CONSISTENCY OF THE CONSTRAINED MAXIMUM LIKELIHOOD ESTIMATOR IN FINITE NORMAL MIXTURE MODELS

Xianming Tan, Jiahua Chen, Runchu Zhang [1]

Nankai University and University of Waterloo

Due to non-regularity of the finite mixture of normal distributions in both mean and variance, the usual maximum likelihood estimator is not well defined for this most important class of mixture models. By requiring the ratio of the variances of any two component distributions being larger than a constant c not depending on sample size, Hathway (1985) established the consistency of the constrained MLE of the mixing distribution. This result, however, can be void when the true distribution does not satisfy the constraint. Thus, it is desirable to let the constant c decreases as the sample size increases to ∞. Two interesting research problems are to determine whether the consistency remains true when c_n decreases to 0 and if so, how fast it is allowed to decrease to 0 as the sample size increases. In this note, we prove the consistency of the new constrained maximum likelihood estimator when c_n is a decreasing sequence satisfying $\log c_n > -k(\log^2 n)$ for any $k > 0$.

1. Introduction. Let the univariate normal density with mean θ and standard deviation σ be denoted as $\varphi(x; \theta, \sigma^2)$. A finite mixture of p univariate normal model in both mean and variance is a class of probability distributions with density function, denoted as $f(x; G)$, defined by

$$f(x; G) = \sum_{j=1}^{p} \pi_j \varphi(x; \theta_j, \sigma_j^2),$$

where $\pi_1, \ldots, \pi_p$ are mixing proportions, $\lambda_j = (\theta_j, \sigma_j^2)$, $j = 1, \ldots, p$ are component parameters, and G is the mixing distribution combining all parameters through

$$G(\lambda) = \sum_{j=1}^{p} \pi_j I(\lambda_j \le \lambda)$$

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

Keywords and phrases. Constrained maximum likelihood, normal mixture model, strong consistency.

with $I(\cdot)$ being the indicator function. The parameter space is conceptually more conveniently specified in terms of G but technically it is sometimes useful to write it as

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

For convenience, we use G to represent both the mixing distribution and the relevant parameters the mixing distribution contains.

Finite mixture models have wide applications in a large number of scientific disciplines. The book by McLachlan and Peel (2000) contains a large collection of application examples. Lindsay (1995), in particular, presented many interesting properties of finite mixture models. Additional references can be found also in Titterington, Smith and Markov (1985). Recently, the finite mixture models become a very popular tool in medical and genetical data analysis, see for examples, Schork, Allison and Thiel (1996) and Tadesse, Sha and Vannucci (2005).

Let $x_1, \ldots, x_n$ be n observed values of independent and identically distributed random variables from a finite normal mixture distribution $f(x; G)$. A fundamental statistical problem is to consistently estimate the mixing distribution G based on the data. In this paper, we consider this problem when 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 $k \neq j,\ k, j = 1, \ldots, p$. Let $l_n(G)$ denote the log-likelihood function based on the random sample $x_1, \ldots, x_n$:

$$l_n(G) = \sum_{i=1}^{n} \log f(x_i, G) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{p} \pi_j \varphi(x_i; \theta_j, \sigma_j^2) \right\}.$$

It is natural to consider estimating G by the maximizer of $l_n(G)$ as in common statistical practice. In this case, however, the likelihood-based method fails due to the unboundedness of $l_n(G)$ on Γ. The unboundedness can be easily seen by noting that when $\theta_1 \to x_1$, $\sigma_1^2 \to 0$, and the rest of parameters remain constant, $l_n(G) \to \infty$. This problem has been noticed for a long time. See also Day (1969), Kiefer and Wolfowitz (1956) and Hathaway (1985). The unboundedness of $l_n(G)$ also causes failures of optimization procedures such as the commonly used EM algorithm (Dempster, Laird and Rubin, 1977) and quasi-Newton type algorithms.

To avoid the difficulty caused by the unboundedness of the likelihood function, various approaches can be used. The classical method of moments was used by Pearson (1894) in fitting the normal mixture model with $p = 2$ to crab data. The methods of moments, however, are known to be less efficient. The moment equations may also produce infeasible solutions. Since the unboundedness of the likelihood function is caused by small values of σ, one class of approaches is to penalize the likelihood at small values of σ. Ciuperca, Ridolfi and Idier (2003) and Chen, Tan and Zhang (2005) discussed a special penalized likelihood approach in that way. Another approach is to place a constraint on the parameter space of G so that the new space is compact. Peters and Walker (1978) and Redner (1981) showed that under certain compact constraints, the constrained MLEs are consistent. Yet, in some applications, compact constraints may appear too restrictive.

Hathaway (1985) proposed a different constraint. A non-compact subset as follows was introduced:

$$\Gamma_c = \{G; \min_{1 \le i,j \le p} (\sigma_i/\sigma_j) \ge c > 0\}$$

with c being a fixed constant. Hathaway's constrained MLE is defined as

$$\hat{G}_n = \arg\max_{G \in \Gamma_c} l_n(G).$$

Note that Γ_c does not indiscriminately rule out any mixing distributions with small value of variances. It was shown that $\hat{G}_n$ is well defined for any given $c > 0$ provided the number of distinct observed values in $x_1, \ldots, x_n$ is larger than p. When the true mixing distribution G_0 belongs to Γ_c, the constrained MLE $\hat{G}_n$ is strongly consistent.

Clearly, the choice of c is not always simple. An improper constant value of c may still exclude the true mixing distribution from Γ_c hence invalidates the consistency results. To this end, Hathaway (1985) asked the following two questions.

1. Is it possible to let c decreases to zero as the sample size increases to infinity while maintaining consistency?
2. If the answer is yes, at what rate can c be decreased to zero?

In this paper, we provide complete answers to these two questions. The consistency is possible when c decreases with n, and the consistency remains true as long as $\log c \ge -k(\log^2 n)$ for arbitrarily fixed $k > 0$.

The paper is organized as follows. In Section 2, we introduce some notations and preparative results mainly taken from Hathaway (1985) and Chen et al. (2005). In Section 3, we state the main result and present our proofs. In Section 4, we provide a brief summary and discussion.

2. Notation and Preliminary Results. We first state a result by Hathaway (1985) regarding the definition of the constrained maximum likelihood estimates.

Theorem 1. *Let* $\{x_1, x_2, ..., x_n\}$ *be a set of observations containing at least* $p+1$ *distinct points. Then for any* c *in* $(0, 1]$*, there exists a constrained global maximizer of* $l_n(G)$ *over* Γ_c*.*

Another important source of useful intermediate results is Chen et al. (2005). The upper bound of the log-likelihood function is crucially related to the number of observations clustered around the location parameter θ in a neighborhood of size σ, or slightly larger, $-\sigma \log \sigma$ when σ is very small. Let $F_n(x) = n^{-1} \sum_{i=1}^{n} I(x_i \leq x)$ be the empirical distribution based on the random sample $x_1, \ldots, x_n$, and $F(x) = E\{F_n(x)\}$. The number of observations close to any fixed θ is then summarized by the quantity $\sup_\theta |F_n(\theta - \sigma \log \sigma) - F_n(\theta)|$. We now introduce a few notations for the purpose of assessing the order of this quantity. Note that the derivative of $F(x)$ is given by $f(x; G_0)$ with G_0 being the true mixing distribution. Let $M = \max\{\sup_{x \in R} f(x; G_0), 8\}$ and $\delta_n(\sigma) = -M\sigma \log(\sigma) + n^{-1}$. It is easily seen that M is finite and $\delta_n(\sigma) \to n^{-1}$ as $\sigma \to 0$. The constants 8 and n^{-1} are needed technically but do not play important rules. The following result is also given in Chen et al. (2005).

Lemma 1. *Except for a zero probability event* E *which is constant in* σ*, we have, as* $n \to \infty$*,*

1. *for all* σ *between* e^{-2} *and* $\frac{8}{nM}$*,*

$$\sup_\theta [F_n(\theta - \sigma \log(\sigma)) - F_n(\theta)] \leq 4\delta_n(\sigma),$$

2. *and for all* σ *between 0 and* $\frac{8}{nM}$*,*

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

Proof. We only provide a scratch proof here, and refer to Chen et. al (2005) for details.

Let $\eta_0 = -\infty;\ \eta_i = F^{-1}(i/n),\ \ i = 1, \ldots, n-1$; and $\eta_n = \infty$. Define

$$\Delta_{nj} = |\{F_n(\eta_j - \sigma \log \sigma) - F_n(\eta_{j-1})\} - \{F(\eta_j - \sigma \log \sigma) - F(\eta_{j-1})\}|,$$

for $j = 1, \ldots, n$. It can be easily verified that

$$\sup_\theta [F_n(\theta - \sigma \log \sigma) - F_n(\theta)] \leq \max_j [\Delta_{nj}] + \delta_n(\sigma). \tag{1}$$

We now prove the lemma in two cases separately.

Case 1. For each $\sigma \in (8/(nM), \exp(-2))$, by the Bernstein inequality (Serfling, 1980, pp 95) and somewhat tedious but straightforward simplification, we can show that

$$P\{\Delta_{nj} \geq \delta_n(\sigma)\} \leq 2n^{-3}.$$

Combined with (1), we get

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

almost surely.

We now tighten up the proof further to show the inequality is valid for all σ with a universal zero-probability event exception.

Let $\widetilde{\sigma}_0 = 8/(nM)$, and choose an increasing sequence $\widetilde{\sigma}_{j+1}$ such that

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

for $j = 1, 2, \ldots$. For 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, we have

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

The upper bound is almost surely less than $2\delta_n(\widetilde{\sigma}_{j+1})$ which is in turn smaller than $4\delta_n(\sigma)$. Since there are only countable j, the event that the inequality fails for at least some σ has zero probability. This proves the first property.

Case 2. For $0 < \sigma < 8/(nM)$, we have $n^{-1}(\log n)^2 > \delta_n(\sigma)$ when n large enough. By the Bernstein inequality again, $P\{\Delta_{nj} \geq n^{-1}(\log n)^2\} \leq n^{-3}$. Thus for all $0 < \sigma < 8/(nM)$,

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

This leads to the second conclusion of the lemma. □

Let us define the following two index sets:

$$\begin{aligned} A &= \{i : |X_i - \theta_1| < |\sigma_1 \log(\sigma_1)|\}, \\ B &= \{i : |X_i - \theta_2| < |\sigma_2 \log(\sigma_2)|\}. \end{aligned}$$

These two sets contain the observations which are the causes of unboundedness of the likelihood in the unconstrained situation for the case $p = 2$. Let $n(A)$ and $n(B)$ be the cardinality of sets A and B. The above lemma implies that for small value of σ_1 such that $\sigma_1 < 8/(nM)$, $n(A) \leq 4 \log^2 n$. For slightly larger σ_1, a non-uniform bound $n(A) \leq 8n\delta(\sigma_1) = -8Mn\sigma_1 \log \sigma_1 + 8$ is applicable. A similar bound applies to $n(B)$.

A key step in our proof is to compactify some unbounded subsets of the parameter space Γ and then consider certain functions on them. To be specific, for $p = 2$ we are going to consider subsets of Γ:

$$\Gamma(\tau) = \{G : \sigma_1 \leq \tau, \sigma_2 \geq \epsilon_0\}$$

with $0 < \tau < \epsilon_0$. Define a distance on $\Gamma(\tau)$ by

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

Under this distance, $\Gamma(\tau)$ is totally bounded finite dimensional set so it can be compactified. For convenience, we use the same notation $\Gamma(\tau)$ for the compactified set.

In spite of the above manipulation, we cannot extend the mixture density to the compactified $\Gamma(\tau)$. Instead, let

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

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

$$K(G) = E_0 \log g(X; G).$$

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

Lemma 2. *For any $\{G_n, n = 1, 2, ...\} \subseteq \Gamma(\tau)$ such that $G_n \to G$, we have*

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

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

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

Note that $\lim_{\rho\to 0} g(x; G, \rho) = g(x; G)$ and $\sup\{g(x; G) : G \in \Gamma(\tau)\} \leq \epsilon_0^{-1}$. 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 $\Gamma(\tau)$, this lemma implies that there exists a $G^* \in \Gamma(\tau)$ such that $K^*(\tau) = K(G^*) = \sup\{K(G) : G \in \Gamma(\tau)\}$. Let

$$\delta = \delta(\tau) = -E_0 \log\{g(X; G^*)/f(X; G_0)\} = K_0 - K^*(\tau).$$

By choosing $\tau < 1$ and using the fact that $\sigma_1 < \tau < \epsilon_0$, we have

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

Using this fact and the Jensen's inequality, we have $\delta(\tau) > 0$. Further, $\delta(\tau)$ is an decreasing function of τ so that we can ensure the value of $\delta(\tau)$ is larger than some positive constant for all small enough τ. Another useful but simple observation is that for any given $\epsilon > 0$,

$$\sup_{G\in\Gamma(\tau)} n^{-1} \sum_{i=1}^{n} \log(g(X_i; G)) < K^*(\tau) + \epsilon \quad \text{a.s. as } n \to \infty \ . \tag{2}$$

With these preparations, we are ready to present our main results in the next section.

3. Main results. For any given $k > 0$, denote

$$\Gamma_n = \left\{ G \in \Gamma \mid \min_{i,j} \sigma_i/\sigma_j \geq \exp(-k \log^2 n) \right\}.$$

It is noted that Γ_n is an increasing subset sequence of the parameter space. When n increases, the true mixing distribution G_0 will eventually be a member of Γ_n. When the parameter space is restricted to Γ_n, the existence of the constraint global maximizer of $l_n(G)$ is guaranteed by Theorem 1 given in the last section (Hathaway, 1985). We now proceed to show that the constrained MLE on Γ_n is strongly consistent.

Theorem 2. *Assume that $X_1, \ldots, X_n$ are independent and identically distributed random variables from a finite mixture of normal distributions with p components, whose density function is $f(x; G_0)$. Let*

$$\bar{G}_n = \bar{G}_n(X_1, \ldots, X_n)$$

be any measurable function of the sample taking values in Γ_n such that for any n and given $a > -\infty$,

$$l_n(\bar{G}_n) - l_n(G_0) \geq a. \tag{3}$$

Then, as $n \to \infty$,

$$\bar{G}_n \to G_0 \text{ almost surely.}$$

Proof. We partition the parameter space into several regions such that the region contains the true mixing distribution is compact. We show that the constrained maximum likelihood estimator, or any estimator satisfying (3), with probability one belongs to the compact region containing the true mixing distribution. Thus, the proof then reduces to the classical result of Wald (1949).

For the sake of simplicity and clarity, we begin with the simplest case when $p = 2$, and will only outline the proof for general p. When $p = 2$, we partition Γ_n into the following three subsets:

$$\begin{aligned}
\Gamma_{n1} &= \{G \in \Gamma_n \,|\, \sigma_1 \leq \sigma_2 \leq \epsilon_0\}, \\
\Gamma_{n2} &= \{G \in \Gamma_n \,|\, \sigma_1 \leq \tau_0, \sigma_2 \geq \epsilon_0\}, \\
\Gamma_{n3} &= \Gamma_n - (\Gamma_{n1} \cup \Gamma_{n2}).
\end{aligned}$$

The constants ϵ_0 and τ_0 in the above partitions are determined when their values become issues in the proof.

The key step in our proof is to provide an order assessment of contributions of various observations to the log-likelihood. For this purpose, for any index set, say S, we define

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

where $\phi(x) = \exp\{-x^2/2\}$, we further denote $n(S)$ as the number of indices belonging to S.

We first consider the case when $G \in \Gamma_{n1}$. Before we move to tedious manipulations, let us point out a few quick facts. First, individual log-likelihood contribution of any observations is no larger than $-\log\sigma_1$. For observations close to θ_2 described by index set B, it is bounded by $-\log\sigma_2$. For observations not close to either θ_1 and θ_2 described by index set A^cB^c, the individual log likelihood contribution is no larger than $-\frac{1}{2}(\log\sigma_2)^2 - \log\sigma_2$. With these facts in mind, we easily see that

$$\begin{aligned} l_n(G) &= l_n(G;A) + l_n(G;A^cB) + l_n(G;A^cB^c) \\ &\le n(A)\log\frac{1}{\sigma_1} + n(A^cB)\log\frac{1}{\sigma_2} + l_n(G;A^cB^c) \\ &= n(A)\log\frac{\sigma_2}{\sigma_1} + n(A\cup B)\log\frac{1}{\sigma_2} + l_n(G;A^cB^c) \\ &\le n(A)\log\frac{\sigma_2}{\sigma_1} + n\log\frac{1}{\sigma_2} - \frac{1}{2}n(A^cB^c)\log^2\sigma_2. \end{aligned} \tag{4}$$

For $0 < \sigma_1 < 8/(nM)$, we have

$$n(A)\log\frac{\sigma_2}{\sigma_1} \le (4\log^2 n) \times (k\log^2 n) = 4k\log^4 n \tag{5}$$

due to the bound on $n(A)$ in Lemma 1 and the constraint $\log(\sigma_2/\sigma_1) \le k\log^2 n$.

By requiring $\epsilon_0 < \exp(-2)$, we have

$$\log(\sigma_2/\sigma_1) \le \log(\epsilon_0/\sigma_1) \le \log(1/\sigma_1).$$

Combining this fact with the other bound for $n(A)$ in Lemma 1, we have

$$n(A)\log\frac{\sigma_2}{\sigma_1} \le 8nM\sigma_1\log^2\sigma_1 + \log n + \log(M/8). \tag{6}$$

The value of ϵ_0 will now be chosen to such

$$\epsilon_0 < \exp(-2),\ 8M\epsilon_0(\log\epsilon_0)^2 \le 1/2,\ \ -\log\epsilon_0 - (\log\epsilon_0)^2/4 \le K_0 - 2. \tag{7}$$

The first two inequality restrictions validate (5) and (6) which in turn imply, for large enough n,

$$n(A)\log\frac{\sigma_2}{\sigma_1} < n, \quad \text{a.s.}$$

We have been very liberal in getting this inequality.

Applying the above inequality to (4) and note that $n(A^cB^c) > n/2$ almost surely, we get

$$l_n(G) \le n(1 - \log\sigma_2 - \log^2\sigma_2/4).$$

Activating the third inequality restriction in (7), and recall that $l_n(G_0) = nK_0 + o(n)$, we have

$$\sup_{\Gamma_{n1}} l_n(G) - l_n(G_0) \le -n \to -\infty \quad \text{a.s. as} \quad n \to \infty.$$

This implies that any $\bar{G}_n$ satisfying (3) almost surely do not belong to Γ_{n1}.

We now consider $G \in \Gamma_{n2}$ which is a subset of

$$\Gamma_2 = \{G \in \Gamma \,|\, \sigma_1 \le \tau_0, \sigma_2 \ge \epsilon_0\}.$$

As in the last section, we can compacify Γ_2 and introduce the function $g(x;G)$ on the compactified Γ_2. Recall that $K(G) = E_0 \log g(X;G)$. By Lemma 2, $K^*(\tau_0) = \sup\{K(G) :\ G \in \Gamma_2\} < K_0$. Furthermore, let $\delta = K_0 - K^*(\tau_0)$, it is seen that $\delta > 0$ and δ is a decreasing function of τ_0. Hence, it is possible to find a small enough τ_0 such that

$$\tau_0 < \min\{\epsilon_0, \exp(-4)\},\ 8M\tau_0(\log\tau_0)^2 \le 2\delta(\epsilon_0)/5 < 2\delta(\tau_0)/5. \tag{8}$$

A value of τ_0 that satisfies (8) will then be selected.

Note that for each observation in A, its likelihood contribution is no larger than $-\log\sigma_1 + \log g(X_i;G)$. For other observations (not in A), 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 less than e^{-4},

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

Consequently,

$$\begin{aligned} l_n(G) &= l_n(G;A) + l_n(G;A^c) \\ &\leq n(A)\log(\frac{1}{\sigma_1}) + \sum_{i=1}^{n} \log g(X_i, G) \end{aligned}$$

Further, we have

$$\log(1/\sigma_1) \leq \log(\sigma_1/\sigma_1) - \log\epsilon_0 \leq k\log^2 n - \log\epsilon_0.$$

Hence, by noting that $8M\tau_0\log^2\tau_0 < 2\delta/5$, for large n, we can show that

$$n(A)\log(\frac{1}{\sigma_1}) \leq 2n\delta(\tau_0)/5 \tag{9}$$

by ignoring a smaller order term $9\log n$. Recall the bound on $\sum_{i=1}^{n}\log g(X_i, G)$ given in (2), we have

$$\sup_{G\in\Gamma_2}\sum_{i=1}^{n}\log g(X_i,G) < n\{K^*(\tau_0) + \delta(\tau_0)/10\} = n\{K_0 - \frac{9}{10}\delta(\tau_0)\} \quad \text{a.s.} \tag{10}$$

Combing (9) and (10), we get

$$\begin{aligned} \sup_{\Gamma_{n2}} l_n(G) - l_n(G_0) &\leq \frac{2}{5}n\delta + n(K_0 - \frac{9}{10}\delta) - nK_0 + o(n) \\ &= -\frac{1}{2}n\delta + o(n) \to -\infty \quad \text{a.s. as} \quad n\to\infty. \end{aligned}$$

That is, any $\bar{G}_n$ satisfying (3) almost surely does not belong to Γ_{n2}.

Based on the above derivations, we now conclude that any $\bar{G}_n$ satisfying (3) must satisfies $\bar{G}(X_1,\ldots,X_n) \in \Gamma_{n3}$ almost surely as n goes to infinity. By noting that $\sigma_2 \geq \sigma_1 \geq \tau_0$ when $G \in \Gamma_{n3}$, the claim that $\bar{G}(X_1,\ldots,X_n)$ is a strong consistent estimator of G_0 is an easy application of the technique used in Wald (1949). This complete the proof for the case $p = 2$.

Proof of consistency for general p can be done in the same manner. We outline the proof below.

Based on p sufficiently small positive constants

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

we partition the parameter space Γ_n into

$$\Gamma_{nl} = \{G \in \Gamma_n :\ \sigma_1 \leq \cdots \leq \sigma_{p-l+1} \leq \epsilon_{l0};\ \epsilon_{(l-1)0} \leq \sigma_{p-l+2} \leq \cdots \leq \sigma_p\},$$

for $l = 1, \ldots, p$ and $\Gamma_{n(p+1)} = \Gamma_n - \cup_{l=1}^{p} \Gamma_{nl}$. Similarly, define

$$\Gamma_l = \{G \in \Gamma : \ \sigma_1 \leq \cdots \leq \sigma_{p-l+1} \leq \epsilon_{l0}; \ \epsilon_{(l-1)0} \leq \sigma_{p-l+2} \leq \cdots \leq \sigma_p\},$$

for $l = 1, \ldots, p$. It is obvious that $\Gamma_{nl} \subseteq \Gamma_l$.

Similar to the case $p = 2$, $\epsilon_{l0}(\ l = 2, \ldots, p)$ will be determined after $\epsilon_{(l-1)0}$ has been selected. The proof starts with compactifying Γ_l and introduce a function similar to g on the compactified space.

The function $K_l(G) = E_0 \log g_l(X; G)$ can be shown to attain its maximum on the compactified Γ_l at some G_l^* so that $K_l^* = K_l(G_l^*) = \max\{K_l(G) : G \in \Gamma_l\}$. Similarly, $\delta_l = K_0 - K_l^* > 0$ and ϵ_{l0} can then be chosen to satisfy two conditions: 1. $\epsilon_{l0} < \epsilon_{(l-1)0}$, and 2. $8(p-l+1)M\epsilon_{l0}(\log \epsilon_{l0})^2 < 2\delta_l/5$. In this way, Γ_{n1}, $\Gamma_{n2}, \cdots, \Gamma_{np}$ are defined one after another.

The proof of the general case can then be done in three general steps. Firstly, we show that the probability of the constrained MLE belonging to Γ_{n1} goes to zero. This is the case when all σ_l's are small. Secondly, we show the same for Γ_{nl}, $l = 2, 3, \ldots, p$. Thirdly, the region $\Gamma_{n(p+1)}$ can be compactified to use the result of Wald (1949) to finish up the proof.

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

$$A_l = \{i : |X_i - \theta_l| \leq |\sigma_l \log \sigma_l|\}.$$

For small enough ϵ_{10}, we have $\sum_{l=1}^{p} n(A_l) < n/2$ and

$$\begin{aligned} l_n(G; A_1^c A_2^c \cdots A_{l-1}^c A_l) &\leq n(A_1^c A_2^c \cdots A_{l-1}^c A_l) \log \frac{1}{\sigma_l} \\ &= n(A_1^c A_2^c \cdots A_{l-1}^c A_l)(\log \frac{\sigma_p}{\sigma_l} + \log \frac{1}{\sigma_p}) \end{aligned}$$

for $l = 1, \ldots, p$. Similar to the case for $p = 2$, for large enough n, we have

$$n(A_1^c A_2^c \cdots A_{l-1}^c A_l) \log \frac{\sigma_p}{\sigma_l} < n(A_l) \log \frac{\sigma_p}{\sigma_l} < n/p.$$

Thus

$$\begin{aligned} l_n(G) &= \sum_{l=1}^{p}\{l_n(G; A_1^c A_2^c \cdots A_{l-1}^c A_l)\} + l_n(G; A_1^c A_2^c \cdots A_p^c) \\ &\leq \sum_{l=1}^{p-1} n(A_l)\log\frac{\sigma_p}{\sigma_l} + n(A_1 \cup A_2 \cup \cdots \cup A_p)\log\frac{1}{\sigma_p} \\ &\quad -n(A_1^c A_2^c \cdots A_p^c)(-\frac{1}{2}\log^2\sigma_p - \log\sigma_p) \\ &\leq n(1 - \log\epsilon_{10} - \frac{1}{4}\log^2\epsilon_{10}). \end{aligned}$$

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

$$l_n(G) - l_n(G_0) < -n$$

almost surely and uniformly on Γ_{n1}. Hence, the maximum likelihood estimate cannot belong to Γ_{n1}. This completes the first step.

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

$$\sup_{\Gamma_{nl}} E_0 \log\{g_l(X; G)/f(X; G_0)\} < 0.$$

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

$$\begin{aligned} \sup_{\Gamma_{nl}} l_n(G) - l_n(G_0) &\leq \sum_{j=1}^{p-l+1} n(A_j)\log\frac{1}{\sigma_j} + \sup_{\Gamma_{nl}}\sum_{i=1}^{n}\log\{g_l(X_i; G)/f(X_i; G_0)\} \\ &\leq -n\delta_l/2. \end{aligned}$$

Therefore, the constrained MLE cannot be in Γ_{nl} for any l except for a zero probability event.

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

In summary, $\bar{G}_n$ is consistent. □

It is beneficial to actually compute the value of $c_n = \exp(-k\log^2 n)$ for some typical choices of k and n. When $n = 100$ and let $k = 2$, this lower bound is of order 10^{-19}. Thus, for the sake of consistency, even a very low value of c_n can be used.

Next, we point out that the constrained maximum likelihood estimator does not suffer from efficiency loss. Let $\hat{G}_n$ denote the global maximizer under the constraint, that is:

$$\hat{G}_n = \arg\max_{G \in \Gamma_n} l_n(G).$$

Theorem 2 clearly implies that $\hat{G}_n$ converge to G_0 almost surely as n goes to infinity.

Theorem 3. *The constrained MLE, $\hat{G}_n$, has the property that as $n \to \infty$,*

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

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. Since the constrained MLE is consistent, it will eventually become an interior point of the parameters space when n becomes large. Hence, we must have

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

By the Taylor's expansion

$$\begin{aligned} 0 \;=\; & \frac{\partial l_n(\hat{G}_n)}{\partial G} = \frac{\partial l_n(G_0)}{\partial G} + \frac{\partial^2 l_n(G_0)}{\partial G \partial G^{\tau}} (\hat{G}_n - G_0) \\ & + \frac{1}{2} (\hat{G}_n - G_0)^{\tau} \frac{\partial^3 l_n(G^*)}{\partial G^3} (\hat{G}_n - G_0), \end{aligned}$$

where $\frac{\partial^3 l_n(G^*)}{\partial G^3}$ 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

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

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

From this expansion, we get

$$\frac{1}{2} \left[(\hat{G}_n - G_0)^{\tau} \frac{\partial^3 l_n(G^*)}{\partial G^3} + \frac{\partial^2 l_n(G_0)}{\partial G \partial G^{\tau}} \right] (\hat{G}_n - G_0) = -\frac{\partial l_n(G_0)}{\partial G}.$$

It is easy to verify that

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

and

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

Thus we find

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

and

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

By central-limit theorem,

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

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 3 implies that the constrained maximum likelihood estimator is asymptotically efficient.

4. Discussion. Ignoring the complex appearance, our proof follows the same general principle in Kiefer and Wolfowitz (1956), Redner (1981) and Hathaway (1985). The most important difference is to assess the influence of potential clusters of observations around the location parameters, particularly when the component variances are allowed to take very small values. This is also a key point in the proof of Hathaway (1985).

The optimization problem associated with the constrained maximum likelihood estimate has been throughly discussed in Hathaway (1985). Whether c depends on n or not does not change the problem. In general, the famous EM can be easily revised to solve the numerical problem. We also refer to Hathaway (1985) for discussions of the finite sample properties of the constrained maximum likelihood estimator in both cases when c is independent or dependent on n. The general idea of this paper applies to other mixture models as well. By far, the normal mixture model is the most important class.

References

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

[2] Chen, J. H. , Tan, X. M. , and Zhang, R. C. (2005). Inference for normal mixture in mean and variance. Manuscript.

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

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

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

[6] 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.

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

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

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

[10] Peters, B.C. and Walker, H.F. (1978). An iterative procedure for obtaining maximum likelihood estiamtion of the parameters for a mixture of normal distributions . *SIAM. J. Appl. Math.* **35**, 362-378.

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

[12] Schork, N. J., Allison, D. B., Thiel, B. (1996). Mixture distributions in human genetics research. *Statistical Methods in Medical Research*, **5** 155-178.

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

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

[15] Titterington, D. M., Smith, A. F. M., and Markov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

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