Statistica Sinica **18**(2008), 443-465

INFERENCE FOR NORMAL MIXTURES IN MEAN AND VARIANCE

Jiahua Chen[1], Xianming Tan[2] and Runchu Zhang[2]

[1]*University of British Columbia and* [2]*LPMC Nankai University*

Abstract: A finite mixture of normal distributions, in both mean and variance parameters, is a typical finite mixture in the location and scale families. Because the likelihood function is unbounded for any sample size, the ordinary maximum likelihood estimator is not consistent. Applying a penalty to the likelihood function to control the estimated component variances is thought to restore the optimal properties of the likelihood approach. Yet this proposal lacks practical guidelines, has not been indisputably justified, and has not been investigated in the most general setting. In this paper, we present a new and solid proof of consistency when the putative number of components is equal to, and when it is larger than, the true number of components. We also provide conditions on the required size of the penalty and study the invariance properties. The finite sample properties of the new estimator are also demonstrated through simulations and an example from genetics.

Key words and phrases: Bernstein inequality, invariant estimation, mixture of normal distributions, penalized maximum likelihood, strong consistency.

1. Introduction

Finite mixture models have wide applications in scientific disciplines, especially in genetics (Schork, Allison and Thiel (1996)). In particular, the normal mixture in both mean and variance was first applied to crab data in Pearson (1894), and is the most popular model for analysis of quantitative trait loci, see Roeder (1994), Chen and Chen (2003), Chen and Kalbfleisch (2005), and Tadesse, Sha and Vannucci (2005). In general, let $f(x, \lambda)$ be a parametric density function with respect to some σ-finite measure and parameter space Λ, usually a subset of some Euclidean space. The density function of a finite mixture model is given by $f(x; G) = \sum_{j=1}^{p} \pi_j f(x; \lambda_j)$ where p is the number of components or the order of the model, $\lambda_j \in \Lambda$ is the parameter of the jth component density, π_j is the proportion of the jth component density, and G is the mixing distribution which can be written as $G(\lambda) = \sum_{j=1}^{p} \pi_j I(\lambda_j \leq \lambda)$ with $I(\cdot)$ being the indicator function.

In this paper, we focus on inference problems related to the univariate normal mixture distribution with parameter λ representing the mean and variance

(θ, σ^2). Let $\phi(x) = 1/(\sqrt{2\pi})\exp\{-x^2/2\}$. In normal mixture models, the component density is given by $f(x;\theta,\sigma) = \sigma^{-1}\phi(\sigma^{-1}(x-\theta))$.

The parameter space of G can be written as

$$\Gamma = \Big\{G = (\pi_1, \ldots, \pi_p, \theta_1, \ldots, \theta_p, \sigma_1, \ldots, \sigma_p) : \\ \sum_{j=1}^{p} \pi_j = 1, \pi_j \geq 0, \sigma_j \geq 0 \ \text{ for } \ j = 1, \ldots, p\Big\}.$$

For convenience, we use G to represent both the mixing distribution and its relevant parameters. We understand that permuting the order of the components does not change the model. Hence, without loss of generality, we assume $\sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_p$.

Let $X_1, \ldots, X_n$ be a random sample from a finite normal mixture distribution $f(x;G)$. A fundamental statistical problem is to estimate the mixing distribution G. Pearson (1894) proposed the method of moments for estimating the parameters in the univariate normal mixture. Many other approaches have been proposed, such as those discussed in McLachlan and Basford (1987) and McLachlan and Peel (2000). The maximum likelihood estimator (MLE), known for its asymptotic efficiency for regular statistical models, is one of the most commonly used approaches (Lindsay (1995)). However, in the case of finite normal mixture distributions in both mean and variance, the MLE is not well defined. Note that the log-likelihood function is

$$l_n(G) = \sum_{i=1}^{n} \log f(X_i; G) = \sum_{i=1}^{n} \log\Big\{\sum_{j=1}^{p} \frac{\pi_j}{\sigma_j}\phi\Big(\frac{X_i - \theta_j}{\sigma_j}\Big)\Big\}.$$

By letting $\theta_1 = X_1$ and $\sigma_1 \to 0$ with the other parameters fixed, we have $l_n(G) \to \infty$. That is, the ordinary maximum-likelihood estimator of G is not well-defined (Day (1969) and Kiefer and Wolfowitz (1956)).

To avoid this difficulty, researchers often turn to estimators on constrained parameter spaces. For example, Redner (1981) proved that the maximum likelihood estimator of G exists and is globally consistent in every compact subparameter space containing the true parameter G_0. When p is known, Hathaway (1985) proposed estimating G by maximizing the likelihood function within a restricted parameter space. Despite the elegant results of Redner (1981) and Hathaway (1985), these methods suffer, at least theoretically, from the risk that the true mixing distribution G_0 may not satisfy the constraint imposed.

We advocate the approach of adding a penalty term to the ordinary log-likelihood function. We define the penalized log-likelihood as

$$pl_n(G) = l_n(G) + p_n(G), \tag{1.1}$$

so that $p_n(G) \to -\infty$ as $\min\{\sigma_j : j = 1, \ldots, p\} \to 0$. We then estimate G with the penalized maximum likelihood estimator (PMLE) $\widetilde{G}_n = \arg\max_G pl_n(G)$. The penalized-likelihood-based method is a promising approach for countering the unboundedness of $l_n(G)$ while keeping the parameter space Γ unaltered. However, to make the PMLE work, one has to consider what penalty functions $p_n(G)$ are suitable. This task proves challenging. Ridolfi and Idier (1999, 2000) proposed a class of penalty functions based on a Bayesian conjugate prior distribution, but the asymptotic properties of the corresponding PMLE were not discussed. Under some conditions on $p_n(G)$ and with p assumed known, Ciuperca, Ridolfi and Idier (2003) provided an insightful proof of strong consistency of the PMLE of G under the normal mixture model. Their proof was for the case where $p = p_0$ is known, and it contains a few loose steps that do not seem to have quick fixes, see Tan (2005).

We use a novel technique to establish the strong consistency of the PMLE for a class of penalty functions, whether or not the true value of p is known. In addition, the proper order of the penalty is established. The paper is organized as follows. We first introduce two important technical lemmas in Section 2, and then present a detailed proof of the strong consistency of the PMLE in Section 3. In Section 4, we present some simulation results and a data example. We summarize the paper in Section 5.

2. Technical Lemmas

To make the penalized likelihood approach work, we use a penalty to counter the effect of observations close to the location parameters. For this purpose, we assess the number of observations falling in a small neighborhood of the location parameters in G.

Let $\Omega_n(\sigma) = \sup_\theta \sum I(0 < X_i - \theta < -\sigma \log \sigma)$ be the number of observations on the positive side of a small neighborhood of θ. We are interested in $\Omega_n(\sigma)$ only when σ is small. The number of observations on the negative side of θ can be assessed in the same way. Let $F_n(x) = n^{-1} \sum_{i=1}^n I(X_i \le x)$ be the empirical distribution function. We have $\Omega_n(\sigma) = n \sup_\theta [F_n(\theta - \sigma \log \sigma) - F_n(\theta)]$. Let $F = E(F_n)$ be the true cumulative distribution function. We take $M = \max\{\sup_x f(x; G_0), 8\}$ and $\delta_n(\sigma) = -M\sigma \log(\sigma) + n^{-1}$, where G_0 is the true mixing distribution. The following lemma uses Bahadur's representation to give an order assessment of $n^{-1}\Omega_n(\sigma)$. With a slight abuse of the probability concept, when an inequality involving random quantities holds as $n \to \infty$ except for a zero probability event, we claim that the inequality is true almost surely. Further, if there is no risk of confusion, we omit the phrase "almost surely."

Lemma 1. *Under the finite normal mixture model assumption, as $n \to \infty$ and almost surely, we have:*

1. *for each given* σ *between* $\exp(-2)$ *and* $8/(nM)$,

$$\sup_\theta [F_n(\theta - \sigma \log \sigma) - F_n(\theta)] \leq 2\delta_n(\sigma);$$

2. *uniformly for* σ *between* 0 *and* $8/(nM)$,

$$\sup_\theta [F_n(\theta - \sigma \log \sigma) - F_n(\theta)] \leq 2\frac{(\log n)^2}{n}.$$

Proof. 1. Let $\eta_0, \eta_1, \ldots, \eta_n$ be such that $\eta_0 = -\infty$, $F(\eta_i) = i/n$, $i = 1, \ldots, n-1$, $\eta_n = \infty$. We have

$$\begin{aligned}
&\sup_\theta [F_n(\theta - \sigma \log \sigma) - F_n(\theta)] \\
&\quad \leq \max_j [F_n(\eta_j - \sigma \log \sigma) - F_n(\eta_{j-1})] \\
&\quad \leq \max_j [\{F_n(\eta_j - \sigma \log \sigma) - F_n(\eta_{j-1})\} - \{F(\eta_j - \sigma \log \sigma) - F(\eta_{j-1})\}] \\
&\qquad + \max_j [F(\eta_j - \sigma \log \sigma) - F(\eta_{j-1})].
\end{aligned}$$

By the Mean Value Theorem and for some $\eta_j \leq \xi_j \leq \eta_j - \sigma \log \sigma$, we have

$$\begin{aligned}
F(\eta_j - \sigma \log \sigma) - F(\eta_{j-1}) &= F(\eta_j - \sigma \log \sigma) - F(\eta_j) + n^{-1} \\
&= f(\xi_j; G_0)|\sigma \log \sigma| + n^{-1} \\
&\leq M|\sigma \log \sigma| + n^{-1} = \delta_n(\sigma).
\end{aligned}$$

In summary, we have $\max_j [F(\eta_j - \sigma \log \sigma) - F(\eta_{j-1})] \leq \delta_n(\sigma)$. Further, for $j = 1, \ldots, n$, define $\Delta_{nj} = |\{F_n(\eta_j - \sigma \log \sigma) - F_n(\eta_{j-1})\} - \{F(\eta_j - \sigma \log \sigma) - F(\eta_{j-1})\}|$. By the Bernstein Inequality (Serfling (1980)), for any $t > 0$ we have

$$P\{\Delta_{nj} \geq t\} \leq 2\exp\Big\{ -\frac{n^2 t^2}{2n\delta_n(\sigma) + \frac{2}{3}nt} \Big\}. \tag{2.1}$$

Since $|\sigma \log \sigma|$ is monotonic in σ for $\exp(-2) > \sigma > 8/(nM)$,

$$|\sigma \log \sigma| \geq \frac{8}{nM} \log \frac{nM}{8} \geq \frac{8 \log n}{nM}.$$

By letting $t = \delta_n(\sigma)$ in (2.1), we obtain

$$\begin{aligned}
P\{\Delta_{nj} \geq \delta_n(\sigma)\} &\leq 2\exp\Big\{ -\frac{3}{8} n\delta_n(\sigma) \Big\} \\
&\leq 2\exp\Big\{ -\frac{3}{8} Mn|\sigma \log \sigma| \Big\} \\
&\leq 2n^{-3}.
\end{aligned}$$

Thus for any σ in this range, $P\{\max_j \Delta_{nj} \geq \delta_n(\sigma)\} \leq \sum P\{\Delta_{nj} \geq \delta_n(\sigma)\} \leq 2n^{-2}$. Linking this inequality back to $\sup_\theta[F_n(\theta - \sigma\log(\sigma)) - F_n(\theta)]$, we get

$$P\Big\{\sup_\theta[F_n(\theta - \sigma\log\sigma) - F_n(\theta)] \geq 2\delta_n(\sigma)\Big\} \leq P\Big\{\max_j \Delta_{nj} \geq \delta_n(\sigma)\Big\} \leq 2n^{-2}.$$

The conclusion then follows from the Borel-Cantelli Lemma.

2. When $0 < \sigma < 8/(nM)$, we choose $t = n^{-1}(\log n)^2$ in (2.1). For n large enough, $2\delta_n(\sigma) < t/3$. Hence $P\{\Delta_{nj} \geq t\} \leq 2\exp\{-nt\} \leq n^{-3}$. The conclusion is then obvious.

The claims in Lemma 1 are made for each σ in the range of consideration. The bounds can be violated by a zero-probability event for each σ and the union of zero-probability events may have non-zero probability as there are uncountably many σ in the range. Our next lemma strengthens the conclusion in Lemma 1.

Lemma 2. *Except for a zero-probability event not depending on σ, and under the same normal mixture assumption, we have for all large enough n*:
1. *for σ between* $\exp(-2)$ *and* $8/(nM)$, $\sup_\theta[F_n(\theta - \sigma\log(\sigma)) - F_n(\theta)] \leq 4\delta_n(\sigma)$*;*
2. *for σ between 0 and* $8/(nM)$, $\sup_\theta[F_n(\theta - \sigma\log\sigma) - F_n(\theta)] \leq 2(\log n)^2/n$.

Proof. Let $\widetilde{\sigma}_0 = 8/(nM)$, choose $\widetilde{\sigma}_{j+1}$ by $|\widetilde{\sigma}_{j+1}\log\widetilde{\sigma}_{j+1}| = 2|\widetilde{\sigma}_j\log\widetilde{\sigma}_j|$ for $j = 0, 1, 2, \ldots$, and let $s(n)$ be the largest integer such that $\widetilde{\sigma}_{s(n)} \leq \exp(-2)$. Simple algebra shows that $s(n) \leq 2\log n$.

By Lemma 1, for $j = 1, \ldots, s(n)$, we have

$$P\Big\{\sup_\theta[F_n(\theta - \widetilde{\sigma}_j\log\widetilde{\sigma}_j) - F_n(\theta)] \geq 2\delta_n(\widetilde{\sigma}_j)\Big\} \leq 2n^{-2}.$$

Define $D_n = \cup_{j=1}^{s(n)}\{\sup_\theta[F_n(\theta - \widetilde{\sigma}_j\log\widetilde{\sigma}_j) - F_n(\theta)] \geq 2\delta_n(\widetilde{\sigma}_j)\}$. It can be seen that

$$\begin{aligned}\sum_{n=1}^\infty P(D_n) &\leq \sum_{n=1}^\infty\sum_{j=1}^{s(n)} P\Big\{\sup_\theta[F_n(\theta - \widetilde{\sigma}_j\log\widetilde{\sigma}_j) - F_n(\theta)] \geq 2\delta_n(\widetilde{\sigma}_j)\Big\}\\ &\leq \sum_{n=1}^\infty 4n^{-2}\log n < \infty.\end{aligned}$$

By the Borel-Cantelli Lemma, $P(D_n \text{ i.o.}) = 0$. The event D_n is defined for a countable number of σ values, and our next step is to allow all σ.

For each σ in the range of consideration, there exists a j such that $|\widetilde{\sigma}_j\log\widetilde{\sigma}_j| \leq |\sigma\log\sigma| \leq |\widetilde{\sigma}_{j+1}\log\widetilde{\sigma}_{j+1}|$. Hence, almost surely,

$$\sup_\theta[F_n(\theta - \sigma\log\sigma) - F_n(\theta)] \leq \sup_\theta[F_n(\theta - \widetilde{\sigma}_{j+1}\log\widetilde{\sigma}_{j+1}) - F_n(\theta)]$$

$$\leq 2\delta_n(\widetilde{\sigma}_{j+1}) \leq 4\delta_n(\sigma).$$

This proves the first conclusion of the lemma.

With the same $\widetilde{\sigma}_0 = 8/(nM)$, we have

$$P\Big\{\sup_{\theta}[F_n(\theta - \widetilde{\sigma}_0 \log \widetilde{\sigma}_0) - F_n(\theta)] \leq 2n^{-1}(\log n)^2\Big\} \leq n^{-3}.$$

That is, almost surely, $\sup_\theta[F_n(\theta - \widetilde{\sigma}_0 \log \widetilde{\sigma}_0) - F_n(\theta)] \leq 2n^{-1}(\log n)^2$. For $0 < \sigma < 8/(nM)$, we always have

$$\sup_{\theta}[F_n(\theta - \sigma \log \sigma) - F_n(\theta)] \leq \sup_{\theta}[F_n(\theta - \widetilde{\sigma}_0 \log \widetilde{\sigma}_0) - F_n(\theta)],$$

and hence the second conclusion of the lemma follows.

In summary, we have shown that almost surely,

$$\sup_{\theta}\sum_{i=1}^{n} I(|X_i - \theta| < |\sigma \log \sigma|) \leq 8n\delta_n(\sigma), \quad \text{for} \quad \sigma \in \Big[\frac{8}{(nM)}, e^{-2}\Big], \tag{2.2}$$

$$\sup_{\theta}\sum_{i=1}^{n} I(|X_i - \theta| < |\sigma \log \sigma|) \leq 4(\log n)^2, \quad \text{for} \quad \sigma \in \Big(0, \frac{8}{(nM)}\Big]. \tag{2.3}$$

It is worth observing that the normality assumption does not play a crucial role in the proofs. Furthermore, the two factors 8 and 4 in (2.2) and (2.3), respectively, carry no specific meaning; they are chosen for simplicity and could be replaced by any positive numbers.

3. Strong Consistency of the PMLE

We now proceed to prove the consistency of the PMLE for a class of penalty functions. The penalty must be large enough to counter the effect of the observations in a small neighborhood of the location parameters, and small enough to retain the optimal properties of the likelihood method. In addition, we prefer penalty functions that enable efficient numerical computation.

3.1. Conditions on penalty functions

We require the penalty functions to satisfy the following conditions.

C1. $p_n(G) = \sum_{j=1}^{p} \widetilde{p}_n(\sigma_j)$.

C2. $\sup_{\sigma>0} \max\{0, \widetilde{p}_n(\sigma)\} = o(n)$ and $\widetilde{p}_n(\sigma) = o(n)$ at any fixed $\sigma > 0$.

C3. For any $\sigma \in (0, 8/(nM)]$, we have $\widetilde{p}_n(\sigma) \leq 4(\log n)^2 \log \sigma$ for large enough n.

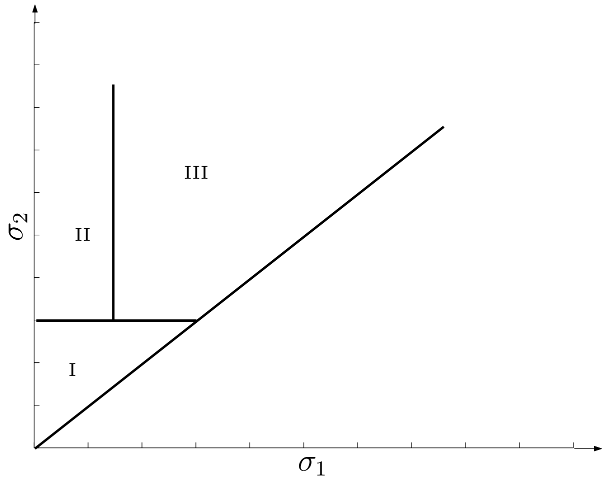

Figure 3.1. Partition of Parameter Space Γ.

When the penalty functions depend on the data, the above conditions are in the sense of almost surely. These three conditions are flexible, and functions satisfying them can easily be constructed. Some examples are given in the simulation section. More specifically, C2 rules out functions that substantially elevate or depress the penalized likelihood at any parameter value. At the same time, C2 allows the penalty to be very severe in a shrinking neighborhood of $\sigma = 0$, C3.

3.2. Consistency of the PMLE when $p = p_0 = 2$

For clarity, we first consider the case where $p = p_0 = 2$. Let $K_0 = E_0 \log f(X; G_0)$, where $E_0(\cdot)$ means expectation with respect to the true density $f(x; G_0)$. It can be seen that $|K_0| < \infty$. Let ϵ_0 be a small positive constant such that

1. $0 < \epsilon_0 < \exp(-2)$;
2. $16M\epsilon_0(\log \epsilon_0)^2 \leq 1$;
3. $-\log \epsilon_0 - (\log \epsilon_0)^2/2 \leq 2K_0 - 4$.

It can easily be seen that as $\epsilon_0 \downarrow 0$, the inequalities are satisfied. Hence, the existence of ϵ_0 is assured. The value of ϵ_0 carries no specific meaning. For some small $\tau_0 > 0$, we define the regions $\Gamma_1 = \{G : \sigma_1 \leq \sigma_2 \leq \epsilon_0\}$, $\Gamma_2 = \{G : \sigma_1 \leq \tau_0, \sigma_2 \geq \epsilon_0\}$, $\Gamma_3 = \Gamma - (\Gamma_1 \cup \Gamma_2)$. See Figure 3.1.

The exact size of τ_0 will be specified later. These three regions represent three situations. One is when the mixing distribution has both scale parameters close to zero. In this case, the number of observations near either one of the location parameters is assessed in the last section; their likelihood contributions are large, but are countered by the penalty so the PMLE has a diminishing probability of being in Γ_1. In the second case, the likelihood has two major sources: the observations near a location parameter with a small scale parameter, and the remaining observations. The first source is countered by the penalty, the likelihood from the second source is not large enough to exceed the likelihood at the true mixing distribution so the PMLE also has a diminishing probability of being in Γ_2.

The following theorem shows that the penalized log-likelihood function is bounded on Γ_1 in some sense.

Theorem 1. *Assume that the random sample is from the normal mixture model with $p = p_0 = 2$, and let $pl_n(G)$ be defined as in* (1.1) *with the penalty function $p_n(G)$ satisfying* C1−C3. *Then $\sup_{G\in\Gamma_1} pl_n(G) - pl_n(G_0) \to -\infty$ almost surely when $n \to \infty$.*

Proof. Let $A_1 = \{i : |X_i-\theta_1| < |\sigma_1 \log\sigma_1|\}$ and $A_2 = \{i : |X_i-\theta_2| < |\sigma_2 \log\sigma_2|\}$. For any index set, say S, we define

$$l_n(G;S) = \sum_{i\in S} \log\left[\frac{\pi}{\sigma_1}\phi\Big(\frac{X_i-\theta_1}{\sigma_1}\Big) + \frac{(1-\pi)}{\sigma_2}\phi\Big(\frac{X_i-\theta_2}{\sigma_2}\Big)\right],$$

hence $l_n(G) = l_n(G;A_1)+l_n(G;A_1^cA_2)+l_n(G;A_1^cA_2^c)$. We investigate the asymptotic order of these three terms. Let $n(A)$ be the number of observations in A. From the fact that the mixture density is no larger than $1/\sigma_1$, we get $l_n(G;A_1) \le -n(A_1)\log\sigma_1$ and, with a slight refinement, we get $l_n(G;A_1^cA_2) \le -n(A_1^cA_2)\log\sigma_2 \le -n(A_2)\log\sigma_2$. By the bounds for $n(A_1)$ and $n(A_1^cA_2)$ given in Lemma 2, almost surely, we have

$$\begin{aligned} l_n(G;A_1) &\le \begin{cases} -4(\log n)^2\log\sigma_1, & 0<\sigma_1\le \dfrac{8}{(nM)},\\ -8\log\sigma_1 + 8Mn\sigma_1(\log\sigma_1)^2, & \dfrac{8}{(nM)} < \sigma_1 < \epsilon_0,\end{cases}\\ l_n(G;A_1^cA_2) &\le \begin{cases} -4(\log n)^2\log\sigma_2, & 0<\sigma_2\le \dfrac{8}{(nM)},\\ -8\log\sigma_2 + 8Mn\sigma_2(\log\sigma_2)^2, & \dfrac{8}{(nM)} < \sigma_2 < \epsilon_0.\end{cases}\end{aligned} \tag{3.1}$$

From (3.1) and condition (C3), we obtain that $l_n(G;A_1) + \widetilde{p}_n(\sigma_1) < 0$ when $0 < \sigma_1 \le 8/(nM)$. When $8/(nM) < \sigma_1 < \epsilon_0$, based on the choice of ϵ_0 we have, almost surely,

$$l_n(G;A_1) + \widetilde{p}_n(\sigma_1) \le 8Mn\sigma_1(\log\sigma_1)^2 - 8\log\sigma_1 \le 8Mn\epsilon_0(\log\epsilon_0)^2 + 9\log n.$$

The two bounds just obtained can be unified as

$$l_n(G;A_1) + \widetilde{p}_n(\sigma_1) \le 8Mn\epsilon_0(\log\epsilon_0)^2 + 9\log n.$$

Similarly, we can show that

$$l_n(G; A_1^c A_2) + \widetilde{p}_n(\sigma_2) \le 8Mn\epsilon_0(\log\epsilon_0)^2 + 9\log n.$$

For observations falling outside both A_1 and A_2, the log-likelihood contributions are bounded by

$$\log\Big\{\pi\sigma_1^{-1}\phi(-\log\sigma_1) + (1-\pi)\sigma_2^{-1}\phi(-\log\sigma_2)\Big\} \le -\log\epsilon_0 - \frac{(\log\epsilon_0)^2}{2},$$

which is negative. At the same time it is easy to show that, almost surely as $n \to \infty$, $n(A_1^c A_2^c) \ge n - \{n(A_1) + n(A_2)\} \ge n/2$. Hence we get the bound $l_n(G; A_1^c A_2^c) \le (n/2)\{-\log\epsilon_0 - (\log\epsilon_0)^2/2\}$.

Combining the three bounds and recalling the choice of ϵ_0, we conclude that when $G \in \Gamma_1$,

$$\begin{aligned} pl_n(G) &= [l_n(G; A_1) + \widetilde{p}_n(\sigma_1)] + [l_n(G; A_1^c A_2) + \widetilde{p}_n(\sigma_2)] + l_n(G; A_1^c A_2^c) \\ &\le 16Mn\epsilon_0(\log\epsilon_0)^2 + \Big(\frac{n}{2}\Big)\Big[-\log\epsilon_0 - \frac{(\log\epsilon_0)^2}{2}\Big] + 18\log n \\ &\le n + \Big(\frac{n}{2}\Big)(2K_0 - 4) + 18\log n \\ &= n(K_0 - 1) + 18\log n. \end{aligned}$$

At the same time, by the Strong Law of Large Numbers, $n^{-1}pl_n(G_0) \to K_0$ almost surely. Hence, $\sup_{G\in\Gamma_1} pl_n(G) - pl_n(G_0) \le -n + 18\log n \to -\infty$ almost surely as $n \to \infty$. This completes the proof.

To establish a similar result for Γ_2, we define

$$g(x; G) = a_1\frac{\pi}{\sqrt{2}}\phi\Big(\frac{x-\theta_1}{\sqrt{2}\sigma_1}\Big) + a_2\frac{(1-\pi)}{\sigma_2}\phi\Big(\frac{x-\theta_2}{\sigma_2}\Big),$$

with $a_1 = I(\sigma_1 \ne 0, \theta_1 \ne \pm\infty)$ and $a_2 = I(\theta_2 \ne \pm\infty)$ for all G in the compacted Γ_2. Note that the first part is not a normal density function as it lacks σ_1 in the denominator of the coefficient. Because of this, the function is well behaved when σ_1 is close to 0 and at 0. It is easy to show that the function $g(x; G)$ has the following properties:

1. $g(x; G)$ is continuous in G almost surely w.r.t. $f(x, G_0)$;
2. $E_0\log\{g(X; G)/f(X; G_0)\} < 0 \;\; \forall G \in \Gamma_2$ (the Jensen inequality);
3. $\sup\{g(x; G) : G \in \Gamma_2\} \le \epsilon_0^{-1}$.

Without loss of generality, we can choose ϵ_0 small enough so that $G_0 \notin \Gamma_2$. Consequently we can easily show, as in Wald (1949), that

$$\sup_{G\in\Gamma_2}\Big\{\frac{1}{n}\sum_{i=1}^{n}\log\Big(\frac{g(X_i; G)}{f(X_i; G_0)}\Big)\Big\} \to -\delta(\tau_0) < 0, \quad \text{a.s.,} \quad \text{as } n \to \infty. \tag{3.2}$$

Note that $\delta(\tau_0) > 0$ is a decreasing function of τ_0. Hence, we can find a τ_0 such that (a) $\tau_0 < \epsilon_0$, and (b) $8M\tau_0(\log\tau_0)^2 \leq 2\delta(\epsilon_0)/5 < 2\delta(\tau_0)/5$. Let τ_0 satisfy these two conditions. Then the PMLE cannot be in Γ_2, as is stated in the following theorem.

Theorem 2. *Assume the conditions of Theorem* 1. *Then* $\sup_{G\in\Gamma_2} pl_n(G) - pl_n(G_0) \to -\infty$ *a.s. as* $n \to \infty$.

Proof. It is easily seen that the log-likelihood contribution of observations in A_1 is no larger than $-\log\sigma_1 + \log g(X_i; G)$. For other observations the log-likelihood contributions are less than $\log g(X_i; G)$. This follows since, if $|x-\theta_1| \geq |\sigma_1 \log\sigma_1|$ and σ_1 is sufficiently small,

$$\frac{1}{\sigma_1}\exp\Big\{-\frac{(x-\theta_1)^2}{2\sigma_1^2}\Big\} \leq \exp\Big\{-\frac{(x-\theta_1)^2}{4\sigma_1^2}\Big\}.$$

Combined with the properties of the penalty function and (3.2), we have

$$\begin{aligned}
&\sup_{\Gamma_2} pl_n(G) - pl_n(G_0)\\
&\quad \leq \sup_{\sigma_1\leq\tau_0}\Big\{\sum_{i\in A_1}\log\Big(\frac{1}{\sigma_1}\Big) + \widetilde{p}_n(\sigma_1)\Big\} + \sup_{\Gamma_2}\sum_{i=1}^{n}\log\Big\{\frac{g(X_i;G)}{f(X_i;G_0)}\Big\} + p_n(G_0)\\
&\quad \leq 8Mn\tau_0(\log\tau_0)^2 + 9\log n - \frac{9\delta(\tau_0)n}{10} + p_n(G_0)\\
&\quad \leq -\frac{\delta(\tau_0)n}{2} + 9\log n + p_n(G_0),
\end{aligned}$$

which goes to $-\infty$ as $n \to \infty$ in view of condition C2 on $p_n(G_0)$.

We now claim the strong consistency of the PMLE.

Theorem 3. *Assume the conditions of Theorem* 1. *For any mixing distribution* $G_n = G_n(X_1,\ldots,X_n)$ *satisfying* $pl_n(G_n) - pl_n(G_0) > c > -\infty$, *we have* $G_n \to G_0$ *a.s. as* $n \to \infty$.

Proof. By Theorems 1 and 2, with probability one, $G_n \in \Gamma_3$ as $n \to \infty$. Confining the mixing distribution G in Γ_3 is equivalent to placing a positive constant lower bound for the variance parameters. Thus, consistency is covered by the result in Kiefer and Wolfowitz (1956). Note that their proof can be modified to accommodate a penalty of size $o(n)$ due to (2.12) of that paper.

Let $\hat{G}_n$ be the PMLE that maximizes $pl_n(G)$. By definition, $pl_n(\hat{G}_n) - pl_n(G_0) > 0$, and therefore $\hat{G}_n \to G_0$ almost surely.

Corollary 1. *Under the conditions of Theorem* 1, *the PMLE* $\hat{G}_n$ *is strongly consistent.*

3.3. Strong consistency of the PMLE when $p = p_0 > 2$

The strong consistency of PMLE for the case where $p_0 > 2$ can be proved in the same manner.

For sufficiently small positive constants $\epsilon_{10} \geq \epsilon_{20} \geq \cdots \geq \epsilon_{p0}$, we partition the parameter space Γ into

$$\Gamma_k = \{G : \sigma_1 \leq \cdots \leq \sigma_{p-k+1} \leq \epsilon_{k0}; \quad \epsilon_{(k-1)0} \leq \sigma_{p-k+2} \leq \cdots \leq \sigma_p\},$$

for $k = 1, \ldots, p$ and $\Gamma_{p+1} = \Gamma - \cup_{k=1}^{p} \Gamma_k$. The choice of ϵ_{k0} $(k = 2, \ldots, p)$ will be given after $\epsilon_{(k-1)0}$ is selected.

Let

$$\begin{aligned} g_k(x; G) = & \sum_{j=1}^{p-k+1} \frac{\pi_j}{\sqrt{2}} \phi\Big(\frac{x - \theta_j}{\sqrt{2}\sigma_j}\Big) I \ \ (\sigma_j \neq 0, \theta_j \neq \pm\infty) \\ & + \sum_{j=p-k+2}^{p} \frac{\pi_j}{\sigma_j} \phi\Big(\frac{x - \theta_j}{\sigma_j}\Big) I \ \ (\theta_j \neq \pm\infty). \end{aligned}$$

We can show that

$$\sup_{G \in \Gamma_k} \left\{ n^{-1} \sum_i \log \Big(\frac{g_k(X_i; G)}{f(X_i; G_0)} \Big) \right\} \to -\delta(\epsilon_{k0}) < -\delta(\epsilon_{(k-1)0}) < 0 \tag{3.3}$$

almost surely as $n \to \infty$. The constants ϵ_{k0} are then chosen so that (a) $\epsilon_{k0} < \epsilon_{(k-1)0}$, and (b) $8(p-k+1)M\epsilon_{k0}(\log \epsilon_{k0})^2 < 2\delta(\epsilon_{(k-1)0})/5$. In this way, $\Gamma_1, \Gamma_2, \ldots$, Γ_p are defined successively. Observe that (3.3) holds since none of $\Gamma_1, \ldots, \Gamma_p$ contains G_0. This fact will be used again later.

The proof of the general case is accomplished in three general steps. First, the probability that the PMLE belongs to Γ_1 goes to zero; this is true because all the σ_k's are small. Second, we show the same for Γ_k, $k = 2, 3, \ldots, p$. Third, when G is confined to Γ_{p+1}, consistency of the PMLE is covered by Kiefer and Wolfowitz (1956).

Step 1. For $k = 1, \ldots, p$, let $A_k = \{i : |X_i - \theta_k| \leq |\sigma_k \log \sigma_k|\}$. For sufficiently small ϵ_{10} and $G \in \Gamma_1$, we have

$$l_n(G; A_1^c A_2^c \cdots A_{k-1}^c A_k) + \tilde{p}_n(\sigma_k) \leq 8M\epsilon_{10}(\log \epsilon_{10})^2 + 9 \log n$$

for $k = 1, \ldots, p$, almost surely. Therefore, the likelihood contribution of the X_i's in $A_1, \ldots, A_p$, plus the penalty term, is

$$\sum_{k=1}^{p} \Big\{ l_n(G; A_1^c A_2^c \cdots A_{k-1}^c A_k) + \tilde{p}_n(\sigma_k) \Big\} \leq 8pM\epsilon_{10}(\log \epsilon_{10})^2 + 9p \log n.$$

At the same time, the total likelihood contributions of the X_i not in $A_1, \ldots, A_p$ are bounded since $l_n(G; A_1^c \cdots A_p^c) \le n/2\{-\log \epsilon_{10} - (\log \epsilon_{10})^2/2\}$. A sufficiently small ϵ_{10} not depending on n can hence be found such that $pl_n(G) - pl_n(G_0) < -n + 9p \log n + p_n(G_0)$ almost surely and uniformly for $G \in \Gamma_1$. The fact that the upper bound goes to $-\infty$ as $n \to \infty$ leads to the conclusion of the first step.

Step 2. The definition of $g_k(x; G)$ is useful in this step. As before, $\sup_{\Gamma_k} E_0 \log\{g_k(X; G)/f(X; G_0)\} < 0$. Hence, using the same idea as for $p_0 = 2$, we get

$$
\begin{aligned}
\sup_{\Gamma_k} pl_n(G) - pl_n(G_0) &\le \sum_{j=1}^{p-k+1} \sup_{\sigma_j < \epsilon_{k0}} \Big[\sum_{i \in A_j} \{-\log \sigma_j\} + \tilde{p}_n(\sigma_j) \Big] \\
&\quad + \sup_{\Gamma_k} \sum_{i=1}^{n} \log \Big\{ \frac{g_k(X_i; G)}{f(X_i; G_0)} \Big\} \\
&\le (p - k + 1) \Big\{ 8Mn\epsilon_{k0} (\log \epsilon_{k0})^2 + 9 \log n \Big\} - \frac{9\delta(\epsilon_{k0}) n}{10} \\
&\le -\frac{\delta(\epsilon_{k0}) n}{2} + 9(p - k + 1) \log n.
\end{aligned}
$$

The last step is a consequence of the choice of these constants.

In conclusion, the PMLE is not in $\Gamma_1, \ldots, \Gamma_p$ except for a zero probability event.

Step 3. Again, confining G to Γ_{p+1} amounts to setting up a positive constant lower bound on σ_k, $k = 1, \ldots, p$. Thus, the consistency proof of the PMLE is covered by Kiefer and Wolfowitz (1956).

In summary, the PMLE of G when $p = p_0 > 2$ is also consistent.

Theorem 4. *Assume that $p_n(G)$ satisfies* C1−C3, *and that $pl_n(G)$ is defined as in* (1.1). *Then for any sequence $G_n = G_n(X_1, \ldots, X_n)$ with $p = p_0$ components satisfying $pl_n(G_n) - pl_n(G_0) > c > -\infty$ for all n, we have $G_n \to G_0$ almost surely.*

3.4. Convergence of the PMLE when $p \ge p_0$.

The exact number of components is often unknown in applications. Thus, it is particularly important to be able to estimate G consistently when only an upper bound p is known. One such estimator is the PMLE with at most p components. Other than in Kiefer and Wolfowitz (1956) and Leroux (1992), whose results do not apply to finite mixture of normal models, there has been limited discussion of this problem.

When $p_0 < p < \infty$, we cannot expect that every part of G converges to that of G_0. We measure their difference by

$$H(G,G_0) = \iint_{\mathcal{R}\times\mathcal{R}^+} |G(\lambda) - G_0(\lambda)| \exp\{-|\theta| - \sigma^2\} d\theta d\sigma^2, \qquad (3.4)$$

where $\lambda = (\theta, \sigma^2)$. It is easily seen that $H(G_n, G_0) \to 0$ implies $G_n \to G_0$ in distribution. An estimator $\hat{G}_n$ is strongly consistent if $H(\hat{G}_n, G_0) \to 0$ almost surely.

Theorem 5. *Under the conditions of Theorem* 1 *except that* $p_0 \le p < \infty$*, for any mixing distribution* $G_n = G_n(X_1, \ldots, X_n)$ *satisfying* $pl_n(G_n) - pl_n(G_0) \ge c > -\infty$*, we have* $H(G_n, G_0) \to 0$ *almost surely as* $n \to \infty$*.*

Most intermediate steps in the proof of consistency of the PMLE when $p = p_0 \ge 2$ are still applicable; some need minor changes. We use many of these results and notations to establish a brief proof.

For an arbitrarily small positive number δ, take $\mathcal{H}(\delta) = \{G : G \in \Gamma, H(G, G_0) \ge \delta\}$. That is, $\mathcal{H}(\delta)$ contains all mixing distributions with up to p components that are at least $\delta > 0$ distance from the true mixing distribution G_0.

Since $G_0 \notin \mathcal{H}(\delta)$, we have $E[\log\{g_k(X;G)/f(X;G_0)\}] < 0$ for any $G \in \mathcal{H}(\delta) \cap \Gamma_k$, $k = 2, 3, \ldots, p$. Thus (3.3) remains valid if revised to

$$\sup_{G\in\mathcal{H}(\delta)\cap\Gamma_k} n^{-1} \sum_{i=1}^{n} \log\left\{\frac{g_k(X_i;G)}{f(X_i;G_0)}\right\} \to -\delta(\epsilon_{k0})$$

for some $\delta(\epsilon_{k0}) > 0$. Because of this, the derivations in Section 3.3 still apply after Γ_k is replaced by $\mathcal{H}(\delta) \cap \Gamma_k$. That is, with proper choice of ϵ_{k0}, we can get $\sup_{G\in\mathcal{H}(\delta)\cap\Gamma_k} pl_n(G) - pl_n(G_0) \to -\infty$ for all $k = 1, \ldots, p$.

With what we have proved, it can be seen that the penalized maximum likelihood estimator of G, $\hat{G}_n$, must almost surely belong to $\mathcal{H}^c(\delta) \cup \Gamma_{p+1}$, where $\mathcal{H}^c(\delta)$ is the complement of $\mathcal{H}(\delta)$. Since δ is arbitrarily small, $\hat{G}_n \in \mathcal{H}^c(\delta)$ implies $H(\hat{G}_n, G_0) \to 0$. On the other hand, $\hat{G}_n \in \Gamma_{p+1}$ is equivalent to putting a positive lower bound on the component variances, which also implies $H(\hat{G}_n, G_0) \to 0$ by Kiefer and Wolfowitz (1956). That is, consistency of the PMLE is also true when $p \ge p_0$.

4. Simulation and a Data Example

In this section, we present some simulation results and a data example.

4.1. The EM algorithm

The EM algorithm is a preferred numerical method in finite mixture models due to its simplicity in coding, and guaranteed convergence to some local maximum under general conditions (Wu (1983)). The EM algorithm can also be easily modified to work with the penalized likelihood method. Often, the penalized log-likelihood function also increases after each EM iteration (Green (1990)) and the algorithm converges quickly.

Let z_{ik} equal 1 when the ith observation is from the kth component, and equal 0 otherwise. The complete observation log-likelihood under the normal mixture model is given by $l_c(G) = \sum_i \sum_k z_{ik}\{\log \pi_k - \log \sigma_k - (2\sigma_k^2)^{-1}(X_i - \theta_k)^2\}$. Given the current parameter value $G^{(m)} = (\pi_1^{(m)}, \ldots, \pi_p^{(m)}, \theta_1^{(m)}, \ldots, \theta_p^{(m)}, \sigma_1^{(m)}, \ldots, \sigma_p^{(m)})$, the EM algorithm iterates as follows.

In the E-Step, we compute the conditional expectation

$$\pi_{ik}^{(m+1)} = E\Big\{z_{ik} \mid \mathbf{x}; G^{(m)}\Big\} = \frac{\pi_k^{(m)}\phi(X_i; \theta_k^{(m)}, \sigma_k^{2(m)})}{\sum_{j=1}^{p} \pi_j^{(m)}\phi(X_i; \theta_j^{(m)}, \sigma_j^{2(m)})}$$

to arrive at

$$\begin{aligned} Q(G; G^{(m)}) &= E\Big\{l_c(G) + p_n(G) \mid \mathbf{x}; G^{(m)}\Big\} \\ &= \sum_{j=1}^{p}(\log \pi_j)\sum_{i=1}^{n}\pi_{ij}^{(m+1)} - \frac{1}{2}\sum_{j=1}^{p}(\log \sigma_j^2)\sum_{i=1}^{n}\pi_{ij}^{(m+1)} \\ &\quad -\frac{1}{2}\sum_{j=1}^{p}\sigma_j^{-2}\sum_{i=1}^{n}\pi_{ij}^{(m+1)}(X_i - \theta_j)^2 + p_n(G). \end{aligned}$$

In the M-step, we maximize $Q(G, G^{(m)})$ with respect to G, and an explicit solution is often possible. For example, when we choose

$$p_n(G) = -a_n\Big\{S_x\sum_{j=1}^{p}(\sigma_j^{-2}) + \sum_{j=1}^{p}\log(\sigma_j^2)\Big\} \tag{4.1}$$

with S_x being a function of the data, $Q(G, G^{(m)})$ is maximized at $G = G^{(m+1)}$ with

$$\begin{cases} \pi_j^{(m+1)} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\pi_{ij}^{(m+1)}, \\ \theta_j^{(m+1)} = \dfrac{\displaystyle\sum_{i=1}^{n}\pi_{ij}^{(m+1)}X_i}{\displaystyle\sum_{i=1}^{n}\pi_{ij}^{(m+1)}}, \\ \sigma_j^{2(m+1)} = \dfrac{2a_nS_x + S_j^{(m+1)}}{\displaystyle\sum_{i=1}^{n}\pi_{ij}^{(m+1)} + 2a_n} \end{cases}$$

where

$$S_j^{(m+1)} = \sum_{i=1}^{n} \pi_{ij}^{(m+1)}(X_i - \theta_j^{(m+1)})^2.$$

It is worth observing here that adding the penalty function (4.1) results in a soft constraint on the scale parameters. From the update formula on σ_j^2, we can easily see that $2a_n S_x/(n + 2a_n) \leq \sigma_j^{2(m)}$. We choose the initial values of the location parameters to be within the data range and then it follows that $\sigma_j^{2(m)} \leq (2a_n S_x + n^2)/2a_n)$. At the same time, from a Bayesian point of view, the penalty function (4.1) puts an Inverse Gamma distribution prior on σ_j^2, where S_x is the mode of the prior distribution or a prior estimate of σ_j^2, and a large value of a_n implies a strong conviction in the prior estimate.

4.2. Simulation results for $p = p_0$

When $p = p_0$, it is meaningful to investigate the bias and variance properties of individual parts of the PMLE. To obtain the results in this section, we generated data from two- and three-component normal mixture models. Two sample sizes, $n = 100$ and $n = 300$, were chosen to examine consistency. We computed the bias and standard deviation of the PMLEs based on 5,000 replicates.

The EM algorithm might miss the global maximum. In our simulation, we used true values as initial values. The EM algorithm was terminated when $\|\lambda^{(m)} - \lambda^{(m+1)}\|_\infty < 5 \times 10^{-6}$, where $\|\mathbf{v}\|_\infty$ denotes the maximal absolute value among the elements of the vector $\mathbf{v}$. We found that the outcomes were satisfactory.

A desirable property of statistical inference for location-scale models is invariance. In this context, given any two real numbers a and b with $a \neq 0$, this means that the PMLE $\tilde{G}$ based on $Y_i = aX_i + b$ and the PMLE $\hat{G}$ based on $X_i, i = 1, \ldots, n$, have the functional relationship $\tilde{G}(a\theta + b, a\sigma) = \hat{G}(\theta, \sigma)$. This is true for the ordinary MLE in general, but is not necessarily true for the PMLE unless we choose our penalty function carefully. For illustration, from a large number of possible penalty functions satisfying conditions C1−C3, we select the penalty functions

P0 $p_n(G) = -\{S_x \sum_{j=1}^{p}(\sigma_j^{-2}) + \sum_{j=1}^{p} \log(\sigma_j^2)\}/n$,

P1 $p_n(G) = -0.4 \sum_{j=1}^{p}(\sigma_j^{-2} + \log \sigma_j^2)$.

Note that C3 is satisfied because when $\sigma < 8/(nM)$, $\sigma^{-2} \approx n^2$. The quantity S_x in P0 is chosen as the sample variance of the observations between two sample quartiles, 25% and 75% in our simulations. Unlike P1, P0 is invariant under location-scale transformations. The choice of the constant 0.4 in P1 is somewhat arbitrary. A sensible choice should depend on n and the (unknown) values of the true variances. Replacing the true variances with an estimate brings us back to P0. In the case of P0, the influence of the penalty is minor when the σ_j's are not close to 0, yet it effectively stops the irregularity. Replacing $1/n$ by $1/\sqrt{n}$ or 1 does not markedly change our simulation results. We included P1 in the simulation to illustrate the importance of the invariance property. For this reason, we computed the PMLEs based on $Y_i = X_i/a$, $i = 1, \ldots, n$, $a = 3.0$, 5.0, 10.0.

In applications, it is a common practice to estimate G with a good local maximum $\hat{G}$ of the likelihood function such that $\hat{\sigma}_j^2 \neq 0$ for all j. Although there are few theoretical guidelines for choosing among the local maxima, we can often identify one that best fits the data by some standard. We regard as the MLE the local maximum located by the EM algorithm with the true mixing distribution as the initial value. When the EM algorithm leads to a local maximum with $\hat{\sigma}_j^2 = 0$ for some j, this outcome is removed; the simulated bias and standard deviation are based on outcomes where no $\hat{\sigma}_j^2 = 0$. The results provide a yardstick for the proposed PMLEs.

Example 1. We consider a two-component normal mixture model with $G_0 = (\pi_0, \theta_{10}, \sigma_{10}^2, \theta_{20}, \sigma_{20}^2) = (0.5, 0, 1, 3, 9)$. The density function of this model has two modes.

The biases and standard deviations (in brackets) of the parameter estimators are presented in Table 1. To make the comparison more sensible, we compute the relative bias and standard deviation of $\hat{\sigma}_j^2$ in terms of $(\hat{\sigma}_j^2 - \sigma_{j0}^2)/\sigma_{j0}^2$ instead of $\hat{\sigma}_j^2 - \sigma_{j0}^2$. The rows marked $\mathrm{P1}^1$, $\mathrm{P1}^2$, $\mathrm{P1}^3$ are the biases and standard deviations of the PMLEs of P1 calculated based on transformed data with $a = 3.0$, 5.0, and 10.0 respectively. In addition, these values were transformed back to the original scale for easy comparisons.

We note that P0 and P1 have similar performance to the MLE and therefore are both very efficient. The PMLE of P1 is not invariant, and its performance worsens as a increases; the invariance consideration is important when selecting appropriate penalty functions. When a decreases, though, the performance of P1 does not deteriorate.

When the sample size increases, all biases and standard deviations decrease, reflecting the consistency of the PMLE. The PMLE based on P1 still suffers from not being invariant but the effect is not as severe.

Probably due to the well separated kernel densities, and the use of the true mixing distribution as initial values, the EM algorithm converged to a reasonable local maximum in all cases.

Example 2. In this example, we choose the two-component normal mixture model with $G_0 = (\pi_0, \theta_{10}, \sigma_{10}^2, \theta_{20}, \sigma_{20}^2) = (0.5, 0, 1, 1.5, 3)$. In contrast to the model used in Example 1, the density function of this model has only one mode.

Table 1. Simulation Results for Example 1, Bias and Standard Deviation.

	$\pi(=0.5)$	$\theta_1(=0)$	$\theta_2(=3)$	$\sigma_1^2(=1)$	$\sigma_2^2(=9)$
	n=100				
MLE	0.052(0.13)	0.038(0.27)	0.519(1.05)	0.126(0.58)	-0.147(0.32)
P0	0.053(0.13)	0.038(0.27)	0.521(1.05)	0.127(0.58)	-0.148(0.32)
P1	0.061(0.13)	0.045(0.24)	0.589(1.09)	0.155(0.60)	-0.179(0.32)
$P1^1$	0.116(0.11)	0.107(0.26)	0.917(1.00)	0.536(0.54)	-0.193(0.32)
$P1^2$	0.167(0.08)	0.184(0.35)	1.188(0.82)	1.072(0.65)	-0.159(0.32)
$P1^3$	0.333(0.15)	0.865(0.56)	1.488(1.07)	5.361(10.7)	2.892(3.97)
	n=300				
MLE	0.015(0.07)	0.006(0.12)	0.145(0.53)	0.027(0.27)	-0.042(0.16)
P0	0.015(0.07)	0.006(0.12)	0.145(0.53)	0.027(0.27)	-0.042(0.16)
P1	0.016(0.07)	0.005(0.12)	0.156(0.54)	0.030(0.27)	-0.050(0.16)
$P1^1$	0.042(0.07)	0.030(0.12)	0.301(0.56)	0.189(0.26)	-0.057(0.18)
$P1^2$	0.079(0.06)	0.074(0.13)	0.534(0.57)	0.446(0.26)	-0.070(0.19)
$P1^3$	0.158(0.05)	0.242(0.14)	0.951(0.47)	1.374(0.36)	-0.011(0.18)

The EM algorithm may not be able to locate a reasonable local maximum. Otherwise, the set-up is the same as in Example 1. The simulation results are presented in Table 2.

The EM algorithm converged to a local maximum with $\hat{\sigma}_j^2 = 0$ in the case of the ordinary MLE 46 of 5,000 times when $n = 100$, even though the true parameter G_0 was used as the initial value; this number was 1 of 5,000 when $n = 300$. We note that the biases and standard deviations decrease when n increases. In general, the precisions of mixing proportion estimators are not high when the two mixing components are close, as documented in Redner and Walker (1984). The performances of $P1^1$, $P1^2$, and $P1^3$ are poor, reaffirming the importance of invariance.

Example 3. In this example, we consider a three-component normal mixture model with

$$\begin{aligned} G_0 &= (\pi_{10}, \theta_{10}, \sigma_{10}^2, \pi_{20}, \theta_{20}, \sigma_{20}^2, \pi_{30}, \theta_{30}, \sigma_{30}^2) \\ &= (0.2, -3.0, 1, 0.5, 0, 0.01, 0.3, 3, 0.5). \end{aligned}$$

The simulation results are presented in Table 3. The performances of the MLE and of the PMLE with P0 or P1 are satisfactory. We note again that the invariance issue is important. Probably due to the well-separated component densities, the EM algorithm converged in all cases.

We remark here that when the component densities are not well separated, much larger sample sizes are needed to achieve precision similar to that in our simulation.

Table 2. Simulation Results for Example 2, Bias and Standard Deviation.

	$\pi(=0.5)$	$\theta_1(=0)$	$\theta_2(=1.5)$	$\sigma_1^2(=1)$	$\sigma_2^2(=3)$
	n=100				
MLE	0.147(0.24)	0.089(0.40)	0.987(1.25)	0.080(0.56)	-0.352(0.44)
P0	0.147(0.24)	0.088(0.40)	0.990(1.25)	0.079(0.56)	-0.354(0.44)
P1	0.173(0.22)	0.105(0.38)	1.108(1.22)	0.116(0.50)	-0.397(0.40)
$P1^1$	0.229(0.17)	0.236(0.33)	0.716(0.73)	0.651(1.15)	0.161(0.57)
$P1^2$	0.395(0.19)	0.571(0.31)	0.705(0.77)	1.887(3.72)	4.007(3.21)
$P1^3$	0.457(0.20)	0.722(0.29)	0.482(1.06)	6.340(19.3)	30.34(7.26)
	n=300				
MLE	0.095(0.20)	0.067(0.23)	0.547(0.87)	0.047(0.38)	-0.192(0.32)
P0	0.095(0.20)	0.067(0.23)	0.548(0.87)	0.047(0.38)	-0.193(0.32)
P1	0.110(0.19)	0.077(0.22)	0.615(0.89)	0.070(0.36)	-0.218(0.31)
$P1^1$	0.163(0.11)	0.129(0.17)	0.566(0.53)	0.264(0.25)	-0.092(0.25)
$P1^2$	0.236(0.11)	0.280(0.20)	0.573(0.36)	0.623(1.09)	0.358(1.10)
$P1^3$	0.474(0.14)	0.706(0.25)	1.214(1.05)	3.495(12.8)	27.21(10.7)

Table 3. Simulation Results for Example 3, Bias and Standard Deviation.

	n=100			n=300		
	$\pi_1(=0.2)$	$\pi_2(=0.5)$	$\pi_3(=0.3)$	$\pi_1(=0.2)$	$\pi_2(=0.5)$	$\pi_3(=0.3)$
MLE	0.000(0.04)	0.001(0.05)	-0.001(0.05)	0.000(0.02)	0.000(0.03)	0.000(0.03)
P0	0.000(0.04)	0.001(0.05)	-0.001(0.05)	0.000(0.02)	0.000(0.03)	0.000(0.03)
P1	-0.002(0.04)	0.002(0.05)	-0.001(0.05)	-0.001(0.02)	0.001(0.03)	0.000(0.03)
$P1^1$	-0.005(0.04)	0.006(0.05)	-0.001(0.05)	-0.002(0.02)	0.002(0.03)	-0.001(0.03)
$P1^2$	-0.004(0.06)	-0.001(0.08)	0.005(0.09)	-0.002(0.02)	0.003(0.03)	-0.001(0.03)
$P1^3$	-0.162(0.19)	-0.500(0.00)	0.662(0.19)	-0.146(0.09)	-0.367(0.22)	0.512(0.31)
	$\theta_1(=-3)$	$\theta_2(=0)$	$\theta_3(=3)$	$\theta_1(=-3)$	$\theta_2(=0)$	$\theta_3(=3)$
MLE	0.005(0.25)	0.000(0.01)	0.001(0.13)	-0.004(0.13)	0.000(0.01)	0.000(0.08)
P0	0.004(0.25)	0.000(0.01)	0.001(0.13)	-0.004(0.13)	0.000(0.01)	0.000(0.08)
P1	-0.016(0.24)	-0.001(0.02)	0.001(0.13)	-0.012(0.13)	0.000(0.01)	0.001(0.08)
$P1^1$	-0.042(0.24)	-0.007(0.02)	0.005(0.13)	-0.027(0.13)	-0.002(0.01)	0.002(0.08)
$P1^2$	0.129(0.46)	-0.028(0.08)	-0.065(0.37)	-2.043(0.21)	-0.003(0.01)	2.010(0.13)
$P1^3$	2.704(0.38)	-0.292(0.39)	-2.696(0.21)	1.970(1.01)	-0.338(0.32)	-2.016(1.16)
	$\sigma_1^2(=1)$	$\sigma_2^2(=0.01)$	$\sigma_3^2(=0.5)$	$\sigma_1^2(=1)$	$\sigma_2^2(=0.01)$	$\sigma_3^2(=0.5)$
MLE	-0.022(0.41)	-0.018(0.21)	-0.044(0.26)	-0.010(0.20)	-0.007(0.12)	-0.013(0.15)
P0	-0.025(0.41)	0.000(0.21)	-0.044(0.26)	-0.011(0.20)	-0.005(0.12)	-0.013(0.15)
P1	-0.067(0.34)	1.600(0.30)	-0.017(0.25)	-0.030(0.19)	0.534(0.13)	-0.006(0.15)
$P1^1$	0.242(0.31)	15.33(2.11)	0.403(0.26)	0.052(0.18)	4.935(0.36)	0.130(0.15)
$P1^2$	1.719(2.72)	94.78(325.)	1.727(2.35)	1.938(0.50)	15.46(1.01)	2.129(0.39)
$P1^3$	95.36(18.2)	$\approx 10^4$(0.01)	17.22(36.4)	73.34(42.6)	$> 10^3(> 10^3)$	7.530(6.12)

4.3. Simulation results for $p \geq p_0$.

In this subsection, we study the properties of the PMLE when $p \geq p_0$, $p_0 = 1$, 2. We generated data from $N(0,1)$ and $0.3N(0,0.1^2)+0.7N(2,1)$, respectively,

Table 4. Number of Degeneracies of the EM Algorithm When Computing the Ordinary MLE.

	$p_0 = 1$			$p_0 = 2$		
n	100	500	2500	100	500	2500
$p = 2$	20	0	0	0	0	0
$p = 3$	209	9	1	16	0	0
$p = 4$	355	33	1	55	0	0
$p = 5$	735	138	23	126	5	0

with $n = 100$, $n = 500$ and $n = 2,500$. In each case, we computed the MLE and the PMLE for $p = p_0, p_0 + 1, \ldots, p_0 + 5$ with penalty function P0. The number of replications was 500.

The EM algorithm was employed to locate the (local) maxima of the $pl_n(G)$. In the EM algorithm, we chose ten initial values; five were in the neighborhood of the true parameter G_0 and the other five were in the neighborhood of some estimates of G_0 without knowledge of p_0. In many cases, the EM algorithm failed to converge when computing the ordinary MLE. A failure was recorded whenever one of the $\hat{\sigma}_j^2$, $j = 1, \ldots, p$, became very large (greater than 10^{32}) or very small (less than 10^{-32}). In all cases, the local maximum (non-degenerate) with the largest likelihood value was considered as the final estimate. The numbers of failures (of 500×10) are given in Table 4. When $n = 100$, $p_0 = 1$, $p = 2$, and $p = 5$, we found two cases where the EM degenerated with all 10 initial values. These were not included in our simulation results.

The distance defined in (3.4) is convenient for theoretical development, but not sensible for measuring the discrepancy between the estimated mixing distribution and the true mixing distribution. To improve the situation, we take a log-transformation on σ^2, and take $H^*(\hat{G}, G_0) = \iint_{[-10,10]\times[-15,5]} |\hat{G}(\lambda) - G_0(\lambda)| d\lambda$ where $\lambda = (\theta, \log \sigma^2)$. This region of integration was chosen because all the PMLEs of θ and $\log \sigma^2$ were within it. The averages of $H^*(\hat{G}, G_0)$ are reported in Table 5.

We first note that it is costly to estimate the mixing distribution with $p = 2$ when $p_0 = 1$. The efficiency of the MLE and the PMLE when $n = 2,500$ is not as good as the MLE with $p = 1$ and $n = 100$. Nevertheless, the mean of $H^*(\hat{G}, G_0)$ apparently decreases when n increases in each case. At the same time, the rate of decrease is not substantial, which might be explained by the result in Chen (1995) that the optimal convergence rate of $\hat{G}$ is at most $n^{-1/4}$ when $p > p_0$.

4.4. A data example.

Liu, Umbach, Peddada, Li, Crockett and Weinberg (2004) analyzed microarray data of the levels of gene expression over time, presented in Bozdech,

Table 5. Average $H^*(\hat{G}, G_0)$ when $p \geq p_0$.

	MLE			PMLE		
n	100	500	2500	100	500	2500
	$p_0 = 1$					
$p = 1$	1.528	0.674	0.305			
$p = 2$	7.331	4.832	2.657	7.348	4.833	2.657
$p = 3$	12.436	8.478	6.409	12.409	8.484	6.404
$p = 4$	17.392	11.546	8.641	17.652	11.514	8.639
$p = 5$	21.360	13.593	9.204	21.243	13.242	9.140
	$p_0 = 2$					
$p = 2$	4.135	1.853	0.851	4.098	1.851	0.851
$p = 3$	8.085	4.658	2.406	8.105	4.856	2.442
$p = 4$	11.601	7.857	4.079	11.690	7.913	4.079
$p = 5$	14.671	11.042	6.673	14.775	11.085	6.732

Table 6. Parameter Estimates for the Data Example.

method	θ_1	θ_2	σ_1^2	σ_2^2	π	$pl_n(\hat{G})$
MLE	38.123	53.057	19.412	21.261	0.676	-8235.8
P0	38.123	53.057	19.412	21.260	0.676	-8235.8
$\sigma_1^2 = \sigma_2^2$	38.200	53.200	19.981	19.981	0.676	-8236.3

Llinas, Pulliam, Wong, Zhu and DeRisi (2003). By employing a random period model, Liu, Umbach, Peddada, Li, Crockett and Weinberg (2004) identified 2,400 cycling transcripts from 3,719 transcripts listed. There is a strong indication that the periods can be modeled by a normal mixture with $p = 2$. By applying a normal mixture model with equal variance, Liu and Chen (2005) found that there is significant evidence for $p = 2$ against $p = 1$, and that the best two-component equal-variance normal mixture model is given by $0.676N(38.2, 4.47^2) + 0.324N(53.2, 4.47^2)$. Figure 5.2 contains the histogram and the density function of the fitted model.

We can also answer the question of whether or not the equal-variance assumption can be justified by testing the hypothesis $H_0 : \sigma_1 = \sigma_2 \leftrightarrow H_1 : \sigma_1 \neq \sigma_2$. We computed the PMLE with penalty P0 as given in Table 6.

It is straightforward that, under the null hypothesis, the penalized likelihood ratio test statistic $R = 2\{\sup_{H_1} pl_n(G) - \sup_{H_0} pl_n(G)\}$ converges in distribution to $\chi^2(1)$. Here $R = 2(8236.3 - 8235.8) = 1.0$ and $P(\chi^2(1) > 1) = 0.317$. Therefore, we have no evidence against the equal-variance assumption.

5. Concluding Remarks.

In this paper, we provide a rigorous proof of the consistency of the penalized MLE, both when the putative number of mixture components $p = p_0$ and when

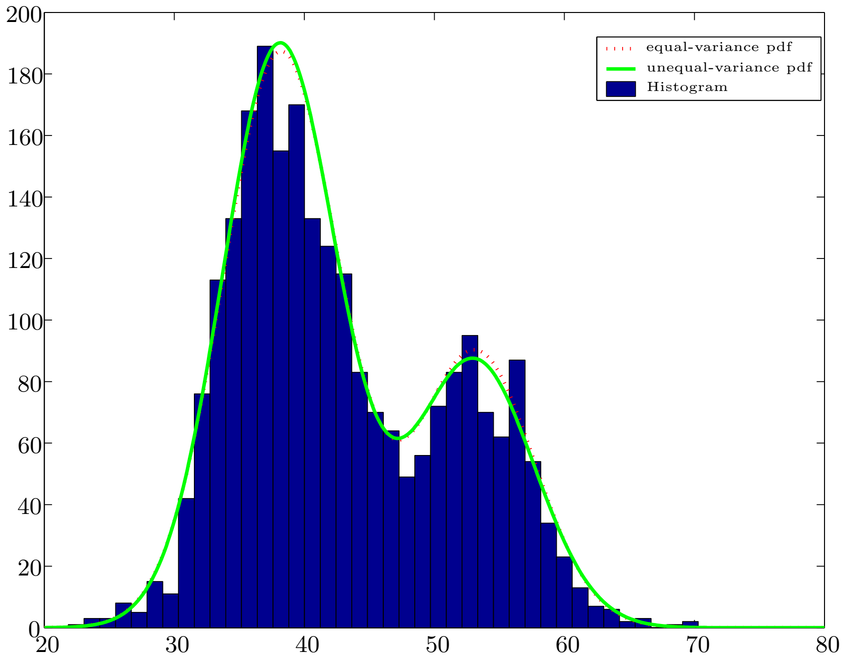

Figure 5.2. The Histogram and Fitted Models of the Data Example.

$p > p_0$. The technique developed could be useful in studying problems of a similar nature such as the consistency of the penalized MLE under a mixture of distributions in location-scale families. The mixture of multivariate normal model is another class of models of practical importance. Its consistency problem remains unsolved, and we believe that further development of our technique may solve this problem.

When $p = p_0$ is known, consistency easily leads to the asymptotic normality of the estimators (Ciuperca, Ridolfi and Idier (2003)). At the same time, the chi-square limiting distribution conclusion for testing equal-component variances is an easy consequence. When p_0 is unknown, the limiting distribution of $\hat{G}$ is not well formulated because of the lack of the corresponding true component parameters in G_0.

Acknowledgement

We would like to thank referees, an associate editor and the Editor for their thoughtful suggestions. This research was supported by a discovery research grant from the NSERC of Canada and a grant from the NSF of China (No. 10601026).

References

Bozdech, Z., Llinas, M., Pulliam, B. L., Wong, E. D., Zhu, J. and DeRisi, J. L. (2003). The transcriptome of the intraerythrocytic developmental cycle of Plasmodium Falciparum. *PLoS Biol.* **1**, 1-16.

Chen, H. and Chen, J. (2003). Test for homogeneity in normal mixtures in the presence of a structural parameter. *Statist. Sinica* **13**, 351-365.

Chen, J. (1995). Optimal rate of convergence for finite mixture models. *Ann. Statist.* **23**, 221-233.

Chen, J. and Kalbfleisch, J. D. (2005). Modified likelihood ratio test in finite mixture models with structural parameter. *J. Statist. Plann. Inference* **129**, 93-107.

Ciuperca, G., Ridolfi, A. and Idier, J. (2003). Penalized maximum likelihood estimator for normal mixtures. *Scand. J. Statist.* **30**, 45-59.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56**, 463-474.

Green, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation. *J. Roy. Statist. Soc. Ser. B* **52**, 443-452.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Ann. Statist.* **13**, 795-800.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27**, 887-906.

Leroux, B. (1992). Consistent estimation of a mixing distribution. *Ann. Statist.* **20**, 1350-1360.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications.* Institute for Mathematical Statistics, Hayward, CA.

Liu, D. and Chen, J. (2005). Heterogeneous periodicities in high-throughput gene expression of synchronized cells. Unpublished manuscript.

Liu, D., Umbach, D. M., Peddada, S. D., Li, L., Crockett, P. W., and Weinberg, C. R. (2004). A random-periods model for expression of cell-cycle genes. *Proc. Natl. Acad. Sci. USA* **101**, 7240-7245.

McLachlan, G. J. and Basford, K. E. (1987). *Mixture Models, Inference and Application to Clustering.* Marcel Dekker, New York.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models.* Wiley, New York.

Pearson, K. (1894). Contribution to the theory of mathematical evolution. *Phil. Trans. Roy. Soc. London A* **186**, 71-110.

Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Ann. Statist.* **9**, 225-228.

Redner, R. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26**, 195-239.

Ridolfi, A. and Idier, J. (1999). Penalized maximum likelihood estimation for univariate normal mixture distributions. In *Actes du* 17^e *colloque GRETSI,* 259-262, Vannes, France.

Ridolfi, A. and Idier, J. (2000). Penalized maximum likelihood estimation for univariate normal mixture distributions. *Bayesian Inference and Maximum Entropy Methods* Maxent Workshops, Gif-sur-Yvette, France, July 2000.

Roeder, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *J. Amer. Statist. Assoc.* **89**, 487-500.

Schork, N., Allison, D. and Thiel, B. (1996). Mixture distributions in human genetics research. *Statist. Meth. Medical Res.* **5**, 155-178.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Tadesse, M., Sha, N. and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *J. Amer. Statist. Assoc.* **100**, 602-617.

Tan, X. M. (2005). Parameter estimation of finite normal mixture models and its application. Unpublished Ph.D. Thesis (in Chinese), School of Mathematical Sciences, Nankai University.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20**, 595-601.

Wu, C. F. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11**, 95-103.

Department of Statistics, University of British Columbia, Vancouver, BC, V6T 1Z2, Canada.

E-mail: jhchen@stat.ubc.ca

LPMC and School of Mathematical Sciences, Nankai University, Tianjin 300071, China.

E-mail: tanxm@nankai.edu.cn

LPMC and School of Mathematical Sciences, Nankai University, Tianjin 300071, China.

E-mail: zhrch@nankai.edu.cn

(Received December 2006; accepted July 2007)