to the definition of $f(x;\theta,\rho)$, we have

$$\begin{aligned}\int \log f^*(x;\gamma,\rho)du_{\gamma_0} &= \int \log\max(1, f(x;\gamma,\rho))du_{\gamma_0} \\ &\le \sum_{j=1}^{g}\int \log\max\left[1, \sup\left\{\sum_{j=1}^{g}\pi_j\varphi_j(x;\mu_j,\Sigma_j)\left[\prod_{j=1}^{g}\pi_j^{(\alpha-1)}\right]^{\lambda} : \delta(\gamma,\gamma')<\rho, \gamma' \in \Gamma\right\}\right]du_{\gamma_0}.\end{aligned}$$

The above inequality follows from Theorem 5 of Redner (1981). Since $\pi'_j s$, α and λ are all finite, $[\prod_{j=1}^{g} \pi_j^{(\alpha-1)}]^{\lambda}$ is finite. Therefore, with the result of Theorem 5 of Redner (1981), the right-hand side of the inequality is finite. This concludes the proof. □

Lemma 2. *$\int |\log f_1(x;\gamma_0)|du_{\gamma_0}$ is finite.*

Proof: According to the definition of $f_1(x;\gamma_0)$ we have

$$\begin{aligned}\int \left|\log f_1(x;\gamma_0)\right|du_{\gamma_0} &= \int \left|\log\sum_{j=1}^{g}\pi_j\varphi_j(x;\mu_j,\Sigma_j)\left[\prod_{j=1}^{g}\pi_j^{(\alpha-1)}\right]^{\lambda}\right|du_{\gamma_0} \\ &= \int \left|\log\sum_{j=1}^{g}\pi_j\varphi_j(x;\mu_j,\Sigma_j) + \lambda(\alpha-1)\sum_{j=1}^{g}\log\pi_j\right|du_{\gamma_0}\end{aligned}$$

$$\leq \int \left| \log \sum_{j=1}^{g} \pi_j \varphi_j(x; \mu_j, \Sigma_j) \right| du_{\gamma_0} + \int \left| \lambda(\alpha - 1) \sum_{j=1}^{g} \log \pi_j \right| du_{\gamma_0}$$

$$\leq \int \left| \log \sum_{j=1}^{g} \pi_j \varphi_j(x; \mu_j, \Sigma_j) \right| du_{\gamma_0} + \int \left| \lambda(\alpha - 1) g \log \varepsilon \right| du_{\gamma_0} \tag{3.1}$$

$$= \int \left| \log \sum_{j=1}^{g} \pi_j \varphi_j(x; \mu_j, \Sigma_j) \right| du_{\gamma_0} + \lambda(\alpha - 1) g \log \varepsilon. \tag{3.2}$$

(3.1) follows from Assumption 3 and the first term in (3.2) is finite following Theorem 5 of Redner (1981). This concludes the proof. □

Redner (1981) defined the quotient space, $\tilde{\Gamma}$, related to Γ and proved that the MLE is consistent in every compact parameter subset $\tilde{\Gamma}$ of Γ that contains γ_0. Therefore, since the results of Lemma 1 and Lemma 2 are also applicable to the quotient space, the nonidentifiable problem is considered to have been resolved hereinafter.

Lemma 3. *For any $\gamma \neq \gamma_0$, we have*

$$E_0 \log f_1(x; \gamma) < E_0 \log f_1(x; \gamma_0). \tag{3.3}$$

Proof: Let us define $\nu = \log f_1(x, \gamma) - \log f_1(x, \gamma_0)$. Then

$$E_0 [e^{\nu}] = E \left[\frac{f_1(x, \gamma)}{f_1(x, \gamma_0)} \right] = \int \left[\frac{\prod_{j=1}^{g} \pi_j^{(\alpha-1)}}{\prod_{j=1}^{g} \pi_{0j}^{(\alpha-1)}} \right]^{\lambda} du_{\gamma_0} = \left[\prod_{j=1}^{g} \left(\frac{\pi_j}{\pi_{0j}} \right)^{(\alpha-1)} \right]^{\lambda}.$$

Thus, we have

$$\log E_0 [e^{\nu}] = \lambda \left[\sum_{j=1}^{g} (\alpha - 1) \log \pi_j - \sum_{j=1}^{g} (\alpha - 1) \log \pi_{0j} \right] < 0.$$

The inequality follows from the Assumption 2. By Jensen's inequality, we obtain $E_0 [\nu] \leq \log E_0 [e^{\nu}] < 0$. Therefore, $E_0 [\nu] < 0$, which is equivalent to (3.3). □

Now, we are ready to prove the consistency of the penalized likelihood estimator.

Theorem 1. *If the Assumptions 1–3 are satisfied, then PMLE of γ_0 in a compact subset of Γ is a consistent estimator.*

Proof: In the previous Lemmas 1–3, we showed the following.

$$\int \log f(x; \gamma, \rho) du_{\gamma_0} < \infty,$$

$$\int \left| \log f_1(x; \gamma_0) \right| du_{\gamma_0} < \infty$$

and

$$E_0 \log f_1(x; \gamma) < E_0 \log f_1(x; \gamma_0).$$

Those correspond to Wald's Assumption 2, Assumption 6 and Lemma 1, respectively. Since all the other assumptions are obviously satisfied, the rest of Proof follows from the Theorem 1 and Theorem 2 of Wald. □

4. Simulation

4.1. Algorithm

In order to estimate parameters maximizing the penalized likelihood proposed as in (2.1), we use the EM algorithm (Dempster *et al.*, 1977). In the EM algorithm, it is assumed that each observation $x_i\,(i = 1, 2, \ldots, n)$ is associated with an unobserved state $z_i\,(i = 1, 2, \ldots, n)$, which is the indicator vector of length g, $z_i = (z_{i1}, z_{i2}, \ldots, z_{ig})'$, and z_{ij} is 1 if and only if x_i is generated by density j and 0 otherwise. The joint distribution of x_i and z_i under normal mixture assumption is (Titterington *et al.*, 1985)

$$f(x_i, z_i; \gamma) = \prod_{j=1}^{g} \left[\pi_j \varphi(x_i; \mu_j, \Sigma_j)\right]^{z_{ij}}.$$

Therefore, the penalized log likelihood for the complete-data based on (2.1) is

$$\sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij} \log \pi_j \varphi(x_i; \mu_j, \Sigma_j) + \lambda n \sum_{j=1}^{g} (\alpha - 1) \log \pi_j. \tag{4.1}$$

Since z_{ij} is unknown, the complete-data log likelihood cannot be used directly. Thus we instead work with its expectation, that is, we apply the EM algorithm. The EM algorithm is defined by cycling back and forth between E-step and M-step until likelihood is no longer improved. In E-step, we find the conditional expectation of the hidden variables as follows.

$$\hat{z}_{ij}^{(k+1)} = \frac{\hat{\pi}_j^{(k)} \varphi\left(x_i; \hat{\mu}_j^{(k)}, \hat{\Sigma}_j^{(k)}\right)}{\sum_{j=1}^{g} \hat{\pi}_j^{(k)} \varphi\left(x_i; \hat{\mu}_j^{(k)}, \hat{\Sigma}_j^{(k)}\right)}.$$

In M-step, we find the optimal parameter values that maximize (4.1). Those are

$$\hat{\pi}_j = \frac{1/n \sum_{i=1}^{n} \hat{z}_{ij}^{(k+1)} + \lambda(\alpha - 1)}{1 + \lambda g(\alpha - 1)},$$
$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} \hat{z}_{ij}^{(k+1)} x_i}{\sum_{i=1}^{n} \hat{z}_{ij}^{(k+1)}}$$

and

$$\hat{\Sigma}_j = \frac{\sum_{i=1}^{n} \hat{z}_{ij}^{(k+1)} \left(x_i - \hat{\mu}_j^{(k)}\right)\left(x_i - \hat{\mu}_j^{(k)}\right)^T}{\sum_{i=1}^{n} \hat{z}_{ij}^{(k+1)}}.$$

The difference from the usual EM algorithm is that we check if the values of some $\pi_j's$ are too small after each iteration step, and if it happens, the corresponding components are eliminated from the model. The current value of the penalized likelihood can drop down temporarily when we eliminate insignificant components, in which case we reset the current likelihood value so that the procedure continues with a model having a smaller number of components.

4.2. The infinite likelihood problem

Theoretically the likelihood function is not bounded above and therefore the algorithm could diverge. This usually happens on the boundary points of the parameter space. We, however, proved the consistency of the estimator with a compact subset by assuming that the density function goes to zero whenever parameters approach a boundary point. This assumption is reasonable except for a situation where one of the components is becoming degenerate on a single data point. If, for component j, $\mu_j \to x_i$ and $|\Sigma_j| \to 0$, then the likelihood goes to infinity.

The proposed model prevents the situation from happening by eliminating the component before it gets stuck on a data point. Since $\hat{z}_{ij}$ is the probability that x_i was generated by component j, $\hat{z}_{ij} \to 1$ and $z_{kj} \to 0\,(k \neq i)$ if $\mu_j \to x_i$ and $|\Sigma_j| \to 0$ as n goes to infinity. Thus, $1/n \sum \hat{z}_{ij} \to 0$ and therefore, $\pi_j \to \lambda(\alpha - 1)/\{1 + \lambda g(\alpha - 1)\}$. If we choose 0.01 and 0.001 for α and λ, respectively, then $\pi_j \to -0.001$ with $g = 5$. It seems awkward that π_j goes to a negative value; however we eliminate the corresponding component before π_j reaches a negative value.

4.3. Simulation results

We present the simulation results with some test distributions, which cover up to the 8-dimension. For up to three dimensional distributions, we started with 10 cases of sample size of 1000 and if the results are not good enough, 10 more cases of sample size of 2000 have been drawn from the same distribution and tested. For higher than three-dimensional distributions, we tested only the standard normal distribution. As a measure of error, we can use the integrated absolute error(IAE), which is defined as

$$\int_{R^d} \left| f_0(x) - \hat{f}(x) \right| dx.$$

However, for the sake of simplicity we used the following formula to approximate IAE.

$$\sum_{x \in R^d} \left| f_0(x) - \hat{f}(x) \right| (\Delta x)^d,$$

which is the sum of the volumes of d-dimensional hypercube multiplied by $|f_0(x) - \hat{f}(x)|$. The sum is over the support of the probability density.

The simulation procedure along with model parameter values is as follows.

1. Set a true underlying distribution.
2. Generate random samples according to the distribution.
3. Find an overfitted model with excessive number of components. Overfitted models can be obtained, for example, by using the adaptive mixture density estimation (Priebe, 1994).
4. Run the PMLE algorithm with $\alpha = 0.01$ and $\lambda = 0.001$ and remove components when $\pi < 0.03$.

Two different bivariate distributions are tested. Those are

1. Unimodal: $N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}2 & .4\\.4 & 2\end{bmatrix}\right)$.
2. Bimodal: $\frac{3}{10}N\left(\begin{bmatrix}-3\\-3\end{bmatrix}, \begin{bmatrix}1 & .8\\.8 & 1\end{bmatrix}\right) + \frac{7}{10}N\left(\begin{bmatrix}3\\3\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$.

Table 1: Results for the bivariate distributions

Type of dist.	Sample size	# of components	# of cases	Average error	Average iter.
Unimodal	1000	1	4	.0588	705
		2	2	.0910	1394
		3	1	.0970	652
		4	1	.1446	322
		5	2	.1440	360
	2000	1	10	.0491	771
Bimodal	1000	2	2	.0678	294
		3	5	.0809	811
		4	1	.0995	545
		5	1	.1358	1394
		8	1	.1513	126
	2000	2	10	.0405	621

Table 2: Results for the trivariate distributions

Type of dist.	Sample size	# of components	# of cases	Average error	Average iter.
Unimodal	1000	1	1	.0670	418
		2	2	.1014	905
		3	2	.1348	1080
		5	1	.2159	551
		8	2	.2888	115
		10	2	.2903	127
	2000	1	8	.0665	879
		2	1	.0813	698
		3	1	.0827	1022
Bimodal	1000	2	1	.0795	395
		3	1	.1126	585
		5	3	.1879	236
		6	1	.1982	60
		9	3	.2194	44
		11	1	.2554	33
	2000	2	6	.0715	819
		3	3	.0843	431
		4	1	.1048	1439

Simulation results are summarized in Table 1. With the sample size of 1000, about half of the unimodal distributions have been poorly estimated and the story is similar for the bimodal distribution; however, both distributions are well estimated with the sample size of 2000.

Now we test two different trivariate distributions. Those are

1. Unimodal: $$N\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}2 & .4 & 0\\ .4 & 1 & -.6\\ 0 & -.6 & 3\end{bmatrix}\right).$$

2. Bimodal: $$\frac{1}{2}N\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}1 & .8 & 0\\ .8 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}\right) + \frac{1}{2}N\left(\begin{bmatrix}3\\3\\3\end{bmatrix}, \begin{bmatrix}1 & -.8 & 0\\ -.8 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}\right).$$

Simulation results are summarized in Table 2. In the trivariate case, the results of the sample size of 1000 look bad. The results of the sample size of 2000 seem acceptable for both distributions, although we might need larger sample sizes to obtain better results for the bimodal case.

As with other maximum-likelihood-type estimation methods, the suggeted method requires large sample sizes for accurate estimation. As we can see in Table 1 and Table 2, the sample size should

Table 3: Results for the higher dimensional distributions

Dimension	Sample Size	Average Error	# of successful cases
4	2000	.3224	4
	3000	.1166	9
	4000	.0681	10
	5000	.0773	10
	6000	.0652	10
5	2000	.2344	2
	3000	.2269	6
	4000	.1045	8
	5000	.0800	10
	6000	.0805	10
6	2000	.3767	0
	3000	.2639	5
	4000	.1035	8
	5000	.0911	10
	6000	.1008	9
7	2000	.4571	1
	3000	.2495	2
	4000	.1466	7
	5000	.1719	8
	6000	.0817	10
8	2000	.4218	0
	3000	.3441	1
	4000	.2693	2
	5000	.1560	6
	6000	.0936	10

be larger than 1000 to have a fairly good estimators of the densities chosen. It is hard to tell *a priori* how big the sample size should be. The simulation results tell us that the likelihood surfaces created with the sample size of 1000 have many spurious local maxima, which could be smoothed out with additional likelihood provided by more sample points.

Now, we present the simulation results for 4- to 8-dimensional distributions. We tested only the standard normal distribution, which is $N(O_{n\times 1}, I_{n\times n})$. For each dimension, total 50 sets of samples were drawn (10 sets of 2000 sample points, 10 sets of 3000 sample points, and so on up to 10 sets of 6000 sample points). And the results are shown in Table 3, where we see average errors and the number of successful cases, which we define to end up with a single number of component for each category of the simulation. With sample sizes of 6000 or more, we were able to correctly estimate the number of components most of the time.

5. Summary and Discussion

This paper investigated a penalized likelihood method for estimating the parameter of normal mixtures in multivariate settings with full covariance matrices. The main idea of our approach to estimating γ is such that we have excessive number of components to start with and try to eliminate spurious components during the process of parameter estimation. In order to do so, we added a penalty term to usual likelihood function and constructed a penalized likelihood function defined as (2.1) with the penalty term to be (2.2). We also added some assumptions over Redner (1981).

To prove consistency of the estimator, we basically followed the technique of Wald (1949) and used the results of Redner (1981) who handled the unidentifiable problem by using the quotient topological space where equivalent solutions are mapped into a single point. However, some of the assumptions of Wald needed to be verified, since we extended the likelihood function to have a specific

penalty term. We verified the assumptions through the Lemmas 1–3.

The proposed model tackled the issues inherent in normal mixtures such as regarding number of components, identifiability problem and boundary point problem.

Finally, the simulation results can be summarized as follows.

1. The algorithm works in the multi-dimensional normal mixture models.
2. Larger sample sizes are needed for distributions having more components.
3. Larger sample sizes are needed for higher dimensional distributions.
4. The number of iterations does not vary much depending on distributions.
5. Proper sample sizes to get reasonably good estimators is suggested.

References

Alexandridis, R., Lin, S. and Irwin, M. (2004). Class discovery and classification of tumor samples using mixture modeling of gene expression data-A unified approach, *Biominfomatics*, **20**, 2545–2552.

Chen, K. and Tan, X. (2009). Inference for multivariate normal mixtures, *Journal of Multivariate Analysis*, **100**, 1367–1383.

Ciuperca, G., Ridolfi, A. and Idier, J. (2003). Penalized maximum likelihood estimator for normal Mixtures, *Scandinavian Journal of Statistics*, **30**, 45–59.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society(B)*, **39**, 1–38.

Hathaway, R. J. (1985). A constrained formulation of maximum likelihood estimation for normal mixture distributions, *Annals of Statistics*, **13**, 795–800.

Ingrassia, S. (2004). A likelihood-based constrained algorithm for multivariate normal mixture models, *Statistical Methods & Applications*, **13**, 151–166.

Priebe, C. E. (1994). Adaptive mixtures, *Journal of American Statistical Association*, **89**, 796–806.

Raftery, A. E. and Dean, N. (1998). Variable selection for model-based clustering, *Journal of American Statistical Association*, **101**, 168–178.

Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions, *Annals of Statistics*, **9**, 225–228.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals, *Journal of the American Statistical Association*, **92**, 894–902.

Solka, J. L., Poston, W. L. and Wegman, E. J. (1995). A visualization technique for studying the iterative estimation of mixture densities, *Journal of Computational and Graphical Statistics*, **4**, 180–198.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*, Wiley.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate, *Annals of Mathematical Statistics*, **20**, 595–601.

Received January 2011; Accepted March 2011