CONSISTENCY OF PENALIZED MLE FOR NORMAL MIXTURES IN MEAN AND VARIANCE

Running Title: Consistency of Estimates in Normal Mixture

By Jiahua Chen, Xianming Tan, Runchu Zhang [1]

University of Waterloo and Nankai University

The finite mixture of normal distributions in both mean and variance parameters have wide applications. It is well known that the likelihood function of this model is unbounded for any given sample size. Hence, the ordinary maximum likelihood estimator is not consistent. At the same time, a local maximum of the likelihood function can often be found to have good statistical properties. In this paper, by introducing a simple penalty function on the component variance parameters, we prove that the penalized maximum likelihood estimator is asymptotically consistent and efficient. The finite sample property of the new estimator is demonstrated through simulations. A genetic data example is also included.

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 it is the most popular model for analysis of quantitative trait loci, see Reoder (1981), 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 which is usually subset of some Euclidean space. The density function of a finite mixture model is given by

$$f(x; G) = \sum_{j=1}^{p} \pi_j f(x; \lambda_j)$$

[1] *AMS 2000 subject Classifications.* Primary 62F10; secondary 62F12.

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

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 univariate normal mixture distribution with λ being mean and variance parameters (θ, σ^2). Let

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2}\}.$$

In normal mixture models, the component density is given by

$$f(x; \theta, \sigma) = \frac{1}{\sigma}\phi(\frac{x-\theta}{\sigma}).$$

The parameter space is conceptually more convenient to be specified in terms of G but technically it is sometimes useful to write it as

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

For convenience, we use G to represent both the mixing distribution and the relevant parameters the mixing distribution contains. We understand that a permutation in the order of the components does not change the model.

Let $x_1, \ldots, x_n$ be n observed values of independent and identically distributed random variables with finite normal mixture distribution $f(x; G)$. A fundamental statistical problem is to estimate the mixing distribution G based on the data. In this study, we assume that p is known so that $\pi_k > 0$ for $k = 1, \ldots, p$ and $(\theta_k, \sigma_k) \neq (\theta_j, \sigma_j)$, for all $k \neq j$, $k, j = 1, \ldots, p$. The first attempt to tackle this problem is attributed to Pearson (1894) which proposed the method of moments for estimating the parameters in the univariate normal mixture. Many other approaches may also be considered, as the ones discussed in McLachlan and Basford (1987) or McLachlan and Peel

(2000). The maximum likelihood estimate (MLE), known for its asymptotic efficiency for regular statistical model, is among the most commonly used approaches (Lindsay, 1995). However, in the case of finite normal mixture distribution in both mean and variance, the MLE is not well defined. Note that the log- likelihood function

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

By letting $\theta_1 = x_1$ and $\sigma_1 \to 0$ with other parameters fixed, it is easily seen that $l_n(G) \to \infty$. It is hence well known that ordinary maximum-likelihood estimate of G is not consistent (Day, 1969; Kiefer and Wolfowitz, 1956)).

In order to avoid this difficulty, authors commonly consider estimates on constrained parameter spaces. For example, Redner (1981) proves that the maximum likelihood estimate of G exists and is globally consistent in every compact sub-parameter space containing the true parameter G_0. Hathaway (1985) proposes to estimate G by maximizing the likelihood function within a restricted parameter space defined by the following constraint:

$$\mathcal{G}_c = \{G : \min_{i,j} \sigma_i/\sigma_j \geq c > 0\}$$

for some constant c. Hathaway's constrained MLE

$$\hat{G}_n = \arg\max_{G \in \mathcal{G}_c} l_n(G)$$

is shown to be strongly consistent provided that the true mixing distribution G_0 belongs to $\mathcal{G}_c$. Despite the elegant results of Redner (1981) and Hathaway (1985), these methods all 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 likelihood as

$$pl_n(G) = l_n(G) + p_n(G) \tag{2}$$

so that $p_n(G) \to -\infty$ as $\min\{\sigma_j : j = 1, \ldots, p\} \to 0$. We then estimate G by the penalized maximum likelihood estimator (PMLE)

$$\widetilde{G}_n = \arg\max_{G} pl_n(G). \tag{3}$$

The penalized likelihood-based method is a promising approach to counter the unboundedness of $l_n(G)$ while keeping the parameter space $\mathcal{G}$ unaltered. However, to make the PMLE work, one has to solve the problem of what kind of penalty functions $p_n(G)$ is eligible.

This task proves to be challenging. Ridolfi and Idier (1999, 2000) propose a class of penalty functions based on a Bayesian conjugate prior distribution, but the asymptotic properties of the corresponding PMLE are not discussed. Under some conditions on $p_n(G)$, Ciuperca et al. (2003) attempt a proof of strong consistency of the PMLE of G under the normal mixture model. The proof however contains a few loose steps which seems hard to be tightened, see Tan (2005).

In this paper, we employ a very different tactic in establishing the strong consistency of the PMLE for a class of penalty functions. In addition, the PMLE is shown asymptotically efficient. The paper is organized as follows. In Section 2, we first present two important technical lemmas. In Section 3, we present detailed proof of strong consistency of the PMLE. The convergence rate and asymptotical efficiency of the PMLE are given in Section 4. In Section 5, we present some simulation results and a real-data example.

2. Technical Lemmas. In a location scale distribution family with parameters θ and σ, the density function at $x = \theta$ increases to infinity as $\sigma \downarrow 0$. When there is only a single observation x_1 from this distribution, the MLE is given by $(\hat{\theta}, \hat{\sigma}) = (x_1, 0)$ which is not sensible. The problem disappears when the number of observations $n \geq 2$ because at this parameter point, while the likelihood contribution of x_1 increases as $\sigma \downarrow 0$, the likelihood contribution of x_2 and others decreases to 0 at a faster rate. In real applications, $n \geq 2$ is generally true and required, the use of MLE is not a problem.

The situation of a mixture of location scale distribution is different. By letting $1 > \pi_1 > 0$, $\theta_1 = x_1$ and $\sigma_1 \downarrow 0$, the likelihood contribution of x_1 increases to infinity. At the same time, the likelihood contributions of $x_2, \ldots, x_n$ do not decrease to 0. Hence, any mixing distribution with $\hat{\theta}_1 = x_1$, $\hat{\sigma}_1 = 0$ and any $0 < \hat{\pi}_1 < 1$ is an MLE of G. Hence, to make the penalized likelihood approach work, one has to use the penalty to counter the effect of observations close to location parameters such as x_1 discussed above. For

this purpose, we need to assess the number of observations fall in a small neighborhood of the location parameters in G.

We first define

$$\Omega_n(\sigma) = \sup_{\theta} \sum_{i=1}^{n} I(0 < x_i - \theta < -\sigma \log \sigma) \tag{4}$$

which is the number of observations fall into the positive side of a small neighborhood of θ. We are only interested in $\Omega_n(\sigma)$ when σ is very small. The number of observations fall into the negative side of θ can be assessed in the same way. Let $F_n(x) = n^{-1} \sum_{i=1}^{n} I(x_i \leq x)$ 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 now define two quantities

$$M = \max\{\sup_{x} f(x; G_0), 8\}, \quad \text{and} \quad \delta_n(\sigma) = -M\sigma \log(\sigma) + n^{-1} \tag{5}$$

where G_0 is the true mixing distribution. The following lemma uses Bahadur's representation to give an order assessment of $n^{-1}\Omega_n(\sigma)$.

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 $8/(nM)$, we have*

$$\sup_{\theta} [F_n(\theta - \sigma \log \sigma) - F_n(\theta)] \leq 2\delta_n(\sigma); \tag{6}$$

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

$$\sup_{\theta} [F_n(\theta - \sigma \log \sigma) - F_n(\theta)] \leq 2(\log n)^2/n. \tag{7}$$

Proof. **1.** Let $\eta_0, \eta_1, \ldots, \eta_n$ be some real numbers such that

$$\eta_0 = -\infty; \; F(\eta_i) = i/n, \quad i = 1, \ldots, n-1; \; \eta_n = \infty.$$

We have

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

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

$$P\{\Delta_{nj} \geq t\} \leq 2\exp\{-\frac{n^2t^2}{2n\delta_n(\sigma) + \frac{2}{3}nt}\}. \tag{8}$$

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

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

By letting $t = \delta_n(\sigma)$ in (8), we obtain

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

Thus for any σ in this range,

$$P\{\max_j \Delta_{nj} \geq \delta_n(\sigma)\} \leq \sum_{j=1}^{n} P\{\Delta_{nj} \geq \delta_n(\sigma)\} \leq 2n^{-2}.$$

Linking this inequality back to $\sup_\theta[F_n(\theta - \sigma\log(\sigma)) - F_n(\theta)]$, we get

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

The conclusion 1 then follows from the Borel-Cantelli lemma.

2. When $0 < \sigma < 8/(nM)$, we choose $t = n^{-1}(\log n)^2$ in (8). For n large enough, $2\delta_n < \frac{1}{3}t$. Hence,

$$P\{\Delta_{nj} \geq t\} \leq 2\exp\{-nt\} \leq n^{-3}.$$

The rest of the proof is similar. □

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 uncountable σ in the range. Our next lemma strengthens the conclusion in Lemma 1.

Lemma 2. *Except for a zero-probability event not depending σ, and under the same normal mixture assumption, we have for all large enough n,*

1. *for each given 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 $\widetilde{\sigma}_0 = 8/(nM)$, and choose $\widetilde{\sigma}_{j+1}$ by

$$|\widetilde{\sigma}_{j+1}\log\widetilde{\sigma}_{j+1}| = 2|\widetilde{\sigma}_j\log\widetilde{\sigma}_j|$$

for $j = 1, 2, \ldots$, and let $s(n)$ in the first integer such that $\widetilde{\sigma}_{s(n)} \geq \exp(-2)$. Simple algebra shows that $s(n) \leq 2\log n$ and $\widetilde{\sigma}_{s(n)} < \exp(-1)$.

By Lemma 1, for $j = 1, 2, \ldots, s(n)$ we have

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

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 is seen that

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

By Borel-Cantelli lemma, D_n almost surely does not occur. The event D_n is defined for a countable number of σ values. Our next step is to allow all σ in the range of consideration.

Since each σ in the range of consideration, there exist 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,

$$\begin{aligned} \sup_\theta [F_n(\theta - \sigma \log \sigma) - F_n(\theta)] &\leq \sup_\theta [F_n(\theta - \widetilde{\sigma}_{j+1} \log \widetilde{\sigma}_{j+1}) - F_n(\theta)] \\ &\leq 2\delta_n(\widetilde{\sigma}_j) \leq 4\delta_n(\sigma). \end{aligned}$$

This proves the first conclusion of the lemma.

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

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

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. □

In summary, we have shown that almost surely,

$$\sup_{\theta} \sum_{i=1}^{n} I(|X_i - \theta| < |\sigma \log \sigma|) \le 8n\delta_n(\sigma), \quad \text{for } \sigma \in [8/(nM), e^{-2}], \tag{9}$$

and

$$\sup_{\theta} \sum_{i=1}^{n} I(|X_i - \theta| < |\sigma \log \sigma|) \le 2(\log n)^2, \quad \text{for } \sigma \in (0, 8/(nM)]. \tag{10}$$

It is worth pointing out that the normality assumption does not play a crucial role in the proofs.

3. Strong consistency of PMLE. We now proceed to prove the consistency of PMLE for a class of penalty functions. The main idea is to make the penalty large enough to counter the effect of the observations in a very small neighborhood of location parameters when the scale parameters decrease. We identify a set of conditions on penalty functions which assure strong consistency of corresponding PMLEs.

3.1 Conditions on penalty functions. To counter the effect of unbounded density function as $\sigma \to 0$, we must have $p_n(G) \to -\infty$ as $\sigma_j \downarrow 0$ for each $j = 1, 2, \ldots, p$. To limit the efficiency loss, $p_n(G)$ should have small derivative at σ not close to zero. Although not necessary, it is a general practice to choose an additive function in σ so that $p_n(G) = \sum_{j=1}^{p} \widetilde{p}_n(\sigma_j)$. Additive penalty functions are convenient for numerical computation by the EM algorithm. Based on these considerations, we require the penalty functions to satisfy

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

C2. $p_n(G_0) = \sum_{j=1}^{p} \widetilde{p}_n(\sigma_{j0}) = o(n)$, where G_0 is the true mixing distribution.

C3. $\widetilde{p}_n(\sigma) \le 4(\log n)^2 \log \sigma$, when $\sigma \le 8/(nM)$ as n is large enough.

These three conditions are flexible and functions satisfying these conditions can be easily constructed. Some examples will be given in the simulation section.

3.2 Consistency of the PMLE when $p = 2$. For the sake of clarity, we begin our proof with two-component normal mixtures. The idea behind our proof has its root in Wald (1949). We first partition the parameter space Γ into three regions, say Γ_1, Γ_2 and Γ_3. We show that with probability 1, the PMLE will not be in the first two regions which does not contain the true mixing distribution. The log-likelihood function can be continuously extended to the compactified Γ_3 and thus Wald's proof can be employed directly to show the strong consistency of PMLE constraint within this region.

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 is seen that $|K_0| < \infty$. Let ϵ_0 be a small positive constant so 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 is easily seen that as $\epsilon_0 \downarrow 0$, the inequalities are satisfied. Hence, the existence of ϵ_0 is assured. The value of ϵ_0 bears no specific meaning. For some small $\tau_0 > 0$, we define three regions as

$$\begin{aligned} \Gamma_1 &= \{G : \sigma_1 \leq \sigma_2 \leq \epsilon_0\}, \\ \Gamma_2 &= \{G : \sigma_1 \leq \tau_0, \sigma_2 \geq \epsilon_0\}, \\ \Gamma_3 &= \Gamma - (\Gamma_1 \cup \Gamma_2). \end{aligned}$$

See Figure 1. The exact size of τ_0 will be specified later. These three regions represent three situations. One is when the mixing distribution having all scale parameters close to zero. In this case, the number of observations near any one of the location parameters is assessed in the last section. Their likelihood contributions are large, but will be countered by the penalty. The contribution of rest of observations are mostly negative. Hence, the PMLE has diminishing probability to be in Γ_1. In the second case, the likelihood has two major sources: the observations near location parameters with small scale parameter, and the remaining observations. The first source is countered by the penalty. We need further to show that the likelihood from the

second source is not large enough to exceed the likelihood at the true mixing distribution. Hence, the PMLE also has diminishing probability to be in Γ_2. Once the possibility of the first two regions is eliminated, the consistency for the PMLE in the third region is the consequence of Wald (1949).

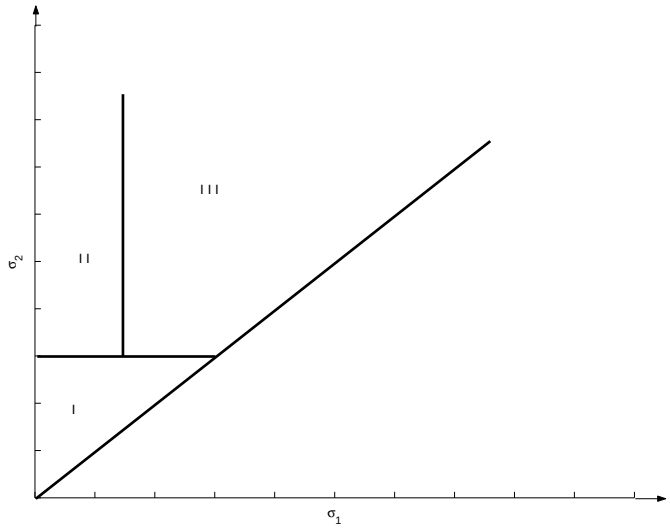

Figure 1: Partition of Γ

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

Theorem 1. *Under the assumption that the random sample is from the normal mixture model with $p = 2$, and let $pl_n(G)$ be defined in (2) with the penalty function $p_n(G)$ satisfying C1-C3. We have that almost surely when $n \to \infty$,*

$$\sup_{G \in \Gamma_1} pl_n(G) - pl_n(G_0) \to -\infty. \tag{11}$$

Proof. Let

$$\begin{aligned} A_1 &= \{i : |x_i - \theta_1| < |\sigma_1 \log \sigma_1|\}, \\ A_2 &= \{i : |x_i - \theta_2| < |\sigma_2 \log \sigma_2|\}. \end{aligned}$$

For any index set, say S, we define

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

hence $l_n(G) = l_n(G; A_1) + l_n(G; A_1^c A_2) + l_n(G; A_1^c A_2^c)$. We now investigate the asymptotic order of these three terms. Let $n(A)$ be the number of samples in set A. From the fact that the mixture density is no larger than $1/\sigma_1$ nor $1/\sigma_2$, we get

$$l_n(G; A_1) \leq -n(A_1) \log \sigma_1,$$

and

$$l_n(G; A_1^c A_2) \leq -n(A_1^c A_2) \log \sigma_2 \leq -n(A_2) \log \sigma_2.$$

By the bounds for $n(A_1)$ and $n(A_1^c A_2)$ given in Lemma 2, except for a zero-probability event, as $n \to \infty$, we have

$$l_n(G; A_1) \leq \begin{cases} -4(\log n)^2 \log \sigma_1, & 0 < \sigma_1 \leq 8/(nM), \\ -8 \log \sigma_1 + 8Mn\sigma_1 (\log \sigma_1)^2, & 8/(nM) < \sigma_1 < \epsilon_0, \end{cases} \tag{12}$$

and

$$l_n(G; A_1^c A_2) \leq \begin{cases} -4(\log n)^2 \log \sigma_2, & 0 < \sigma_2 \leq 8/(nM), \\ -8 \log \sigma_2 + 8Mn\sigma_2 (\log \sigma_2)^2, & 8/(nM) < \sigma_2 < \epsilon_0. \end{cases} \tag{13}$$

From (12), (13), the conditions on the penalty functions, and the way we selected ϵ_0, we arrive at the following bounds:

$$\begin{aligned} l_n(G; A_1) + \widetilde{p}_n(\sigma_1) &\leq 8Mn\sigma_1 (\log \sigma_1)^2 - 8 \log \sigma_1 \leq 8Mn\epsilon_0 (\log \epsilon_0)^2 + 9 \log n; \\ l_n(G; A_1^c A_2) + \widetilde{p}_n(\sigma_2) &\leq 8Mn\sigma_2 (\log \sigma_2)^2 - 8 \log \sigma_2 \leq 8Mn\epsilon_0 (\log \epsilon_0)^2 + 9 \log n. \end{aligned}$$

For observations fall outside both A_1 and A_2, their likelihood contributions are bounded by

$$\log\{\pi \sigma_1^{-1} \phi(-\log \sigma_1) + (1 - \pi)\sigma^{-1} \phi(-\log \sigma_2)\} \leq -\log \epsilon_0 - (\log \epsilon_0)^2/2$$

which is negative. It is easy to show that

$$n(A_1^c A_2^c) \geq n - \{n(A_1) + n(A_2)\} \geq n/2,$$

almost surely. Hence we get the third bound

$$l_n(G; A_1^c A_2^c) \leq (n/2)\{-\log \epsilon_0 - (\log \epsilon_0)^2/2\}.$$

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

$$\begin{aligned} pl_n(G) &= [l_n(G; A_1) + \widetilde{p}_n(\sigma_1)] + [l_n(G; A_1^c A_2) + \widetilde{p}_n(\sigma_2)] + l_n(\gamma | A_1^c A_2^c) \\ &\leq 16Mn\epsilon_0(\log \epsilon_0)^2 + (n/2)[-\log \epsilon_0 - (\log \epsilon_0)^2/2] + 18 \log n \\ &\leq n + (n/2)(2K_0 - 4) + 18 \log n \\ &= n(K_0 - 1) + 18 \log n, \quad \text{a.s.}. \end{aligned}$$

At the same time, by the strong low of large numbers, $n^{-1} pl_n(G_0) \to K_0$ almost surely. Hence, in

$$\sup_{G \in \Gamma_1} pl_n(G) - pl_n(G_0) \leq -n + 18 \log n \to -\infty$$

almost surely as $n \to \infty$. This completes the proof. □

Unlike Γ_1, we have an unbounded Γ_2. Our first step is to compactify it. Define a distance on Γ_2 by

$$d(G, G') = \arctan |\pi - \pi'| + \sum_{i=1}^{2} \arctan |\theta_i - \theta_i'| + \sum_{i=1}^{2} \arctan |\sigma_i - \sigma_i'|.$$

Under this distance, Γ_2 is totally bounded finite dimensional set so it can be compactified. For convenience, we use the same notation Γ_2 for the compactified set. We further define

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

on Γ_2 with $a_1 = I(\sigma_1 \neq 0, \theta_1 \neq \pm\infty)$ and $a_2 = I(\theta_2 \neq \pm\infty)$. For any $G \in \Gamma_2$, let

$$K(G) = E_0 \log g(X; G). \tag{14}$$

Note that for almost all x, $g(x; G)$ is continuous in G. Consequently, $K(G)$ has the following property.

Lemma 3. *For any $\{G_n, n = 1, 2, ...\} \subseteq \Gamma_2$ such that $G_n \to G$,*

$$\overline{\lim}_{n \to \infty} K(G_n) \leq K(G).$$

Proof. For any $\rho > 0$, define

$$g(x; G, \rho) = \sup\{g(x; G') : d(G, G') < \rho, G' \in \Gamma_2\}.$$

Note that $\lim_{\rho \to 0} g(x; G, \rho) = g(x; G)$ and $\sup\{g(x; G) : G \in \Gamma_2\} \leq \epsilon_0^{-1}$. By the dominate convergence theorem,

$$\lim_{\rho \to 0} E_0 \log g(X; G, \rho) = E_0 \log g(X; G) = K(G).$$

Let $\rho_n = d(G, G_n)$, we have $K(G_n) \leq E_0 \log g(X; G, \rho_n)$. Therefore

$$\overline{\lim}_{n \to \infty} K(G_n) \leq \overline{\lim}_{n \to \infty} E_0 \log g(X; G, \rho_n) = E_0 \log g(X; G) = K(G).$$

This complete the proof. □

Due to the compactness of Γ_2, this lemma implies that there exists a $G^* \in \Gamma_2$ such that $K^* = K(G^*) = \sup\{K(G) : G \in \Gamma_2\}$. Let $\delta = \delta(\tau_0) = -E_0 \log\{g(X; G^*)/f(X; G_0)\} = K_0 - K^*$. Since $\sigma_1 < \tau_0$ which will be chosen small,

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

which is a normal mixture density. By Jensen's inequality, we have $\delta(\tau_0) > 0$. Note that $\delta(\tau_0)$ is an decreasing function of τ_0. Hence, we can find a τ_0 such that 1. $\tau_0 < \epsilon_0$ and 2. $8M\tau_0(\log \tau_0)^2 \leq 2\delta(\epsilon_0)/5 < 2\delta(\tau_0)/5$. We will then assume τ_0 satisfies these two conditions. Now we proceed to show that PMLE cannot be in Γ_2 either, as is stated in the following theorem.

Theorem 2. *Assume that $p_n(G)$ satisfies C1-C3 and $pl_n(G)$ is defined as in (2). As $n \to \infty$, we have almost surely that*

$$\sup_{G \in \Gamma_2} pl_n(G) - pl_n(G_0) \to -\infty. \tag{15}$$

Proof. For any $G \in \Gamma_2$,

$$\lim_{\rho \to 0} E_0 \log g(X; G, \rho) = E_0 \log g(X; G) \leq K^*.$$

Hence for each $G \in \Gamma_2$, there exists a $\rho(G) > 0$, such that

$$E_0 \log g(X; G, \rho(G)) < K^* + \delta/10 = K_0 - 9\delta/10.$$

Let $B(G;\rho(G)) = \{G' \in \Gamma_2 : d(G,G') < \rho(G)\}$. Then $B(G;\rho(G))$ form an open cover of Γ_2. From the compactness of Γ_2, there are finite number of G_k, ρ_k with $k = 1, 2, \ldots, K$, such that $\bigcup_{k=1}^{K} B(G_k, \rho_k) = \Gamma_2$. Hence,

$$\sup\Big\{\sum_{i=1}^{n} \log g(X_i;G) : G \in \Gamma_2\Big\} \le \max_k \Big\{\sum_{i=1}^{n} \log g(X_i;G_k,\rho_k)\Big\}.$$

For each k, by the law of large numbers,

$$\sum_{i=1}^{n} \log g(X_i, G_k, \rho_{G_k}) \le n\{K_0 - 9\delta/10 + o(1)\}.$$

Consequently,

$$\sup\Big\{\sum_{i=1}^{n} \log g(X_i;G) : G \in \Gamma_2\Big\} \le n\{K_0 - 9\delta/10 + o(1)\}.$$

The likelihood contribution of observations in A_1 is no larger than $-\log\sigma_1 + \log g(X_i;G)$. For other observations, their likelihood contributions are less than $\log g(X_i;G)$. This is seen by the fact that when $|x-\theta_1| \ge |\sigma_1 \log\sigma_1|$ and σ_1 is sufficiently small,

$$\frac{1}{\sigma_1}\exp\left\{-\frac{(x-\theta_1)^2}{2\sigma_1^2}\right\} \le \exp\left\{-\frac{(x-\theta_1)^2}{4\sigma_1^2}\right\}.$$

Hence, combined with the properties of the penalty function, we have

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

This leads to the conclusion. □

We now claim the strong consistency of PMLE.

Theorem 3. *Assume that $p_n(G)$ satisfies C1-C3 and $pl_n(G)$ is defined as in (2). 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 that $G_n \to G_0$ almost surely as $n \to \infty$.

Proof. By Theorems 1 and 2, with probability one, $G_n \in \Gamma_3$ as $n \to \infty$. Since $\sigma_1 \geq \epsilon_0$ and $\sigma_2 \geq \tau_0$, we can easily compactify Γ_3 and extend the definition of the mixture density function $f(x; G)$ to the compactified Γ_3. Hence, the result is an easy application of Wald (1949). □

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.

3.3 Proof of consistency for general p. The strong consistency of PMLE for the case when $p > 2$ can be proved in the same manner. The only huddle is to produce a clear presentation.

We assume that in the normal mixture model with p components, $\pi_j \neq 0, j = 1, \ldots, p$ and $(\theta_i, \sigma_i^2) \neq (\theta_j, \sigma_j^2)$ whenever $i \neq j$. For p sufficiently small positive constants

$$\epsilon_{10} \geq \epsilon_{20} \geq \cdots \geq \epsilon_{p0},$$

we partition the parameter space Γ into

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

for $k = 1, \ldots, p$ and $\Gamma_{p+1} = \Gamma - \cup_{k=1}^{p} \Gamma_k$.

Similar to the case $p = 2$, $\epsilon_{k0}(\ k = 2, ..., p)$ is determined after $\epsilon_{(k-1)0}$ has been selected. We again compactify Γ_k and use the same notation. Further, on compactified Γ_k, we define

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

As before, the function $K_k(G) = E_0 \log g_k(X; G)$ attains its maximum on the compactified Γ_k^* at some G_k so that $K_k^* = K_k(G_k^*) = \max\{K_k(G) : G \in \Gamma_k^*\}$. Similarly, $\delta_k = K_0 - K_k^* > 0$ and ϵ_{k0} can then be chosen to satisfy two conditions: 1. $\epsilon_{k0} < \epsilon_{(k-1)0}$, and 2. $8(p-k+1)M\epsilon_{k0}(\log \epsilon_{k0})^2 < 2\delta_k/5$. In this way, $\Gamma_1, \Gamma_2, \cdots, \Gamma_p$ are defined one after another.

The proof of the general case can also be done in three general steps. Firstly, we show that the probability of the PMLE belonging to Γ_1 goes to zero. This is the case when all σ_k's are small. Secondly, we show the same for Γ_k, $k = 2, 3, \ldots, p$. Thirdly, the region Γ_{p+1} can be compactified to use the result of Wald (1949) to finish up the proof.

Step 1. For $k = 1, \ldots, p$, define

$$A_k = \{i : |X_i - \theta_k| \leq |\sigma_k \log \sigma_k|\}.$$

For small enough ϵ_0, we have $\sum_{k=1}^p n(A_k) < n/2$ and

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

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

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

At the same time, the total likelihood contributions of X_i not in $A_1, \ldots, A_p$ are bounded as follows:

$$l_n(G; A_1^c \cdots A_p^c) = \frac{1}{2} n\{-\log \epsilon_{10} - (\log \epsilon_{10})^2\}.$$

A sufficiently small ϵ_{10} not depending on n can hence be found such that

$$pl_n(G) - pl_n(G_0) < -n + 9p \log n$$

almost surely and uniformly on Γ_1. This completes the first step.

Step 2. The definition of $g_k(x; G)$ is used in this step. Similar to the case of $p = 2$, for each k, it is seen that

$$\sup_{\Gamma_k} E_0 \log\{g_k(X; G)/f(X; G_0)\} < 0.$$

Hence, using the same idea as for $p = 2$, we get

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

Therefore, the PMLE cannot be in Γ_k for any k except for a zero probability event.

Step 3. We compactify Γ_{p+1} and use the resulf of Wald (1949) to complete the third step.

In summary, the PMLE is consistent.

Theorem 4. *Assume that $p_n(G)$ satisfies C1-C3 and $pl_n(G)$ is defined as in (2). Then for any sequence*

$$G_n = G_n(X_1, \ldots, X_n)$$

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 Possible extensions. Having unbounded likelihood function is not the patent of normal mixture models. Examples include all finite mixture models of location-scale families such as the double exponential distribution and Cauchy distribution families. In each case, the likelihood becomes infinity when a component location parameter equals some observed values and the corresponding scale parameter equals 0. In these mixture models, the maximum likelihood estimator is not consistent. Extension of our results on finite normal mixture models to general mixtures involves no major obstacles other than substantial amount of technical details. Thus, we intend to continue our research and present the result in a future paper.

Another interesting situation is when the order p of the mixture model is unknown. Clearly, this is a more practical problem than the one being discussed. This paper, however, solves an important problem and provides a meaningful step stone for the solution of the more general finite mixture problems. The problem with unknown p can be discussed in a similar fashion.

4. Asymptotic Normality of the PMLE. A typical technique for studying asymptotic distributions is the Taylor's expansion. To make it work, we now place some differentiability conditions on the penalty function $p_n(G)$.

C4 There exists a neighborhood Γ_0 of G_0 such that $p_n(G)$ is continuously differentiable with respect to parameters in G at all $G \in \Gamma_0$ up to order 3 and $\|p_n^{(s)}(G)\| = o(\sqrt{n})$ for $s = 1, 2, 3$.

Theorem 5. *Suppose that $p_n(G)$ satisfies the conditions specified in C1-C4, and $\hat{G}_n = \arg\max_{G \in G} pl_n(G)$, then*

$$\sqrt{n}(\hat{G}_n - G_0) \to N(0, I^{-1}(G_0))$$

in distribution where

$$I(G_0) = \left[E_0 \frac{\partial \log f(X; G_0)}{\partial G} \right]^{\tau} \left[E_0 \frac{\partial \log f(X; G_0)}{\partial G} \right]$$

is the Fisher information matrix.

Proof. From the smoothness of $pl_n(G)$ and the fact that the PMLE is consistent, we must have

$$\frac{\partial pl_n(\hat{G}_n)}{\partial G} = 0.$$

By the Taylor's expansion

$$\begin{aligned} 0 \;=\; & \frac{\partial pl_n(\hat{G}_n)}{\partial G} = \frac{\partial pl_n(G_0)}{\partial G} + \frac{\partial^2 pl_n(G_0)}{\partial G \partial G^{\tau}} (\hat{G}_n - G_0) \\ & + (\hat{G}_n - G_0)^T \frac{\partial^3 pl_n(G^*)}{\partial G^3} (\hat{G}_n - G_0), \end{aligned}$$

where $\frac{\partial^3 pl_n(G^*)}{\partial G^3}$ is a 3-dimensional array with its jth$(j = 1, ..., 3p-1)$ page be a $(3p-1) \times (3p-1)$ matrix whose (k,l)th element equals to

$$\frac{\partial^3 pl_n(G^{*j})}{\partial G_j \partial G_k \partial G_l},\ k, l = 1, ..., 3p-1,$$

where G^{*j} is a mixing distribution between $\hat{G}_n$ and G_0.

From this expansion, we get

$$\left[(\hat{G}_n - G_0)^\tau \frac{\partial^3 pl_n(G^*)}{\partial G^3} + \frac{\partial^2 pl_n(G_0)}{\partial G \partial G^\tau}\right](\hat{G}_n - G_0) = -\frac{\partial pl_n(G_0)}{\partial G}.$$

It is easy to verify that

$$\frac{1}{n}\frac{\partial^3 pl_n(G^*)}{\partial G^3} = O(1),$$

and

$$\frac{1}{n}\frac{\partial^2 pl_n(G_0)}{\partial G \partial G^T} = I(G_0) + o(1).$$

Thus we find

$$\left[(\hat{G}_n - G_0)^\tau \frac{\partial^3 pl_n(G^*)}{\partial G^3} + \frac{\partial^2 pl_n(G_0)}{\partial G \partial G^\tau}\right](\hat{G}_n - G_0) = n\{I(G_0) + o(1)\}(\hat{G}_n - G_0),$$

and

$$\sqrt{n}(\hat{G}_n - G_0) = -\{I^{-1}(G_0) + o(1))\}[\frac{1}{\sqrt{n}}\frac{\partial pl_n(G_0)}{\partial G}].$$

By central-limit theorem,

$$\frac{1}{\sqrt{n}}\frac{\partial pl_n(G_0)}{\partial G} \to N(0, I(G_0))$$

in distribution, which implies

$$\sqrt{n}(\hat{G}_n - G_0) \xrightarrow{L} N(0, I^{-1}(G_0)).$$

□

Since the asymptotic variance is $I^{-1}(G_0)$, Theorem 5 implies that the PMLE is asymptotically efficient.

5. Simulation and Real-data Example. In this section, we will present some simulation results and a real-data example. We also briefly some methods for numerical computation.

5.1 The EM algorithm. The numerical problem associated with the finite mixture model is often be dealt by the famous EM-algorithm. For some choices of penalty function, the EM-algorithm can be easily modified. Often, the penalized log-likelihood function also increases after each EM iteration (Green, 1990) and the algorithm converges as quickly.

Let z_{ik} be the indicator variable such that it equals 1 when the ith observation is from the k component of the normal mixture model, and equals 0 otherwise. The complete observation likelihood is then

$$l_c(G) = \sum_{i=1}^{n}\sum_{k=1}^{p} z_{ik}\left\{\log \pi_k - \log \sigma_k - \frac{(x_i-\theta_k)^2}{2\sigma_k^2}\right\}.$$

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 E-Step, we compute the conditional expectation

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

and arrive at

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

In M-step, we maximize $Q(G, G^{(m)})$ with respect to G. Explicit solution in this step is often possible. For example, if

$$p_n(G) = -\frac{cS_x}{n}\sum_{j=1}^{p}(\sigma_j^{-2}) - \frac{1}{n}\sum_{j=1}^{p}\log(\frac{\sigma_j^2}{S_x})$$

for some constant c, where S_x is certain function depends only on the sample $x_1, \ldots, x_n$,then $Q(G, G^{(m)})$ is maximized at $G = G^{(m+1)}$ such that

$$\begin{cases} \pi_j^{(m+1)} = \dfrac{1}{n}\sum_{i=1}^{n} \pi_{ij}^{(m+1)}, \\ \theta_j^{(m+1)} = \dfrac{\sum_{i=1}^{n} \pi_{ij}^{(m+1)} x_i}{\sum_{i=1}^{n} \pi_{ij}^{(m+1)}}, \\ \sigma_j^{2\,(m+1)} = \dfrac{2cS_x/n + S_j^{(m+1)}}{\sum_{i=1}^{n} \pi_{ij}^{(m+1)} + 2/n} \end{cases}$$

where

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

5.2 Simulation Results. We investigated the bias and variance properties of the PMLE based on simulated data $x_1, \ldots, x_n$ from two-component or three-component normal mixture models. The PMLEs were computed by the EM-algorithm discussed earlier with the true mixing distribution used as the initial value. Two sample sizes, $n = 100$ and $n = 300$ are chosen to examine the consistency. We compute the bias and standard deviation of the PMLEs based on 5000 replicates.

A desirable property of statistical inference for location-scale models is invariance. In this context, given any two real numbers a and b with $a \neq 0$, we require the PMLE $\tilde{G}$ based on $y_i = ax_i + b,\ i = 1, \ldots, n$ 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 necessary for PMLE unless we choose our penalty function carefully. For illustration, among a large number of possible penalty functions satisfying the conditions C1-C3, we select two penalty functions as follows:

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

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

The quantity S_x in P0 is chosen as the sample variance of the observations between two sample quartiles, which are set to be 25% and 75% respectively and $c > 0$ is an adjusting factor which is set to be 0.1 in our simulation. The choice of the quantiles are based more on some intuition. The constant c will not affect the theoretical result, but may have an impact in applications. In these situations, one may try a few values to check the sensitivity. Our simulation indicates that a choice of c anywhere between 0.05 and 10 will not alter our simulation conclusions. The main advantage of $P0$ is its invariance under location-scale transformation. Both P0 and P1 can be motivated from a Bayes point of view: they represent a Gamma prior distribution for σ^2's. The constant 0.4 in P1 is directly adopted from Ciuperca et al. (2003). To illustrate invariance property, we also computed the PMLEs based on $y_i = x_i/10,\ i = 1, \ldots, n$. As expected, the PMLE using P0 is invariant but the PMLE using P1 is not invariant. In our simulation, we transformed the bias and standard deviation of PMLE of P1 based on y_i's back to original scale for easy comparison.

In applications, it is a common practice to estimate G with a local maximum $\hat{G}$ of the likelihood function such that $\hat{\sigma}_j^2 \neq 0$. Although there are few theoretical guidances on choosing among the local maxima if they are not unique, we can often identify one that fits the data well. Thus, we decided to also compute the local maxima as MLE located by EM-algorithm with the true mixing distribution as initial value. It could happen for some data sets that the EM-algorithm leads to the local maximum with $\hat{\sigma}_j^2 = 0$. Nevertheless, we compute the simulated bias and standard deviation based on outcomes when none of $\hat{\sigma}_j^2 = 0$. They at least provide a yard stick for the proposed PMLEs.

Example 1. We generated data $x_1, \ldots, x_n$ from two-component normal mixture model with $G_0 = (\pi_0, \theta_{10}, \sigma_{10}^2, \theta_{20}, \sigma_{20}^2) = (0.5, 0, 1, 3, 9)$. Since the two components $N(0, 1)$ and $N(3, 9)$ are well-separated, the density function of this model has two modes. This example is also used in Ciuperca, et al. (2003).

The simulation results for Example 1 are presented in Table 1.The rows marked as MLE, $P0$, $P1$ and $P1^*$ contain the bias' and standard deviations of the corresponding parameter estimators. The columns are marked for the

parameters to be estimated. The numerical entries of each cell are the bias plus the standard deviation placed in brackets.

Table 1 shows that the PMLEs presented in rows of P0 and P1 have almost the same bias and standard deviation as the revised MLE. This shows that they are both close to efficient. The results in $P1^*$ are very poor. Choosing a smaller penalty, say reduce the constant from 0.4 to 0.04, in the penalty function $P1$ will reduce this devastating effect but the problem persists. Our result shows for the least that the invariance consideration is very important in selecting penalty functions.

We note that when the sample size increases, all bias' and standard deviations decrease indicating the consistency of the PMLE. The PMLE based on P1 still suffer from not being invariant but the effect is not as severe.

Probably due to the well separated kernel densities, by using the true mixing density as initial values, EM-algorithm converged to an reasonable local maximum in all cases in this example.

Table 1: The Bias(Std) of MLE, PMLEs for Two-Component Mixture

Penalty	$\pi(=0.5)$	$\theta_1(=0)$	$\theta_2(=1)$	$\sigma_1^2(=3)$	$\sigma_2^2(=9)$
	n=100				
MLE	.030 (.13)	.136 (.84)	.348 (1.03)	.059 (.487)	-1.15 (2.70)
P0	.030 (.13)	.135 (.84)	.349 (1.03)	.058 (.487)	-1.15 (2.70)
P1	.034 (.14)	.157 (.93)	.386 (1.07)	.073 (.460)	-1.39 (2.72)
P1*	.347 (.12)	.895 (.50)	1.46 (1.20)	4.15 (1.65)	27.2 (36.2)
	n=300				
MLE	.015 (.07)	.006 (.12)	.143 (.527)	.026 (.265)	-.374 (1.44)
P0	.015 (.07)	.006 (.12)	.143 (.527)	.025 (.265)	-.375 (1.44)
P1	.016 (.07)	.005 (.12)	.154 (.541)	.029 (.264)	-.446 (1.47)
P1*	.158 (.05)	.242 (.14)	.951 (.472)	1.37(.355)	-.098 (1.63)

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. It is expected the EM-algorithm may not be able to locate the most reasonable local maximum all the times. Other than this, the rest of the set up is the same as Example 1. The simulation results are presented in Table 2.

As expected, the EM-algorithm converged to a local maximum with $\hat{\sigma}_j^2 = 0$ in the case of ordinary MLE 46 out of 5000 times when $n = 100$, even though the true parameter G_0 was used as the initial values. This number decreases to 1 out of 5000 when $n = 300$. Further, it is noted that under the reduced scale, the penalty in P1 played a strong hand. The simulation indicates that most of the times, the best fitted mixture model has practically only one component. Recall that this distribution has only one mode. Our results clear demonstrate the importance of invariance consideration.

Table 2: The Bias(Std) of MLE, PMLEs for Two-Component Mixture

Penalty	$\pi(=0.5)$	$\theta_1(=0)$	$\theta_2(=1)$	$\sigma_1^2(=3)$	$\sigma_2^2(=9)$
	n=100				
MLE	-.104 (.25)	1.09 (1.83)	-.163 (.896)	-.278 (.427)	-.636 (.994)
P0	-.109 (.25)	1.14 (1.91)	-.174 (.890)	-.289 (.428)	-.642 (.987)
P1	-.108 (.25)	1.26 (1.92)	-.174 (.949)	-.223 (.343)	-.785 (.947)
P1*	.499 (.004)	.749 (.16)	.455 (1.11)	2.30 (.390)	95.1 (10.6)
	n=300				
MLE	-.024 (.21)	.493 (1.25)	.041 (.735)	-.074 (.338)	-.420 (.752)
P0	-.025 (.21)	.496 (1.26)	.041 (.736)	-.075 (.338)	-.421 (.751)
P1	-.021 (.22)	.548 (1.30)	.057 (.777)	-.055 (.311)	-.489 (.759)
P1*	.493 (.02)	.729 (.107)	1.19 (1.12)	1.74(.262)	83.4 (30.0)

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

$$\begin{aligned} G_0 &= (\pi_{10}, \theta_{10}, \sigma_{10}^2, \pi_{20}, \theta_{20}, \sigma_{20}^2, \pi_{30}, \theta_{30}, \sigma_{30}^2) \\ &= (0.5, 0, 0.25, 0.17, 1, 0.25, 0.33, 2, 6.25). \end{aligned}$$

Similar to other two examples, we also examined the invariance property by

computing the PMLEs after scale down the observations by a factor of 10. The simulation results are presented in Table 3. As in Example 2, the EM-algorithm failed to converge to local maxima with $\sigma_j^2 > 0$ for the ordinary MLE 28 and 1 out of 5000 replicates when $n = 100$ and 300 respectively. In general, the PMLE based on P0 works well, and the PMLE based on P1 is also satisfactory, but its performance on the scaled down observations is poor.

Table 3: The Bias(Std) of MLE, PMLEs for Three-Component Mixture

	$\pi_1(=0.5)$	$\pi_2(=0.17)$	$\pi_3(=0.33)$	$\pi_1(=0.5)$	$\pi_2(=0.17)$	$\pi_3=(0.33)$
	n=100			n=300		
MLE	-.011 (.17)	.025 (.16)	-.014 (.10)	-.033 (.17)	.041 (.16)	-.008 (.06)
P0	-.011 (.17)	.024 (.16)	-.013 (.10)	-.032 (.17)	.040 (.16)	-.008 (.06)
P1	-.002 (.19)	.058 (.20)	-.056 (.11)	-.057 (.15)	.087 (.14)	-.030 (.06)
P1*	.442 (.05)	-.112 (.05)	-.330 (.00)	.305 (.03)	.024 (.03)	-.329 (.002)
	$\theta_1(=0)$	$\theta_2(=1)$	$\theta_3(=2)$	$\sigma_1^2(=0.25)$	$\sigma_2^2(=0.25)$	$\sigma_3^2(=6.25)$
	n=100					
MLE	.443 (.943)	-.262 (.927)	.178 (.868)	-.153 (.099)	.128 (.319)	-.416 (2.33)
P0	.429 (.922)	-.234 (1.08)	.174 (.862)	-.157 (.097)	.122 (.308)	-.436 (2.32)
P1	.288 (1.20)	-.042 (1.74)	.364 (1.14)	.069 (.120)	.413 (.344)	-.862 (2.79)
P1*	.630 (.207)	2.64 (1.29)	1.21 (1.33)	2.53 (.669)	44.0 (37.3)	93.7 (.888)
	n=300					
MLE	.416 (.716)	-.326 (.752)	.061 (.421)	-.117 (.102)	.146 (.334)	-.039 (1.35)
P0	.413 (.714)	-.324 (.758)	.062 (.420)	-.118 (.101)	.145 (.333)	-.041 (1.34)
P1	.151 (.573)	-.134 (1.09)	.149 (.565)	.023 (.091)	.348 (.434)	-.031 (1.49)
P1*	.349 (.073)	1.85 (.434)	-.646 (1.54)	.956 (.141)	7.76 (1.54)	90.5 (13.8)

5.3 Real-data Example. Liu et al. (2004) analyze a microarray data of the levels of gene expression over time presented in Bozdech et al. (2003). By employing a random period model, Liu et al. (2004) identify 2400 cycling transcripts from 3719 transcripts listed. There is a strong indication that the

periods can be modeled by a normal mixture with $p = 2$. By applying normal mixture model with equal variance, Liu and Chen (2005) find that there is significant evidence for $p = 2$ against $p = 1$ and 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 2 contain the histogram and the density function of the fitted model.

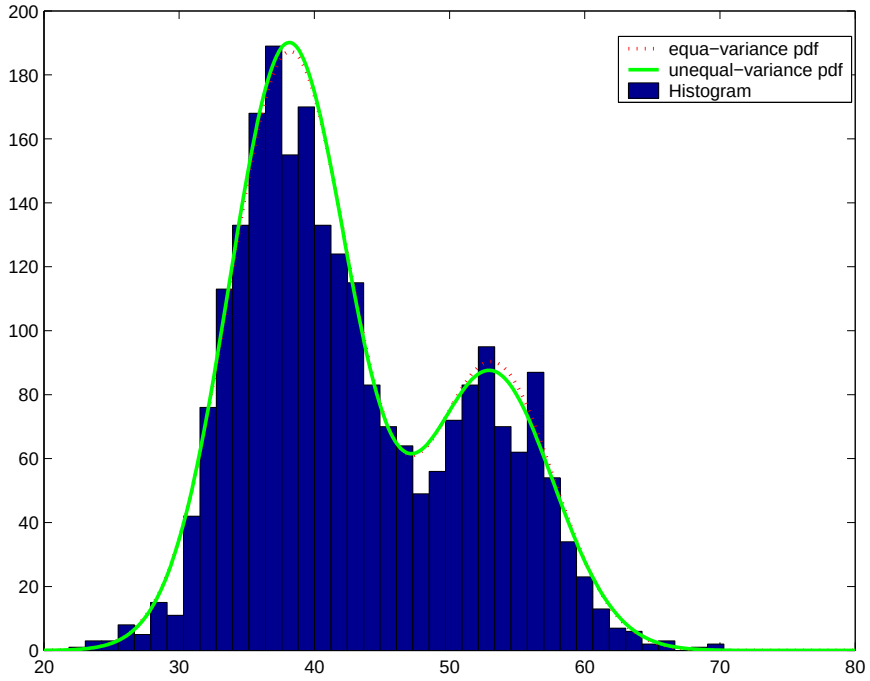

Figure 2: Goodness of fit

We are interested in answering the question whether the equal variance assumption can be justified. We hence test the hypothesis

$$H_0 : \sigma_1 = \sigma_2 \quad \leftrightarrow \quad H_1 : \sigma_1 \neq \sigma_2.$$

We computed the PMLE with the penalty $P0$ as in the following table.

Table 4 PMLE

method	θ_1	θ_2	σ_1^2	σ_2^2	π	$pl_n(\hat{G})$	no. of iter.
MLE	38.123	53.057	19.412	21.261	0.67569	-8235.8	82
P0	38.123	53.057	19.412	21.260	0.67570	-8235.8	77
equi-variance	38.200	53.200	19.981	19.981	0.67600	-8236.3	N/A

It is straightforward that the penalized likelihood ratio test statistic

$$\lambda = 2\left(\sup_{H_1} pl_n(G) - \sup_{H_0} pl_n(G)\right)$$

converges in distribution to $\chi^2(1)$. Here $\lambda = 2(8236.3 - 8235.8) = 1.0$ and $P(\chi^2(1) > 1) = 0.317$. Therefore, we have no evidence against the equal variance assumption.

References

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

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

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

[4] Ciuperca, G., Ridolfi, A. and Idier, J. (2003). Penalized Maximun Likelihood Estimator for Normal Mixtures. *Scand. J. Statist.* **30**, 45-59.

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

[6] Green, P. J. (1990). On Use of the EM Algorithm for Penalized Likelihood Estimation. *J. Roy. Statist. Soc. Ser. B* **52**, 443-452.

[7] Hathaway, R. J. (1985). A Constrained Formulation of Maximun-Likelihood Estimation for Normal Mixture Distributions. *Ann. Statist.* **13**, 795-800.

[8] Kiefer, J. and Wolfowitz, J. (1956). Consistency on the Maximum Likelihood Estimator in the Presence of Infinitly many incidental parameters. *Ann. Math. Statist.* **27**, 887-906.

[9] Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications.* Hayward: Institute for mathematical Statistics.

[10] Liu, D. and Chen, J. (2005). Heterogeneous Periodicities in High-throughput Gene Expression of Synchronized Cells. Unpublished manuscript.

[11] 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.

[12] MaLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models.* Wiley, New York.

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

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

[15] Redner, R. (1981). Note on the Consistency of the Maximum Likelihood Estimate for Nonidentifiable Distributions. *Ann. Statist.* **9**, 225-228.

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

[17] 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.

[18] 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, Juli 2000.

[19] Schork, N., Allison, D., and Thiel, B. (1996). Mixture distributions in human genetics research. *Stat. Methods Med. Res,* **5**, 155-178.

[20] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons.

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

[22] Tan, X. M. (2005). Parameter Estimation of Finite Normal Mixure Models and Its Application. Unpublished Ph.D. Thesis (In Chinese), School of Mathematical Sciences, Nankai University.

[23] Wald, A. (1949). Note on the Consistency of the Maximum Likelihood Estimate. *Ann. Math. Statist.* **20**, 595-601.