

Consistency of the MLE under Mixture Models

Jiahua Chen

Abstract. The large-sample properties of likelihood-based statistical inference under mixture models have received much attention from statisticians. Although the consistency of the nonparametric MLE is regarded as a standard conclusion, many researchers ignore the precise conditions required on the mixture model. An incorrect claim of consistency can lead to false conclusions even if the mixture model under investigation seems well behaved. Under a finite normal mixture model, for instance, the consistency of the plain MLE is often erroneously assumed in spite of recent research breakthroughs. This paper streamlines the consistency results for the nonparametric MLE in general, and in particular for the penalized MLE under finite normal mixture models.

Key words and phrases: Nonparametric MLE, identifiability, Kiefer–Wolfowitz approach, penalized MLE, Pfanzagl approach.

1. INTRODUCTION

A family of distributions, parametric or nonparametric, is regarded as a probability or statistical model. A parametric model can hence be a family of density functions in the form $\{f(x; \theta) : \theta \in \Theta\}$ where Θ is a subset of \mathbb{R}^d for some positive integer d . The measure with respect to which these densities are defined is μ , but this will be de-emphasized. In applications, statisticians must select one or a set of these distributions to characterize the uncertainty displayed in a random sample from a population.

A specific distribution family or model is often tentatively chosen in accordance with the scientific background of the application. For instance, the Poisson distribution/model is a textbook choice for the number of annual car accidents of a policy holder, but it may not be ideal because risk levels vary. It can therefore be helpful to divide the population into several subpopulations, each modeled by its own Poisson distribution. Thus, a finite mixture of Poisson distributions is a better choice for the pooled accident data. This leads to a

generic finite mixture model in the form

$$(1.1) \quad f(x; G) = \sum_{j=1}^m \alpha_j f(x; \theta_j).$$

In this formulation, $f(x; G)$ is a finite convex combination of the component density functions $f(x; \theta)$. Each $f(x; \theta_j)$ is the density for a subpopulation. The recipe for the mixture is summarized by G , specified by its cumulative distribution function (c.d.f.)

$$(1.2) \quad G(\theta) = \sum_{j=1}^m \alpha_j I(\theta_j \leq \theta),$$

where $I(\cdot)$ is an indicator function. The model $f(x; G)$ in the above definition has order m , even if some θ_j are equal or some $\alpha_j = 0$. A finite mixture model of order m is hence also a degenerate order $m + 1$ mixture.

For any c.d.f. G on Θ , the following integral

$$(1.3) \quad f(x; G) = \int f(x; \theta) dG(\theta)$$

is a well-defined density function. When G has only a finite number of support points, as in (1.2), $f(x; G)$ reduces to the finite mixture model (1.1). When the form of G is unspecified, (1.3) is regarded as a nonparametric mixture model. The space of all the mixing distributions on Θ is denoted \mathbb{G} , and the space of those with at most m support points is \mathbb{G}_m . It can

Jiahua Chen is Professor of Statistics, Research Institute of Big Data, Yunnan University, Kunming, 650021, China and Department of Statistics, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada (e-mail: jhchen@stat.ubc.ca).

be seen that $\mathbb{G}_m \subset \mathbb{G}_{m+1}$. See [Titterington, Smith and Makov \(1985\)](#), [Lindsay \(1995\)](#) and [McLachlan and Peel \(2000\)](#) for the general theory and applications of mixture models.

Research into mixture models has a long history. The most cited early publication is [Pearson \(1894\)](#); he used a two-component normal mixture model for a biometric data set. This model was motivated by the apparent skewness in the ratio of forehead-width to body-length of 1000 crabs from Naples. It was suspected that the population contained two unidentified species. [Pearson \(1894\)](#) employed the method of moments for the parameter estimation and provided a detailed account of the numerical calculation in the absence of the modern computer. Nowadays, we would estimate the parameters by the maximum likelihood estimator (MLE). The MLE is favored for its asymptotic efficiency under regular parametric models. The unpleasant numerical obstacle is now a history.

Mixture distributions form a distinct class of non-regular statistical models. They are notorious for presenting statisticians with serious technical challenges as well as many pleasant surprises. Given an i.i.d. sample of size n from a mixture model, [Hartigan \(1985\)](#) showed that the likelihood ratio test statistic for homogeneity is stochastically unbounded defying the classical chi-squared limiting distribution. [Chen \(1995\)](#) showed that the optimal rate for estimating the mixing distribution is $O_p(n^{-1/4})$ as compared with $O_p(n^{-1/2})$ for parameter estimation under regular models. [Quinn, McLachlan and Hjort \(1987\)](#) interpreted some of these abnormal results by the degenerated Fisher information when the order of the finite mixture model is unknown. At the same time, [Lindsay \(1983\)](#) showed that the nonparametric MLE of G has at most k support points, the number of distinct observed values. Given the order of the finite mixture model and an initial mixing distribution $G^{(0)}$, [Wu \(1983\)](#) found that the famous EM-algorithm leads to a sequence of $G^{(k)}$ that converges to a local maximum of the likelihood function and it can be a locally consistent MLE [Peters and Walker \(1978\)](#). Both [Lindsay \(1983\)](#) on the geometric property and [Wu \(1983\)](#) on the algorithmic convergence place only nominal conditions on $f(x; G)$.

There is also good news about the large-sample properties of the MLE. [Kiefer and Wolfowitz \(1956\)](#) and [Pfanzagl \(1988\)](#) proved that the nonparametric MLE of G is strongly consistent in an i.i.d. setting. Using quotient topology, [Redner \(1981\)](#) gave another consistency proof under finite mixture models.

These consistency results are encouraging, but careful discussions of the relationships between these proofs are lacking. Why are there multiple proofs for a single consistency result? How do these proofs relate to each other? Without a full understanding, researchers may cite a paper when its specific proof/result is not applicable to the target problem. This paper streamlines the consistency proofs of the nonparametric MLE under mixture models and crystallizes their applicability. Another topic is consistent estimation under the finite normal mixture model: the consistency of the MLE under this model is often erroneously assumed. This paper aims to popularize the consistent estimator based on the penalized likelihood. The ultimate goal of this exercise is to provide a solid basis for research into the large-sample properties of mixture models.

The remainder of the paper consists of two major sections. Section 2 addresses classical consistency results for the nonparametric MLE under a mixture model. Section 3 addresses results related to the consistency of the penalized (or regularized) MLE under the finite normal mixture model. Section 4 provides concluding remarks.

2. CLASSICAL CONSISTENCY RESULTS

Let x_1, \dots, x_n be a set of i.i.d. observations of size n from a nonparametric mixture model $\{f(x; G) : G \in \mathbb{G}\}$. The likelihood function of G is given by

$$L_n(G) = \prod_{i=1}^n f(x_i; G).$$

The log-likelihood function of G is hence

$$(2.1) \quad \ell_n(G) = \sum_{i=1}^n \log f(x_i; G).$$

Both $L_n(G)$ and $\ell_n(G)$ are functions defined on \mathbb{G} . The nonparametric MLE \hat{G}_n of G is a c.d.f. on Θ such that

$$(2.2) \quad \ell_n(\hat{G}) = \sup\{\ell_n(G) : G \in \mathbb{G}\}.$$

Most rigorously, \hat{G} is related to the sample in a measurable fashion, and it is potentially one of many possible global maximum points of $\ell_n(\cdot)$. In addition, \hat{G} can be a limiting point of a sequence of mixing distributions G_j such that $\ell_n(G_j) \rightarrow \sup\{\ell_n(G) : G \in \mathbb{G}\}$ as $j \rightarrow \infty$.

For ease of presentation, (2.2) is regarded as an unambiguous definition of MLE. When G is confined to \mathbb{G}_m , this becomes the MLE of G under finite mixture

models. When $m = 1$ in \mathbb{G}_m , (2.2) defines the ordinary parametric MLE of θ under the model $\{f(x; \theta) : \theta \in \Theta\}$.

As acknowledged by Kiefer and Wolfowitz (1956), their proof of the consistency of \hat{G} is a technical clone of that of Wald (1949). One key technical preparation for this proof is the following well-known inequality.

LEMMA 2.1 (Jensen's inequality). *Let X be a random variable such that $E|X| < \infty$ and let $\varphi(t)$ be a convex function. Then*

$$E\{\varphi(X)\} \geq \varphi(E(X)).$$

Suppose X is a random variable with density function $f(x; \theta)$ in general and θ^* is the true value of the parameter. Let $Y = f(X; \theta)/f(X; \theta^*)$. It can be seen that

$$\begin{aligned} E^*(Y) &= \int \{f(x; \theta)/f(x; \theta^*)\} f(x; \theta^*) d\mu(x) \\ &\leq \int f(x; \theta) d\mu(x) = 1, \end{aligned}$$

where E^* is the expectation with respect to the distribution with the specific θ^* parameter value. Applying Jensen's inequality to Y and $\varphi(t) = -\log(t)$, we get

$$(2.3) \quad \begin{aligned} -E^* \log\{f(X; \theta)/f(X; \theta^*)\} \\ \geq -\log[E^*\{f(X; \theta)/f(X; \theta^*)\}] \geq 0. \end{aligned}$$

If $E^*|\log f(X; \theta^*)| < \infty$, then

$$E^*\{\log f(X; \theta)\} \leq E^*\{\log f(X; \theta^*)\}.$$

The equality holds if and only if $f(x; \theta) \equiv f(x; \theta^*)$, or it is equal except for a zero-probability set of x in terms of $f(x; \theta^*)$. The expectation $E^*[\log\{f(X; \theta^*)/f(X; \theta)\}]$ is referred to as the Kullback–Leibler information between the two distributions.

The following lemma may seem trivial, but it is the basis of most proofs for generic consistency.

LEMMA 2.2 (Trivial consistency of MLE). *Suppose the model under investigation is $\{f(x; \theta), \theta \in \Theta\}$ where $\Theta = \{\theta^*, \theta_1, \dots, \theta_M\}$. In addition, $f(x; \theta) = f(x; \theta^*)$ except for a zero-probability set of x with respect to $f(x; \theta^*)$ implies $\theta = \theta^*$.*

Then the MLE $\hat{\theta}_n$ of θ equals θ^ almost surely as $n \rightarrow \infty$.*

PROOF. Let X be a random variable from $f(x; \theta^*)$. Then

$$\max_{1 \leq j \leq M} E^* \log\{f(X; \theta_j)/f(X; \theta^*)\} < 0.$$

By the strong law of large numbers, this inequality implies

$$(2.4) \quad \max_{1 \leq j \leq M} \sum_{i=1}^n \log f(x_i; \theta_j) < \sum_{i=1}^n \log f(x_i; \theta^*)$$

almost surely. By the definition of the MLE, $\hat{\theta}_n = \theta^*$ almost surely as $n \rightarrow \infty$. \square

Surprisingly, most generic MLE consistency proofs are variations or novel upgraded versions of this lemma. These include the situations where Θ is a subset of the Euclidean space \mathbb{R}^d , the space of all mixing distributions \mathbb{G} or \mathbb{G}_m or any abstract space.

In Section 2.1, we first present a trivialized Wald theorem, Theorem 2.1. It is universal consistency result though not directly applicable to many practical models. However, the full Wald theorem can be easily explained from this angle.

Theorem 2.1 is then used as basis to understand and prove the consistency results of Kiefer and Wolfowitz (1956), Pfanzagl (1988) in the context of mixture models subsequently presented in Sections 2.2–2.4.

2.1 Essentials of the Wald Consistency Proof

Let $\{f(x; \theta), \theta \in \Theta\}$ be the probability model under investigation, where Θ is an abstract but metric parameter space. For any subset B of Θ , define

$$f(x; B) = \sup_{\theta \in B} f(x; \theta).$$

Denote the distance on Θ as $\rho(\cdot, \cdot)$. For any $\varepsilon > 0$, let

$$B_\varepsilon(\theta^*) = \{\theta : \rho(\theta, \theta^*) < \varepsilon\}$$

be an open ball of radius ε centred at θ^* . Let B^c be the complement of B .

THEOREM 2.1 (Trivialized Wald theorem). *Let x_1, \dots, x_n be an i.i.d. sample from $f(x; \theta^*)$, a member of $\{f(x; \theta), \theta \in \Theta\}$. Let $\hat{\theta}_n$ be an MLE of θ as defined in (2.2).*

Suppose that for any $\varepsilon > 0$, there exists a finite number of subsets B_1, B_2, \dots, B_J of Θ such that $B_\varepsilon^c(\theta^) \subset \{\cup_{j=1}^J B_j\}$ and for each j*

$$(2.5) \quad E^* \log\{f(X; B_j)/f(X; \theta^*)\} < 0.$$

Then $\hat{\theta}_n \rightarrow \theta^$ almost surely as $n \rightarrow \infty$.*

The proof is identical to that of Lemma 2.2. The technicalities of Wald (1949) amount to specifying conditions on the model that lead to the generalized Jensen's inequality (2.5). What are the conditions placed on $f(x; \theta)$ by Wald (1949)? Here is a reorganized and slightly altered list:

(W1) Identifiability: Let $F(x; \theta)$ be the cumulative distribution function of $f(x; \theta)$. $F(x; \theta) = F(x; \theta^*)$ for all x implies $\theta = \theta^*$.

Clearly, if $F(x; \theta) = F(x; \theta^*)$ for all x , there will be no stochastic difference between the data sets generated from these two distributions. Hence, consistent estimation of θ based only on data is not possible if $\theta \neq \theta^*$. Identifiability is therefore a necessary condition for the consistent estimation of θ .

(W2) Continuity and slightly more: For all x , $\lim_{\theta \rightarrow \theta_0} f(x; \theta) = f(x; \theta_0)$ for any given θ_0 and $\lim_{|\theta| \rightarrow \infty} f(x; \theta) = 0$.

Here, $|\theta|$ can be the Euclidean norm or any norm appropriate in the context of the problem. Technically, continuity may not be required for a zero-probability set of x in terms of $f(x; \theta_0)$. The models used in applications are usually well behaved. Hence, this extra generality is not generally needed if one takes appropriate care and, therefore, is not included as part of the condition.

For the consistent estimation of a parameter, the distributions with close θ values must be similar. Therefore, the continuity condition is indispensable.

(W3) Finite Kullback–Leibler information: Let $[\cdot]^+$ denote the positive part of the quantity in the brackets. For any $\theta \neq \theta^*$, there exists an $\varepsilon > 0$ such that

$$E^*[\log\{f(X; B_\varepsilon(\theta))/f(X; \theta^*)\}]^+ < \infty$$

and there exists a large enough $r > 0$ such that

$$E^*[\log\{f(X; B_r^c(\theta^*))/f(X; \theta^*)\}]^+ < \infty.$$

(W4) Closeness: The parameter space Θ is a closed subset of \mathbb{R}^d .

The use of these conditions is demonstrated in the following lemma.

LEMMA 2.3. The conditions of Theorem 2.1 are satisfied under (W1)–(W4).

PROOF. Note that $\log\{f(X; B_\varepsilon(\theta))/f(X; \theta^*)\}$ is a monotone increasing function of ε . The continuity condition (W2) ensures that $\lim_{\varepsilon \rightarrow 0^+} f(X; B_\varepsilon(\theta)) = f(X; \theta)$ as $\varepsilon \rightarrow 0^+$, that is, when ε decreases to 0. Thus, this condition validates the dominated convergence theorem in the following way:

$$\begin{aligned} & \lim_{\varepsilon \rightarrow 0^+} E^*[\log\{f(X; B_\varepsilon(\theta))/f(X; \theta^*)\}]^+ \\ &= E^*[\log\{f(X; \theta)/f(X; \theta^*)\}]^+. \end{aligned}$$

For the negative counterpart of this expectation, the famous Fatou's lemma implies, under the continuity condition (W2),

$$\begin{aligned} & \liminf_{\varepsilon \rightarrow 0^+} E^*[\log\{f(X; B_\varepsilon(\theta))/f(X; \theta^*)\}]^- \\ & \geq E^*[\log\{f(X; \theta)/f(X; \theta^*)\}]^-. \end{aligned}$$

The monotonicity on the left-hand side in ε ensures that the limit exists, rather than merely \liminf . Hence,

$$\begin{aligned} & \lim_{\varepsilon \rightarrow 0^+} E^*[\log\{f(X; B_\varepsilon(\theta))/f(X; \theta^*)\}] \\ & \leq E^*[\log\{f(X; \theta)/f(X; \theta^*)\}] < 0 \end{aligned}$$

with the strict < 0 implied by the identifiability condition (W1).

The above result shows that for each $\theta \neq \theta^*$, there is a small enough ε_θ such that

$$E^* \log\{f(X; B_{\varepsilon_\theta}(\theta))/f(X; \theta^*)\} < 0.$$

Since $f(x; \theta) \rightarrow 0$ as $|\theta| \rightarrow \infty$, we can similarly show that there exists r such that

$$E^* \log\{f(X; B_r^c(\theta^*))/f(X; \theta^*)\} < 0.$$

For any $\varepsilon > 0$, let $\bar{B}_r(\theta^*)$ be the closure of $B_r(\theta^*)$. Then

$$A = \{B_\varepsilon^c(\theta^*)\} \cap \{\bar{B}_r(\theta^*)\}$$

is bounded and closed and, therefore, compact. It is trivial to see that

$$\left[\bigcup_{\theta \in A} \{B_{\varepsilon_\theta}(\theta)\} \right] \supset A.$$

When a compact set is covered by the union of a collection of open sets, it is covered by a finite number of such sets. Applying this result here, we have a finite number of $B_{\varepsilon_\theta}(\theta)$ whose union covers the compact set A . Let these be B_1, \dots, B_{J-1} , and reserve B_J for $B_r(\theta^*)$.

All the conditions in Theorem 2.1 are satisfied. Hence, the MLE is consistent under (W1)–(W4). \square

We have successfully established the consistency result of Wald (1949) stated as Lemma 2.3. The proof is markedly simplified and it is obtained without requiring

$$E^*|\log f(X; \theta^*)| < \infty.$$

When Θ is not a closed set, it is often possible to continuously extend the range of the density function $f(x; \theta)$ to the closure of Θ , $\bar{\Theta}$. The Wald theorem may

then be applicable to the expanded model. Take the Poisson distribution family as an example:

$$f(x; \theta) = \frac{\theta^x}{x!} \exp(-\theta)$$

for $x = 0, 1, \dots$ and $\Theta = (0, \infty)$. It can be seen that Θ is not a closed set. However, by defining $f(x; 0) = 0$ for all x except for $f(x; 0) = 1$ when $x = 0$, we extend the model to $\bar{\Theta} = [0, \infty)$. Consequently, the conditions for a consistent MLE are satisfied and the MLE is consistent. Of course, the consistency of the MLE can be easily established without utilizing the generic Wald theorem.

2.2 Consistency of Nonparametric MLE under Mixture Model: Kiefer–Wolfowitz (KW) Approach

We now illustrate and understand the contribution of Kiefer and Wolfowitz (1956) through its connection to Theorem 2.1. To discuss the consistence of the nonparametric MLE, we need to choose a distance $D(\cdot, \cdot)$ on \mathbb{G} such as

$$D_{KW}(G_1, G_2) = \int_{\Theta} |G_1(\theta) - G_2(\theta)| \exp(-|\theta|) d\theta,$$

where $|\theta|$ is interpreted as $|\theta_1| + \dots + |\theta_d|$ and $d\theta$ as $d\theta_1 \dots d\theta_d$ when $\Theta \subset \mathbb{R}^d$. Note that there are two d 's involved: one is the dimension of θ and the other is the differential symbol for the integration.

At this moment, a generic notion of distance suffices. We say that $G \rightarrow G_0$ if $D(G, G_0) \rightarrow 0$. Suppose $G^* \in \mathbb{G}$ is the true mixing distribution and \hat{G} is an estimator. Then \hat{G} is strongly consistent when $D(\hat{G}, G^*) \rightarrow 0$ almost surely.

(KW1) Identifiability: *Let $F(x; G)$ be the cumulative distribution function of $f(x; G)$. If $F(x; G) = F(x; G^*)$ for all x , then $D(G, G^*) = 0$.*

Suppose $D_{KW}(\cdot, \cdot)$ is chosen as the distance on \mathbb{G} . When $f(x; \theta)$ is confined to the Poisson distribution, (KW1) is satisfied. When $f(x; \theta)$ is the normal distribution, this condition is violated. When $f(x; \theta)$ is binomial, it is also violated in general. When \mathbb{G} is reduced to \mathbb{G}_m , the normal mixture satisfies (KW1); the binomial mixture satisfies (KW1) when m is small. See Teicher (1961, 1963).

For \mathbb{G}_m , a mixing distribution can be expressed by two vectors: one for the component parameter values and one for the corresponding mixing proportions. Suppose the Euclidean distance on this vector space is chosen. Then $f(x; G)$ remains the same when the entries of the two vectors are permuted. This is the

loss of identifiability due to label switching. Quotient topology as suggested by Redner (1981) can be used to avoid this problem. Label switching leads to technical difficulties for Bayesian analysis; see Frühwirth-Schnatter (2011) and Stephens (2000).

(KW2) Continuity: *The component parameter space Θ is a closed set. For all x and any given G_0 ,*

$$\lim_{G \rightarrow G_0} f(x; G) = f(x; G_0).$$

(KW3) Finite Kullback–Leibler information: *For any $G \neq G^*$, there exists an $\varepsilon > 0$ such that*

$$E^*[\log\{f(X; B_\varepsilon(G))/f(X; G^*)\}]^+ < \infty.$$

(KW4) Compactness: *The definition of the mixture density $f(x; G)$ in G can be continuously extended to a compact space $\bar{\mathbb{G}}$ while retaining the validity of (KW3).*

THEOREM 2.2. *Let x_1, \dots, x_n be an i.i.d. sample from $f(x; G^*)$, a member of $\{f(x; G), G \in \mathbb{G}\}$. Suppose conditions (KW1)–(KW4) are satisfied. Then the nonparametric MLE \hat{G}_n is strongly consistent. Namely, $D_{KW}(\hat{G}_n, G^*) \rightarrow 0$ almost surely.*

PROOF. Under the theorem conditions and following the proof of Theorem 2.1, there must exist a $\delta > 0$ for each $G \neq G^*$ such that

$$E^*[\log\{f(X; B_\delta(G))/f(X; G^*)\}] < 0.$$

This leads to a finite open cover of the compact set $B_\varepsilon^c(G^*)$ for any $\varepsilon > 0$. Hence, by the law of large numbers,

$$\max_{G \notin B_\varepsilon(G^*)} \ell_n(G) \leq \ell_n(G^*)$$

almost surely. The arbitrariness of ε implies that \hat{G}_n is within an infinitesimal neighborhood of G^* almost surely as $n \rightarrow \infty$ and is therefore consistent for G^* . \square

The proof of the consistency result in Theorem 2.2 is not fundamentally different from that of Kiefer and Wolfowitz (1956). The simplicity comes from requiring high level conditions (KW3) and (KW4), and from not repeating some steps in Lemma 2.3. The current proof promotes the understanding of the essence of their proof. At the same time, (KW3) and (KW4) can be established by (a): introducing a specific distance on \mathbb{G} ; (b): extending the domain of $f(x; G)$ in G continuously to a compact $\bar{\mathbb{G}}$; (c): verifying that Jensen's inequality hold on $\bar{\mathbb{G}}$. Going over these steps brings back the complexity. We illustrate this point subsequently based on D_{KW} . It is known that $G_m \rightarrow G$ in distribution if and only if $D_{KW}(G_m, G) \rightarrow 0$ as $m \rightarrow \infty$.

2.2.1 *Compactification of \mathbb{G} .* Based on $D_{\text{KW}}(\cdot, \cdot)$, the distance is no more than $\int_{\Theta} \exp(-|\theta|) d\theta < \infty$. To compactify \mathbb{G} , we introduce ρG as a subdistribution for any $\rho \in [0, 1)$. Let $\bar{\mathbb{G}}$ be \mathbb{G} supplemented with all the subdistributions. Clearly, $D_{\text{KW}}(\cdot, \cdot)$ is easily extended to $\bar{\mathbb{G}}$. Similarly, we naturally extend the range of $f(x; G)$ to $G \in \bar{\mathbb{G}}$ by defining

$$f(x; \rho G) = \rho \int_{\Theta} f(x; \theta) dG(\theta).$$

Although we have technically expanded \mathbb{G} to $\bar{\mathbb{G}}$, the likelihood cannot be maximized on $\bar{\mathbb{G}} - \mathbb{G}$ since if $f(x; G) \neq 0$ then $f(x; \rho G) < f(x; G)$ when $\rho < 1$. Hence, the MLE on $\bar{\mathbb{G}}$ is always a proper distribution on Θ .

Since Θ is a closed subset of \mathbb{R}^d from condition (KW2), the limit of any Cauchy sequence in \mathbb{G} in terms of $D_{\text{KW}}(\cdot, \cdot)$ is a subdistribution on Θ . Hence, $\bar{\mathbb{G}}$ is a closed set. With the addition of total boundedness, $\bar{\mathbb{G}}$ is compact.

The extension from $\{f(x; G) : G \in \mathbb{G}\}$ to $\{f(x; G) : G \in \bar{\mathbb{G}}\}$ is largely symbolic. The real issues are continuity and (KW3) on $\bar{\mathbb{G}}$. For instance, the extension would fail on a normal mixture because the component density function is not defined at $\sigma = 0$. Otherwise, the continuous extension is feasible.

2.2.2 *Continuity of $f(x; G)$ on $\bar{\mathbb{G}}$.* Here is the sufficient and likely also necessary condition for the continuity of $f(x; G)$ on $\bar{\mathbb{G}}$ based on D_{KW} distance. Recall that $D_{\text{KW}}(G_m, G) \rightarrow 0$ if and only if $G_m \rightarrow G$ in distribution.

LEMMA 2.4. *Under (W2) and (W4), the extended mixture model $\{f(x; G) : G \in \bar{\mathbb{G}}\}$ is continuous in G for all given x .*

PROOF. Recall that $G_m \rightarrow G_0$ in distribution if and only if $\int h(\theta) dG_m(\theta) \rightarrow \int h(\theta) dG_0(\theta)$ for all bounded and continuous functions $h(\cdot)$, according to one of many equivalent definitions. By condition (W2), $\lim_{|\theta| \rightarrow \infty} f(x; \theta) = 0$. Thus, for given x , $f(x; \theta)$ is continuous and bounded on Θ . Therefore, this definition leads to

$$\begin{aligned} f(x; G_m) &= \int f(x; \theta) dG_m(\theta) \rightarrow \int f(x; \theta) dG_0(\theta) \\ &= f(x; G_0) \end{aligned}$$

whenever $G_m \rightarrow G_0$ in distribution for all $G_0 \in \bar{\mathbb{G}}$. \square

REMARK. Here the convergence in distribution includes subdistributions.

2.2.3 *Generalized Jensen's inequality.* Technicality is inevitable when it comes to establishing the generalized Jensen's inequality for the mixture model on $\bar{\mathbb{G}}$. The user must decide whether or not (KW3) holds on $\bar{\mathbb{G}}$ for each specific mixture model. Two examples will help to explain the issue.

EXAMPLE 2.1. *If $G^*(M) = 1$ for some $M < \infty$, then the Poisson mixture model satisfies all the conditions of Theorem 2.2.*

PROOF. Let θ_0 be a support point of G^* . There must be a positive constant δ such that

$$f(x; G^*) \geq \delta \frac{(\theta_0)^x}{x!} \exp(-\theta_0).$$

Therefore, we have

$$\begin{aligned} E^*\{\log f(X; G^*)\} &\geq \log(\theta_0)E^*(X) - E^*\{\log(X!)\} \\ &\quad + \log \delta - \theta_0. \end{aligned}$$

The condition $G^*(M) = 1$ easily leads to the finiteness of both $E^*(X)$ and $E^*\log(X!)$. Hence, $E^*\{\log f(X; G^*)\} > -\infty$. Since the probability mass function is bounded from above, we also have $E^*\{\log f(X; G^*)\} < \infty$. Therefore, $E^*|\log f(X; G^*)| < \infty$.

Since $f(x; B_\varepsilon(G)) < 1$ for any G and $\varepsilon > 0$,

$$\begin{aligned} E^*[\log\{f(x; B_\varepsilon(G))/f(X; G^*)\}] \\ \leq -E^*\{\log f(X; G^*)\} < \infty. \end{aligned}$$

This is (KW3) for all $G \in \bar{\mathbb{G}}$. \square

Condition (KW3) is not always satisfied even when the component distribution is Poisson. Let $G(\{A\})$ be the probability of $\theta \in A$ under mixing distribution G .

EXAMPLE 2.2. *Let G^* be a mixing distribution such that*

$$G^*(\{\log n\}) = c\{n(\log n)(\log \log n)^2\}^{-1}$$

with some normalizing positive constant c , for $n = 20, 21, \dots$. Then under a Poisson mixture model, we have

$$E^*\{\log f(X; G^*)\} = -\infty.$$

Consequently, for any $\varepsilon > 0$,

$$E^*\{\log f(X, B_\varepsilon(1))/f(X; G^*)\} = \infty,$$

where $B_\varepsilon(1)$ is the set of all mixing distributions within ε -distance of $f(x; 1)$.

PROOF. The size of c in this example does not affect the proof, so we take $c = 1$. We choose this specific mixing distribution because $E^*\{\theta\} = \infty$. In addition, it can be seen that

$$\begin{aligned} f(x; G^*) &= \sum_{n=20}^{\infty} \left[\{n(\log n)(\log \log n)^2\}^{-1} \right. \\ &\quad \left. \cdot \frac{(\log n)^x}{x!} \exp(-\log n) \right] \\ &= \frac{1}{x!} \sum_{n=20}^{\infty} \frac{(\log n)^{x-1}}{(n \log \log n)^2}. \end{aligned}$$

It is seen that

$$\sum_{n=20}^{\infty} \frac{(\log n)^{x-1}}{(n \log \log n)^2} \approx \int_{t=20}^{\infty} \frac{(\log t)^{x-1}}{(t \log \log t)^2} dt.$$

The approximation is so precise that an error assessment is possible but unnecessary. Changing the variable via $u = \log t$, we find

$$\begin{aligned} \int_{t=20}^{\infty} \frac{(\log t)^{x-1}}{(t \log \log t)^2} dt &\leq \int_{u=0}^{\infty} u^{x-1} \exp(-u) du \\ &= (x-1)!. \end{aligned}$$

The choice of $\log(20)$ as the lowest support point of G ensures that $\log(\log(20)) > 1$, which avoids some technicalities in the above inequality. Hence,

$$\log f(x; G^*) \leq \log\{(x-1)!/x!\} = -\log x.$$

Using $E^* = E_{G^*}E_{\theta}$, we find

$$E^*\{\log(X)\} \geq E_{G^*} \log\{E_{\theta} X\} = E_{G^*}\{\log \theta\} = \infty.$$

Therefore,

$$E^*\{\log f(X; G^*)\} \leq -E^* \log(X) = -\infty.$$

Let δ_x be a distribution with all the probability mass at x . It can be seen that

$$D_{\text{KW}}(\delta_1, (1-\varepsilon)\delta_1 + \varepsilon\delta_x) \leq \varepsilon \int \exp(-\theta) d\theta = \varepsilon.$$

Therefore, $(1-\varepsilon)\delta_1 + \varepsilon\delta_x \in B_{\varepsilon}(1)$ for any x . Hence, for any x value,

$$f(x; B_{\varepsilon}(1)) \geq \varepsilon f(x; x) = \frac{\varepsilon x^x}{x!} \exp(-x) \approx \frac{\varepsilon}{\sqrt{2\pi x}}$$

by the Stirling formula. Because the Stirling formula is very accurate, this implies that

$$\begin{aligned} \log\{f(x; B_{\varepsilon}(1))/f(x; G^*)\} \\ \geq \log(\varepsilon/\sqrt{2\pi}) + (1/2)\log x. \end{aligned}$$

Hence, for any $\varepsilon > 0$, we have

$$\begin{aligned} E^*[\log\{f(X; B_{\varepsilon}(1))/f(X; G^*)\}] \\ \geq \log(\varepsilon/\sqrt{2\pi}) + (1/2)E^*\{\log X\} = \infty. \quad \square \end{aligned}$$

The point is that condition (KW3) on a compact $\bar{\mathbb{G}}$ places a severe restriction on the KW proof. Luckily, the Pfanzagl proof is free of this restriction.

2.3 Consistency of Nonparametric MLE under Mixture Model: Pfanzagl Approach

The most demanding condition in the consistency proof of Kiefer and Wolfowitz (1956) is (KW3), needed to validate the generalized Jensen's inequality (2.5). Unfortunately, (KW3) under a mixture model is hard to verify, as seen in the Poisson mixture example. The Pfanzagl approach requires merely (W2) in comparison. Here is the inequality in Pfanzagl (1988) that takes over the role of Jensen's inequality.

LEMMA 2.5. *Let $f(x)$ and $f^*(x)$ be density functions of any two distributions with respect to some σ -finite measure μ . For any $u \in (0, 1)$, we have*

$$E^* \log\{1 + u[f^*(X)/f(X) - 1]\} \geq 0,$$

and equality holds if and only if $f^*(x) = f(x)$ almost surely with respect to the f^* distribution.

PROOF. Let $Y = \{1 + u[f^*(X)/f(X) - 1]\} = (1-u) + u\{f^*(X)/f(X)\}$. It can be seen that

$$\begin{aligned} \log Y &\geq (1-u)\log(1) + u\log\{f^*(X)/f(X)\} \\ &= u\log\{f^*(X)/f(X)\}. \end{aligned}$$

Hence,

$$E^* \log(Y) \geq uE^*[\log\{f^*(X)/f(X)\}] \geq 0,$$

as required. \square

Now consider the mixture model $f(x; G)$ with its domain already extended to $\bar{\mathbb{G}}$ as in the last subsection.

LEMMA 2.6. *Assume that the mixture model is identifiable and (KW2) is satisfied on $\bar{\mathbb{G}}$. Then, for any $G \in \bar{\mathbb{G}}$ and $G \neq G^*$, there exists an $\varepsilon > 0$ such that*

$$(2.6) \quad \begin{aligned} E^* \log\{1 + u[f(X; G^*)/f(X; B_{\varepsilon}(G)) - 1]\} \\ > 0, \end{aligned}$$

where G^* is the true mixing distribution.

PROOF. Note that $f(x; B_{\varepsilon}(G))$ is a monotone increasing function of ε , or it decreases to $f(x; G)$ for all x as $\varepsilon \rightarrow 0^+$. In addition, given $u \in (0, 1)$,

$$\begin{aligned} \log\{1 + u[f(X; G^*)/f(X; B_{\varepsilon}(G)) - 1]\} \\ \geq \log(1-u) > -\infty. \end{aligned}$$

Namely, this function has a finite lower bound, which enables the use of Fatou's lemma. Hence,

$$\begin{aligned} & \lim_{\varepsilon \rightarrow 0^+} E^* \log\{1 + u[f(X; G^*)/f(X; B_\varepsilon(G)) - 1]\} \\ & \geq E^* \log\{1 + u[f(X; G^*)/f(X; G) - 1]\} > 0. \end{aligned}$$

Consequently, there exists a positive ε value at which (2.6) holds. \square

This lemma is the counterpart of the generalized Jensen's inequality in the KW approach. It is challenging to verify the validity of the generalized Jensen's inequality under a general mixture model. In contrast, Lemma 2.6 is valid provided the parameterization is continuous and identifiable.

THEOREM 2.3. *Assume that the mixture model is identifiable and (KW2) is satisfied on $\bar{\mathbb{G}}$. We have*

$$D_{\text{KW}}(\hat{G}_n, G^*) \rightarrow 0.$$

PROOF. For any $\delta > 0$, let $B_\delta^c(G^*)$ be the distributions in $\bar{\mathbb{G}}$ that are at least a distance δ from G^* . Because $\bar{\mathbb{G}}$ is compact, so is $B_\delta^c(G^*)$. Thus, the continuity condition implies that there exists a finite number of $G_k, k = 1, \dots, J$, with corresponding ε_k such that

$$(2.7) \quad B_\delta^c(G^*) \subset \bigcup_{k=1}^J B_k,$$

where $B_k = \{G : D_{\text{KW}}(G, G_k) < \varepsilon_k\}$, and

$$(2.8) \quad E^* \log\{1 + u[f(X; G^*)/f(X; B_k) - 1]\} > 0.$$

By the strong law of large numbers, (2.8) implies

$$n^{-1} \sum_{i=1}^n \log\{1 + u[f(x_i; G^*)/f(x_i; B_k) - 1]\} > 0$$

almost surely for $k = 1, 2, \dots, J$. Consequently, we have

$$\begin{aligned} 0 & < \sum_{i=1}^n \log\{1 + u[f(x_i; G^*)/f(x_i; B_k) - 1]\} \\ & \leq \inf_{G \in B_k} \sum_{i=1}^n \log\{1 + u[f(x_i; G^*)/f(x_i; G) - 1]\} \end{aligned}$$

almost surely for each $k = 1, 2, \dots, J$. Combining this inequality with (2.7), we get

$$0 < \inf_{G \notin B_\delta(G^*)} \sum_{i=1}^n \log\{1 + u[f(x_i; G^*)/f(x_i; G) - 1]\}$$

almost surely. By interpreting the summation in terms of the log-likelihood function, we find

$$\ell_n(uG^* + (1-u)G) > \ell_n(G)$$

for all $G \in B_\delta^c(G^*)$ almost surely. Since the likelihood function at any $G \notin B_\delta(G^*)$ is smaller than the likelihood value at another mixing distribution $uG^* + (1-u)G$ that is a member of $\bar{\mathbb{G}}$, the members of $B_\delta^c(G^*)$, all of which are at least a δ -distance away from G^* , cannot possibly attain the supremum of $\ell_n(G)$. Hence, the nonparametric MLE must reside in the δ -neighborhood of G^* almost surely. The arbitrarily small size of δ implies that $D_{\text{KW}}(\hat{G}, G^*) \rightarrow 0$ almost surely as $n \rightarrow \infty$. This completes the proof. \square

In this proof, Pfanzagl (1988) took tactical advantage of the linearity of the mixture model in mixing distributions:

$$uf(x; G^*) + (1-u)f(x; G) = f(x; uG^* + (1-u)G),$$

which is the density function of another mixture distribution.

There is a limitation in the Pfanzagl result. Consider the finite mixture model where $\bar{\mathbb{G}}$ is replaced by \mathbb{G}_m for a given m . The Pfanzagl result is no longer applicable because $uG^* + (1-u)G$ likely has more than m support points even if both G^* and G have only m support points. In contrast, the KW proof leads to the consistency of the MLE under finite mixture models provided the corresponding conditions are satisfied. In addition, the KW conditions under finite mixture models are simple to verify, and they hold widely.

For finite mixture models, there is another widely cited paper on the consistency of the MLE.

2.4 Consistency of the MLE under Finite Mixture Model: Redner Approach

The generic results in Redner (1981) are not restricted to the finite mixture model: the paper examines models lacking full identifiability. Without identifiability, the parameter estimator is inconsistent in general. However, in many situations, the estimator may be regarded as consistent from a different angle.

Suppose a probability model has its density function given by

$$f(x; \theta_1, \theta_2) = \theta_1 \exp(-\theta_1 x)$$

for $x > 0$. Of course, this is simply an exponential distribution with rate parameter θ_1 ; parameter θ_2 is irrelevant. Given a set of i.i.d. samples from a distribution in this model, the MLE of θ_1 is consistent, while there is no way to have θ_2 consistently estimated. At the same time, the consistent estimation of θ_2 is unnecessary: it has no role in this population. Let the distance between two vectors be

$$\rho((\theta_1, \theta_2), (\eta_1, \eta_2)) = |\theta_1 - \eta_1|.$$

Then the MLE would satisfy, almost surely,

$$\rho((\hat{\theta}_1, \hat{\theta}_2), (\theta_1^*, \theta_2^*)) \rightarrow 0$$

as the sample size $n \rightarrow \infty$.

Finite mixture models may not appear as trivial as in this example. The density function of the two-component normal mixture in mean parameter is given by

$$f(x; \alpha, \theta_1, \theta_2) = (1 - \alpha)\phi(x - \theta_1) + \alpha\phi(x - \theta_2).$$

The mixture with $\eta = (\alpha, \theta_1, \theta_2) = (0.3, 1, 2)$ is identical to the mixture with $\eta = (0.7, 2, 1)$. These two parameter vectors are apparently distinct in \mathbb{R}^3 . Hence, identifiability is lost when the model is parameterized with this scheme. When η is regarded as a vector in \mathbb{R}^3 equipped with the Euclidean distance, the MLE is not consistent.

Each η in the above model has a corresponding mixing distribution G . Let us define a distance for \mathbb{R}^3 (with the first component in $[0, 1]$) as

$$\rho(\eta_1, \eta_2) = D_{KW}(G_1, G_2).$$

Under this distance definition, we have

$$\rho(\hat{\eta}, \eta^*) \rightarrow 0$$

as $n \rightarrow \infty$ under appropriate conditions.

Since $\rho(\eta_1, \eta_2) = 0$ does not lead to $\eta_1 = \eta_2$ in \mathbb{R}^3 , $\rho(\cdot, \cdot)$ is not a mathematical distance. Regarding several distinct members of \mathbb{R}^3 as identical leads to quotient topology. This turns out to be the core of Redner (1981).

2.5 Summary

The consistency result of Pfanzagl (1988) seems perfect except for the consistency of the MLE under finite mixture models. The proof of Kiefer and Wolfowitz (1956) contains conditions that are not user-friendly. However, these conditions become simple under finite mixture models. Hence, the two papers perfectly complement each other. Redner (1981) resorts to quotient topology to resolve nonidentifiability and thereby provides another consistency proof for the finite mixture model. Redner (1981), however, assumes that Θ is a compact subset of \mathbb{R}^d , making the result weaker.

These papers do not consider only mixture models, as this section may have suggested. In this section, we have substantially streamlined the conditions and conclusions and provided additional insight in the context of mixture models.

3. CONSISTENCY UNDER FINITE NORMAL MIXTURE MODEL

One common requirement for the consistency of the MLE under a mixture model is that $f(x; G)$ can be continuously extended to \mathbb{G} . The first step of this extension is to have $f(x; \theta)$ continuously extended to include all θ on the boundary of Θ . This turns out to be impossible for the normal model with density function

$$\phi(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}.$$

Because of this, none of the three approaches in the last section is applicable to normal mixture models.

The above issue is not the only obstacle. The normal mixture model is not identifiable on \mathbb{G} unless that space is reduced to \mathbb{G}_m for a prespecified m . This section is devoted to the consistent estimation of G under the finite normal mixture model.

3.1 Finite Normal Mixture Model with Equal Variances

Consider the finite normal mixture model where the component distributions share an equal but unknown variance:

$$(3.1) \quad f(x; G; \sigma^2) = \sum_{j=1}^m \alpha_j \phi(x; \theta_j, \sigma^2).$$

Let G be the mixing distribution in component mean on $\Theta = \mathbb{R}$. The common variance σ is structural with parameter space \mathbb{R}^+ .

The log-likelihood function based on a set of i.i.d. samples is given by

$$\ell_n(G; \sigma^2) = \sum_{i=1}^n \log f(x_i; G; \sigma^2).$$

Here is a preliminary result similar to but much strengthened over that in Chen and Chen (2003). We no longer confine the mean parameter in a finite interval.

LEMMA 3.1. *Let $(\hat{G}, \hat{\sigma}^2)$ be a global maximum point of the likelihood function $\ell_n(G; \sigma^2)$. Then there exist constants $0 < \varepsilon < \Delta < \infty$ such that as $n \rightarrow \infty$; the event sequence $\{\varepsilon \leq \hat{\sigma}^2 \leq \Delta\}$ occurs almost surely.*

PROOF. It can be seen that $f(x; G; \sigma^2) \leq 1/\sigma$ for all x and G . When $\sigma^2 > \Delta$,

$$\ell_n(G; \sigma^2) \leq -(1/2)n \log \Delta.$$

Let \bar{x} be the sample mean, and $s_n^2 = n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$. Then

$$\ell_n(\bar{x}; s_n^2) \geq -n \log(s_n) - (n/2).$$

It can then be seen that

$$\begin{aligned} \ell_n(\bar{x}, s_n^2) - \ell_n(G; \sigma^2) & \\ & \geq \{-n \log(s_n) - (n/2)\} - \{-(1/2)n \log \Delta\} \\ & = n\{\log \Delta - \log(s_n) - (1/2)\} \end{aligned}$$

uniformly for all $\sigma^2 > \Delta$.

Let X be a random variable with the true finite normal mixture distribution. By the strong law of large numbers, s_n^2 almost surely converges to $\text{VAR}(X)$ as $n \rightarrow \infty$. Hence, when $\log \Delta > \log\{\text{VAR}(X)\} + (1/2)$, we have

$$\ell_n(G; \sigma^2) < \ell_n(\bar{x}, s_n^2)$$

almost surely for all $\sigma^2 > \Delta$. This proves that the MLE for σ^2 is below this finite value Δ almost surely.

Next, because of the algebraic form of the normal density,

$$(3.2) \quad \log f(x; G, \sigma^2) \leq -\log(\sigma)$$

regardless of the actual value of x . At the same time,

$$f(x; G; \sigma^2) = \sum_{j=1}^m \alpha_j \phi(x; \theta_j, \sigma^2) \leq \max_j \phi(x; \theta_j, \sigma^2).$$

Hence, for any G and σ^2 , there is another upper bound:

$$\log f(x; G; \sigma^2) \leq -\log \sigma - (2\sigma^2)^{-1} \min_{1 \leq j \leq m} (x - \theta_j)^2.$$

Let M be an arbitrary positive number and denote the truncated θ value as

$$\tilde{\theta} = \begin{cases} -M, & \theta < -M; \\ \theta, & |\theta| < M; \\ M, & \theta > M. \end{cases}$$

Let $\theta = (\theta_1, \dots, \theta_m)$ and $\tilde{\theta} = (\tilde{\theta}_1, \dots, \tilde{\theta}_m)$. The space of $\tilde{\theta}$ is clearly compact given finite M . For any $|x| \leq M$, we have

$$(3.3) \quad \begin{aligned} \log f(x; G, \sigma^2) & \\ & \leq -\log \sigma - (2\sigma^2)^{-1} \min_{1 \leq j \leq m} (x - \tilde{\theta}_j)^2. \end{aligned}$$

Applying (3.2) for $|x_i| > M$ and (3.3) for $|x_i| \leq M$,

$$(3.4) \quad \begin{aligned} \ell_n(G, \sigma^2) & \leq -n \log \sigma - (2\sigma^2)^{-1} \\ & \cdot \sum_{i=1}^n \left\{ \min_{1 \leq j \leq m} (x_i - \tilde{\theta}_j)^2 \right\} I(|x_i| \leq M). \end{aligned}$$

Now we focus on the stochastic quantity

$$h(\tilde{\theta}, X) = \left\{ \min_{1 \leq j \leq m} (X - \tilde{\theta}_j)^2 \right\} I(|X| \leq M).$$

Clearly, $h(\tilde{\theta}, x) \leq M^2$ and it is equicontinuous in $\tilde{\theta}$ for all x . Because the space of $\tilde{\theta}$ is compact, by the uniform strong law of large numbers of Rubin (1956),

$$n^{-1} \sum_{i=1}^n h(\tilde{\theta}, x_i) \rightarrow E^* \{h(\tilde{\theta}, X)\}$$

almost surely and uniformly in $\tilde{\theta}$. Because $E^* \{h(\tilde{\theta}, X)\}$ is smooth in $\tilde{\theta}$ and it is clearly nonzero at each $\tilde{\theta}$, this implies

$$\inf E^* \{h(\tilde{\theta}, X)\} = \delta > 0,$$

where the infimum is over the compact space of $\tilde{\theta}$. Applying this result to (3.3), we find

$$\ell_n(G, \sigma) \leq -n \{\log \sigma + \delta / \sigma^2\}$$

almost surely for all σ and, therefore,

$$(3.5) \quad \begin{aligned} \ell_n(G, \sigma) - \ell_n(\bar{x}, s_n^2) & \\ & \leq -n \{\log \sigma - \log(s_n) + \delta / \sigma^2 - 1/2\} \end{aligned}$$

almost surely. When σ^2 is small enough, the upper bound goes to negative infinity as $n \rightarrow \infty$. Hence, the maximum value of $\ell_n(G, \sigma)$ must be attained when $\sigma > \varepsilon$ for some $\varepsilon > 0$. This completes the proof. \square

The key improvement of this lemma over that in Chen and Chen (2003) is that here the parameter space of θ is the noncompact \mathbb{R} .

Based on this result, under the finite normal mixture model with equal variance, the effective component parameter space for (θ, σ^2) is $\mathbb{R} \times [\varepsilon, \Delta]$ from the asymptotic point of view. Restricting the space of σ in this way leads to a compact parameter space. On this component parameter space, we have

$$\lim_{|\theta| \rightarrow \infty} f(x; \theta, \sigma^2) = 0.$$

Hence, conditions (W2) and (W3) are satisfied after this restriction and the KW approach is applicable.

THEOREM 3.1. *Under the finite normal mixture model (3.1) with m known, the MLE $(\hat{G}, \hat{\sigma}^2)$ is strongly consistent.*

The proof is simple. Lemma 3.1 implies that the effective component parameter space is $\mathbb{R} \times [\varepsilon, \Delta]$. The KW conditions on the reduced component parameter space are satisfied. Hence, Theorem 2.2 can be applied to give the consistency result.

3.2 Finite Normal Mixture Model with Unequal Variances

The unequal-variance assumption does not exclude the possibility that the true component variances are all equal. The density function is now given by

$$(3.6) \quad f(x; G) = \sum_{j=1}^m \alpha_j \phi(x; \theta_j, \sigma_j^2).$$

In this case, the mixing distribution G is bivariate on $\mathbb{R} \times \mathbb{R}^+$ and it mixes both mean and variance. The log-likelihood has the same symbolic form:

$$\ell_n(G) = \sum_{i=1}^n \log f(x_i; G).$$

Consider the case where $m = 2$. Let $(\theta_1, \sigma_1) = (0, 1)$, $\alpha_1 = \alpha_2 = 0.5$. Let $\theta_2 = x_1$ and $\sigma_2 = 1/2k$ for $k = 1, 2, \dots$. Let G_k be the corresponding mixing distribution. This setup creates a sequence of mixing distributions $\{G_k\}_{k=1}^\infty$. It can be seen that

$$f(x_1; G_k) = \frac{0.5}{\sqrt{2\pi}}(2k) + \frac{0.5}{\sqrt{2\pi}} \exp\left(-\frac{x_1^2}{2}\right) \geq \frac{k}{2\pi}$$

and that for $i \geq 2$,

$$\begin{aligned} f(x_i; G_k) &= \frac{0.5}{\sqrt{2\pi}}(2k) \exp(-2k^2(x_i - x_1)^2) \\ &\quad + \frac{0.5}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2}\right) \\ &\geq \frac{1}{2\pi} \exp\left(-\frac{x_i^2}{2}\right). \end{aligned}$$

Consequently, we have

$$\ell_n(G_k) \geq \log(k) - \frac{1}{2} \sum_{i=2}^n x_i^2 - n \log(2\pi).$$

Clearly, $\ell_n(G_k) \rightarrow \infty$ as $k \rightarrow \infty$. Hence, the limiting point of G_k is one of the MLEs of G , which is inconsistent.

There are some misconceptions in the literature. Since $\text{PR}(X_1 = \theta) = 0$ for any given θ value, one may suggest that the probability of having a degenerate MLE is zero. This is false because $\theta = x_1$, in which x_1 is an observed value, is no longer random after the observation.

In applications, EM-algorithm can be used to locate many nondegenerate local maxima of $\ell_n(G)$. The one with the largest likelihood value is a locally consistent MLE of G (Peters and Walker, 1978; Redner and

Walker, 1984). One may also use a consistent estimator, possibly via method of moments, as an initial mixing distribution for the EM-algorithm. Gan and Jiang (1999) developed an approach to test for global maximum, which can be useful in the current context. Experience shows that such an estimator has good statistical properties, so we should not write off this practice.

The inconsistency conclusion may not be an obstacle in many applications. Nonetheless, it is more satisfactory to have a foolproof method with solid underlying theory that performs well in applications. In the literature, there are two approaches to consistent estimation based on likelihood. One is the constraint MLE proposed by Hathaway (1985). Simply put, it reduces the component parameter space of σ . The result of the last subsection may be regarded as its simplest case.

This subsection focuses on the penalty method applied to σ . The penalty is also a prior on σ or a regularization measure. The first largely successful proof of consistency for the penalized MLE is in Ciuperca, Ridolfi and Idier (2003), following its proposal in Ridolfi and Idier (1999). Chen, Tan and Zhang (2008) provided a successful complete proof that is simplified and improved here.

3.2.1 Penalized likelihood. The inconsistency of the MLE under the finite normal mixture model is largely due to nonregularity. Hence, regularizing the likelihood is a natural way to gain consistency of the altered MLE. The regularization is itself in the form of a penalized likelihood as follows:

$$\tilde{\ell}_n(G) = \ell_n(G) + p_n(G).$$

The mixing distribution is then estimated by one of the global maxima of $\tilde{\ell}_n(G)$ over \mathbb{G}_m :

$$\tilde{G} = \arg \max \tilde{\ell}_n(G).$$

The uniqueness is a natural consequence of the subsequent discussion. We denote the component means in \tilde{G} as $\tilde{\theta}_j$ and so on.

3.2.2 Technical lemmas. The following lemma provides a technical basis for the size of the penalty.

LEMMA 3.2. *Let x_1, \dots, x_n be a set of n i.i.d. observations from an absolute continuous distribution F with density function $f(x)$. Assume that $f(x)$ is continuous and $M = \sup_x f(x) < \infty$. Let $F_n(x) = n^{-1} \sum_{i=1}^n I(x_i \leq x)$ be the empirical distribution function.*

Then, as $n \rightarrow \infty$ and almost surely, for any given $\varepsilon > 0$,

$$\sup_{\theta \in \mathbb{R}} \{F_n(\theta + \varepsilon) - F_n(\theta)\} \leq 2M\varepsilon + 8n^{-1} \log n.$$

PROOF. Since $F(x)$ is continuous, there exist $\eta_0, \eta_1, \dots, \eta_n$ such that $F(\eta_i) = i/n$ for $0 < i < n$ with $\eta_0 = -\infty$ and $\eta_n = \infty$. This ensures that for any θ value, there exists a j such that $\eta_{j-1} < \theta \leq \eta_j$. Therefore,

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

The task of the proof is to find appropriate bounds for these two terms. First,

$$\begin{aligned}
& F(\eta_j + \varepsilon) - F(\eta_{j-1}) \\
& \leq \{F(\eta_j + \varepsilon) - F(\eta_j)\} + \{F(\eta_j) - F(\eta_{j-1})\}.
\end{aligned}$$

Since $F(\eta_j) - F(\eta_{j-1}) = n^{-1}$ and $F(\eta_j + \varepsilon) - F(\eta_j) \leq M\varepsilon$ by the mean value theorem, we have

$$\max_j \{F(\eta_j + \varepsilon) - F(\eta_{j-1})\} \leq M\varepsilon + n^{-1}.$$

Let $Y_i = I(\eta_{j-1} < x_i \leq \eta_j + \varepsilon)$ and write $\Delta_j = n^{-1} \sum_i \{Y_i - EY_i\}$. The first term in (3.7) equals $\max_j \Delta_j$. By applying Bernstein's inequality to Δ_j followed by the Borel–Cantelli lemma as in [Serfling \(1980\)](#), we get

$$\max_j \Delta_j < M\varepsilon + 8n^{-1} \log n.$$

Combining the two bounds leads to

$$\begin{aligned}
(3.8) \quad & \sup_{\theta} \{F_n(\theta + \varepsilon) - F_n(\theta)\} \\
& \leq 2M\varepsilon + n^{-1} + 8n^{-1} \log n
\end{aligned}$$

almost surely. Because n^{-1} is a high-order term compared to $n^{-1} \log n$, it is absorbed into the latter. \square

The proof remains solid when ε depends on n . Because of this, ε is allowed to take an arbitrarily small value without invalidating the inequality. Technically, Lemma 3.2 leaves a zero-probability event for each value of ε on which the upper bound is violated. The union of these zero-probability events over ε does not have to be a zero-probability event. However, since $\sup_{\theta} \{F_n(\theta + \varepsilon) - F_n(\theta)\}$ is monotone in ε , this technicality is easily resolved.

LEMMA 3.3. *The upper bound in Lemma 3.2 after a minor alteration,*

$$\sup_{\theta \in \mathbb{R}} \{F_n(\theta + \varepsilon) - F_n(\theta)\} \leq 2M\varepsilon + 10n^{-1} \log n,$$

holds uniformly for all $\varepsilon > 0$ almost surely.

This lemma shows that i.i.d. observations from a population with a bounded density function spread out evenly almost surely. This result extends a corresponding result in [\(Chen, Tan and Zhang, 2008\)](#) to cover generic distribution F .

3.2.3 *Choice of penalty.* Imposing the following three properties on the penalty function makes the penalized MLE consistent:

P1. Additivity: $p_n(G) = \sum_{j=1}^m \tilde{p}_n(\sigma_j)$.

P2. Uniform upper bound: $\sup_{\sigma > 0} [\tilde{p}_n(\sigma)]^+ = o(n)$; individual lower bound: $\tilde{p}_n(\sigma) = o(n)$ for each $\sigma > 0$.

P3. Sufficiently severe: $\tilde{p}_n(\sigma) < (\log n)^2 \log(\sigma)$ for $\sigma < n^{-1} \log n$ when n is large enough.

The first property allows for a simple discussion and straightforward numerical solution to the penalized MLE. The upper and lower bounds in **P2** prevent the likelihood from being seriously inflated or deflated at any G . Property **P3** requires the size of the penalty to be large enough to prevent $\sigma_j \approx 0$ in the penalized MLE of G . One possible \tilde{p}_n is

$$\tilde{p}_n(\sigma) = -n^{-1} \{\sigma^{-2} + \log \sigma^2\}.$$

This penalty function goes to negative infinity when $\sigma \rightarrow 0$ or $\sigma \rightarrow \infty$. It is minimized when $\sigma = 1$. The upper-bound condition in **P2** on $[\tilde{p}_n(\sigma)]^+$ is certainly satisfied. The lower-bound condition in **P2** is clearly satisfied. As for **P3**, when $\sigma \rightarrow 0^+$, $\sigma^{-2} \rightarrow \infty$ much faster than $|\log \sigma|$. Hence, the penalty is much more severe than the required $4(\log^2 n) \log \sigma$. In applications, σ should be replaced by σ/s_n , where s_n is the sample variance, to retain scale invariance. This penalty function also represents a prior Gamma distribution placed on σ^{-2} . This form of penalty is very convenient for the EM-algorithm popularly used for numerical computation and on Bayes analysis as in [Redner, Hathaway and Bezdek \(1987\)](#), [Ridolfi and Idier \(1999\)](#).

3.2.4 *Consistency of the penalized MLE.* We now outline the proof for $m = 2$; the general case is similar and omitted. Let $K^* = E\{\log f(X; G^*)\}$ and $M = \sup_x f(x; G^*)$. Select a sufficiently small ε_0 such that

- (1) $4M\varepsilon_0(\log \varepsilon_0)^2 \leq 1$;
- (2) $(1/2)(\log \varepsilon_0)^2 + \log(\varepsilon_0) \geq 4 - 2K^*$.

Without loss of generality, we assume $\sigma_1 \leq \sigma_2$. Partition the mixing distribution space \mathbb{G}_2 into

$$\Gamma_1 = \{G : \sigma_1 \leq \sigma_2 \leq \varepsilon_0\};$$

$$\Gamma_2 = \{G : \sigma_1 < \tau_0, \sigma_2 > \varepsilon_0\}$$

for some constant $\tau_0 < \varepsilon_0$ to be specified;

$$\Gamma_3 = \{\Gamma_1 \cup \Gamma_2\}^c.$$

The overall strategy is to show that the penalized MLE is not in Γ_1 or Γ_2 but in Γ_3 . The final conclusion follows from the fact that the finite normal mixture model on Γ_3 fits into the KW proof.

Step I. The following lemma says it all.

LEMMA 3.4. *In the current setting,*

$$\sup_{G \in \Gamma_1} \tilde{\ell}_n(G) - \ell_n(G^*) \rightarrow -\infty$$

almost surely as $n \rightarrow \infty$.

PROOF. Define $A_j = \{i : |x_i - \theta_j| < |\sigma_j \log \sigma_j|\}$ for $j = 1, 2$. Partition the entries in $\ell_n(G)$ to get

$$\ell_n(G) = \ell_n(G; A_1) + \ell_n(G; A_1^c A_2) + \ell_n(G; A_1^c A_2^c),$$

where $\ell_n(G; A) = \sum_{i \in A} \log f(x_i; G)$.

Denote the number of observations in set A as $n(A)$. Since the mixture density function is bounded by σ_1^{-1} for $G \in \Gamma_1$,

$$\ell_n(G; A_1) \leq -n(A_1) \log(\sigma_1).$$

Applying Lemma 3.3 with $\varepsilon = \sigma_1 \log(1/\sigma_1)$, we get

$$(3.9) \quad n(A_1) \leq -2nM\sigma_1 \log(\sigma_1) + 10 \log n.$$

Hence,

$$(3.10) \quad \begin{aligned} \ell_n(G; A_1) &\leq 2nM\sigma_1(\log \sigma_1)^2 \\ &\quad - 10(\log n) \log(\sigma_1). \end{aligned}$$

By **P3**, $\tilde{p}_n(\sigma_1) < (\log n)^2 \log(\sigma_1)$. Hence, (3.10) leads to

$$\begin{aligned} \ell_n(G; A_1) + \tilde{p}_n(\sigma_1) &\leq 2nM\sigma_1(\log \sigma_1)^2 \\ &\quad - \{10(\log n) - (\log n)^2\} \log(\sigma_1) \\ &\leq 2nM\sigma_1(\log \sigma_1)^2. \end{aligned}$$

Similarly,

$$\ell_n(G; A_1^c A_2) + \tilde{p}_n(\sigma_2) \leq 2nM\sigma_2(\log \sigma_2)^2.$$

Let $\phi(\cdot)$ be the density function of the standard normal. The observations falling outside both A_1 and

A_2 have log-likelihood contributions that are bounded by

$$\begin{aligned} \log \left\{ \frac{\alpha_1}{\sigma_1} \phi(-\log \sigma_1) + \frac{\alpha_2}{\sigma_2} \phi(-\log \sigma_2) \right\} \\ \leq -\log(\varepsilon_0) - \frac{1}{2}(\log \varepsilon_0)^2. \end{aligned}$$

At the same time, by (3.9) and for small enough ε_0 and sufficiently large n ,

$$n(A_1^c A_2^c) \geq n - \{n(A_1) + n(A_2)\} \geq \frac{n}{2}.$$

Hence, we get the third bound:

$$\ell_n(G; A_1^c A_2^c) \leq -\frac{n}{2} \left\{ \log(\varepsilon_0) + \frac{1}{2}(\log \varepsilon_0)^2 \right\}.$$

Combining the three bounds and remembering how ε_0 was selected, we conclude that when $G \in \Gamma_1$,

$$\begin{aligned} \tilde{\ell}_n(G) &= \{\ell_n(G; A_1) + \tilde{p}_n(\sigma_1)\} \\ &\quad + \{\ell_n(G; A_1^c A_2) - \tilde{p}_n(\sigma_2)\} \\ &\quad + \ell_n(G; A_1^c A_2^c) \\ &\leq 4Mn\varepsilon_0(\log \varepsilon_0)^2 \\ &\quad - \frac{n}{2} \left\{ \frac{1}{2}(\log \varepsilon_0)^2 + \log(\varepsilon_0) \right\} \\ &\leq n - \frac{n}{2}(4 - 2K^*) \\ &= n(K^* - 1). \end{aligned}$$

The last few inequalities hold by the tactical choice of ε_0 .

By the strong law of large numbers, $n^{-1} \tilde{\ell}_n(G^*) \rightarrow K^*$ almost surely. The last inequality is then simplified to

$$\sup_{G \in \Gamma_1} \tilde{\ell}_n(G) - \tilde{\ell}_n(G^*) \leq -n \rightarrow -\infty.$$

This completes the proof. \square

REMARK. I have omitted ‘‘almost surely’’ in the proof for ease of presentation.

Step II. The penalized MLE of G is almost surely not inside Γ_2 , for an appropriately chosen τ_0 . The choice may depend on G^* but not on the sample size n . Let $\bar{\Gamma}_2$ be a compactified Γ_2 allowing $\sigma_1 = 0$ and $\alpha_1 + \alpha_2 < 1$. Define, for any $G \in \bar{\Gamma}_2$,

$$(3.11) \quad g(x; G) = \alpha_1 \phi(x; \theta_1, 2\varepsilon_0^2) + \alpha_2 \phi(x; \theta_2, \sigma_2^2).$$

On $\bar{\Gamma}_2$, σ_2 has a nonzero lower bound. Thus, $g(x; G)$ is bounded although $\sigma_1 = 0$ is allowed.

Without loss of generality, τ_0 is small enough such that the true mixing distribution $G^* \notin \Gamma_2$. Hence, applying Jensen's inequality, we also have

$$E^* \log\{g(X; G)/f(X; G^*)\} < 0$$

for all $G \in \bar{\Gamma}_2$. Using a slightly different symbol from ℓ , we define

$$l_n(G) = \sum_{i=1}^n \log\{g(x_i; G)\}$$

on $\bar{\Gamma}_2$. By the strong law of large numbers and the newly established Jensen's inequality,

$$\begin{aligned} n^{-1}\{l_n(G) - \ell_n(G^*)\} \\ \rightarrow E^* \log\{g(X; G)/f(X; G^*)\} < 0. \end{aligned}$$

Further exploring this key conclusion leads to the following lemma.

LEMMA 3.5. *Consider a set of n i.i.d. observations from $f(x; G^*)$ and the function $g(x; G)$ defined by (3.11). Let $l_n(G) = \sum_{i=1}^n \log\{g(x_i; G)\}$. We then have*

$$(3.12) \quad l_n(G) - \ell_n(G^*) \leq -n\delta(\varepsilon_0)$$

for some $\delta(\varepsilon_0) > 0$ almost surely.

PROOF. Let

$$\begin{aligned} g(x; G, \varepsilon) \\ = \sup\{g(x; G') : G' \in \bar{\Gamma}_2, D_{\text{KW}}(G', G) < \varepsilon\}. \end{aligned}$$

Because $\sigma_2 > \tau_0 > 0$, we still have $g(x; G, \varepsilon) \leq 1 + \tau_0^{-1}$. Hence, $E^* \log g(X; G, \varepsilon) < \infty$.

Clearly, $E^* \log g(X; G, \varepsilon) > -\infty$. Therefore,

$$E^* \log\{g(X; G, \varepsilon)/f(X; G^*)\}$$

is well defined. Let $\varepsilon \rightarrow 0^+$; by the monotone convergence theorem we find

$$\begin{aligned} \lim_{\varepsilon \downarrow 0} E^* \log\{g(X; G, \varepsilon)/f(X; G^*)\} \\ \leq E^* \log\{g(X; G)/f(X; G^*)\} < 0. \end{aligned}$$

Next, note that $\bar{\Gamma}_2$ is compact based on the distance $D_{\text{KW}}(\cdot, \cdot)$. There is a finite number of G and ε such that

$$\bar{\Gamma}_2 \subset \bigcup_{j=1}^J \{G : D_{\text{KW}}(G, G_j) \leq \varepsilon_j\}$$

and for each $j = 1, \dots, J$,

$$E^* \{g(X; G_j, \varepsilon_j)/f(X; G^*)\} < 0.$$

This leads to the claim of this lemma:

$$l_n(G) - \ell_n(G^*) \leq -n\delta(\varepsilon_0)$$

for some $\delta(\varepsilon_0) > 0$ whose size depends on the size of ε_0 . \square

Let us connect $l_n(G)$ to $\ell_n(G)$ on the space $\bar{\Gamma}_2$ and refine this result to obtain the major result of this step.

LEMMA 3.6. *As $n \rightarrow \infty$,*

$$\sup_{G \in \bar{\Gamma}_2} \tilde{\ell}_n(G) - \tilde{\ell}_n(G^*) \rightarrow -\infty.$$

PROOF. Retain the definition $A_1 = \{i : |x_i - \theta_1| \leq \sigma_1 \log(1/\sigma_1)\}$. For each $i \in A_1$, we have

$$f(x_i; G) \leq (1/\sigma_1)g(x_i; G).$$

Therefore, the log-likelihood contribution of each observation in A_1 is

$$\log\{f(x_i; G)\} \leq \log(1/\sigma_1) + \log\{g(x_i; G)\}.$$

The observed values of the observations not in A_1 satisfy $|x - \theta_1| \geq |\sigma_1 \log \sigma_1|$. For these x values, we have

$$\begin{aligned} \frac{(x - \theta_1)^2}{2\sigma_1^2} &\geq \frac{(x - \theta_1)^2}{4\sigma_1^2} + \frac{1}{4}(\log \sigma_1)^2 \\ &\geq \frac{(x - \theta_1)^2}{4\varepsilon_0^2} + \frac{1}{4}(\log \sigma_1)^2. \end{aligned}$$

Consequently,

$$\begin{aligned} \frac{1}{\sigma_1} \exp\left\{-\frac{(x - \theta_1)^2}{2\sigma_1^2}\right\} \\ \leq \exp\left\{-\frac{(x - \theta_1)^2}{4\varepsilon_0^2}\right\} \\ (3.13) \quad \times \exp\left\{-\frac{1}{4}(\log \sigma_1)^2 - \log \sigma_1\right\} \\ = \exp\left\{-\frac{(x - \theta_1)^2}{4\varepsilon_0^2}\right\} \\ \times \exp\left\{-\frac{1}{4}(\log \sigma_1 + 2)^2 + 1\right\}. \end{aligned}$$

The factor $1/\sigma_1$ has been turned into $\exp(-\log \sigma_1)$ in the first line of the above derivation. For a small enough ε_0 ,

$$\exp\left\{-\frac{1}{4}(\log \sigma_1 + 2)^2 + 1\right\} \leq \frac{1}{2\varepsilon_0}$$

when $\sigma_1 \leq \varepsilon_0$. Hence, (3.13) leads to, for those x values not in the set A_1 ,

$$\phi(x; \theta_1, \sigma_1^2) \leq \phi(x; \theta_1, 2\varepsilon_0^2)$$

and, therefore,

$$\begin{aligned} f(x; G) &= \alpha_1 \phi(x; \theta_1, \sigma_1^2) + \alpha_2 \phi(x; \theta_2, \sigma_2^2) \\ &\leq \alpha_1 \phi(x; \theta_1, 2\varepsilon_0^2) + \alpha_2 \phi(x; \theta_2, \sigma_2^2) \\ &= g(x; G). \end{aligned}$$

In summary, when $i \notin A_1$, its log-likelihood contributions

$$\log f(x_i; G) \leq \log\{g(x_i; G)\}.$$

Combining the cases for the observations in and not in A_1 , we find

$$\ell_n(G) \leq n(A_1) \log(1/\sigma_1) + \sum \log\{g(x_i; G)\}.$$

This leads to

$$\begin{aligned} \sup_{G \in \Gamma_2} \tilde{\ell}_n(G) &\leq \sup_{G \in \Gamma_2} \{l_n(G) + \tilde{p}_n(\sigma_2)\} \\ &\quad + \sup_{G \in \Gamma_2} \{n(A_1) \log(1/\sigma_1) + \tilde{p}_n(\sigma_1)\}. \end{aligned}$$

Reusing the bound (3.9) on $n(A_1)$ together with **P3**, we get

$$\sup_{G \in \Gamma_2} \{n(A_1) \log(1/\sigma_1) + \tilde{p}_n(\sigma_1)\} < 2Mn\tau_0(\log \tau_0)^2.$$

Hence, the proof of the lemma is reduced to showing that

$$(3.14) \quad \begin{aligned} &\left[\sup_{\Gamma_2} \{l_n(G) + \tilde{p}_n(\sigma_2)\} + 2Mn\tau_0(\log \tau_0)^2 \right] \\ &\quad - \tilde{\ell}_n(G^*) < 0. \end{aligned}$$

Because $[p_n(\sigma_2)]^+ = o_p(n)$ by choice, it suffices to show that

$$(3.15) \quad \sup_{\Gamma_2} l_n(G) - \ell_n(G^*) \leq -\delta n$$

for some $\delta > 2M\tau_0(\log \tau_0)^2$. This is implied by Lemma 3.5 when a sufficiently small τ_0 is chosen, after the choice of ε_0 . \square

The $g(x; G)$ used here is more convenient than that in Chen, Tan and Zhang (2008). The proofs so far have successfully excluded the possibility that the penalized MLE of G falls in $\Gamma_1 \cup \Gamma_2$. Completing the consistency proof is a simple task.

THEOREM 3.2. *The penalized MLE of G is consistent: $\tilde{G} \rightarrow G^*$ almost surely as $n \rightarrow \infty$.*

Once the mixing distribution is restricted to Γ_3 , the KW conditions are satisfied. Hence, the restricted MLE is consistent. On this space, the penalty is of size $o_p(n)$. Hence, the penalized MLE remains consistent. Note that the KW conditions are easy to verify for finite normal mixture models on Γ_3 .

3.3 Summary

The finite normal mixture model does not satisfy the KW conditions, and the MLE as defined by (2.2) together with subsequent remarks is inconsistent. The consistency of the penalized MLE has only recently been solidly proved. Yet these facts are often overlooked; we have made these results more accessible to researchers in various disciplines. The penalized MLE under a multivariate normal mixture has also been shown to be consistent by Chen and Tan (2009) with a minor correction by Alexandrovich (2014). Note that the proof of the multivariate case can be substantially simplified using the new techniques in this paper. Finally, being consistent is a minimum requirement in statistical data analysis. The proper estimation of the mixing distribution under a finite mixture model requires a very large sample size when the subpopulations are not well separated, as indicated by the simulation study of Chen and Tan (2009).

4. CONCLUDING REMARKS

We have discussed the consistency of the MLE under mixture models given i.i.d. observations.

When there are no restrictions on the mixing distribution, the nonparametric MLE is consistent under the minimum identifiability condition (KW1) and the continuity condition (KW2) based on Pfanzagl's proof. In addition, (KW2) is implied by (W2) and (W4).

The conclusion via the KW proof is most useful when applied to finite mixture models. In this case, the KW conditions are implied by Wald conditions (W2)–(W4) and (KW1). The MLE of the mixing distribution under a finite mixture is consistent.

Under the finite normal mixture model with equal variances, the MLE is consistent. Under the finite normal mixture model in both mean and variance, the MLE is not consistent. When a penalty satisfying **P1**–**P3** is applied to the log-likelihood function, the penalized MLE is consistent.

It is curious that the consistency proof given by Pfanzagl for the nonparametric MLE under a mixture model

requires a somewhat weaker set of conditions than its parametric counterpart given by Wald: (KW1), (W2) and (W4), but not (W3). Because Pfanzagl's result is not applicable to finite mixture models, it does not cover the result of Wald or that of KW. The KW proof makes that of Wald a special case.

This paper is novel at streamlining the conditions, conclusions and the proofs related to the consistency of the MLE under mixture models. The proofs of Kiefer and Wolfowitz (1956), Pfanzagl (1988), Redner (1981), Wald (1949) are substantially simplified by connecting all of them with Theorem 2.1. The proofs focus on essential ideas and leaves complex conditions out for separate discussions. The Wald consistency result is strengthened by de-requiring $E^*|\log f(X; \theta^*)| < \infty$. The KW consistency conclusion is found most useful to finite mixture models. Its generic conclusion is less useful because its (KW3) condition is difficult to verify or not satisfied as testified by a Poisson mixture example. The paper follows the existing line of proofs for the consistency of the MLE under finite normal mixture models. When the variance is a structural parameter, the conclusion is strengthened by de-require the space of the mean parameter being compact. The consistency proof of the penalized MLE under finite normal mixture models is simplified though developing a more generic concentration inequality in Lemma 3.3 and introducing a more convenient function $g(x; G, \varepsilon)$ in Lemma 3.5.

ACKNOWLEDGMENTS

This research is supported by funding from Yunnan University through the plan of one thousand talents, and through the funding from the Natural Science and Engineering Research Council of Canada, RGPIN-2014-03743.

REFERENCES

- ALEXANDROVICH, G. (2014). A note on the article 'Inference for multivariate normal mixtures' by J. Chen and X. Tan. *J. Multivariate Anal.* **129** 245–248. [MR3215995](#)
- CHEN, H. and CHEN, J. (2003). Tests for homogeneity in normal mixtures with 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. [MR1331665](#)
- CHEN, J. and TAN, X. (2009). Inference for multivariate normal mixtures. *J. Multivariate Anal.* **100** 1367–1383. [MR2514135](#)
- CHEN, J., TAN, X. and ZHANG, R. (2008). Inference for normal mixtures in mean and variance. *Statist. Sinica* **18** 443–465. [MR2432278](#)
- CIUPERCA, G., RIDOLFI, A. and IDIER, J. (2003). Penalized maximum likelihood estimator for normal mixtures. *Scand. J. Stat.* **30** 45–59. [MR1963892](#)
- FRÜHWIRTH-SCHNATTER, S. (2011). Dealing with label switching under model uncertainty. In *Mixture Estimation and Applications* (K. L. Mengersen, C. P. Robert and D. M. Titterton, eds.) 193–218. Wiley, New York.
- GAN, L. and JIANG, J. (1999). A test for global maximum. *J. Amer. Statist. Assoc.* **94** 847–854.
- HARTIGAN, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* **2** 807–810, Wadsworth, Belmont, CA.
- HATHAWAY, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Ann. Statist.* **13** 795–800. [MR0790575](#)
- KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Ann. Math. Stat.* **27** 887–906.
- LINDSAY, B. G. (1983). The geometry of mixture likelihoods: A general theory. *Ann. Statist.* **11** 86–94. [MR0684866](#)
- LINDSAY, B. G. (1995). *Mixture Models: Theory, Geometry and Applications*. Institute for Mathematical Statistics, Hayward.
- MCLACHLAN, G. J. and PEEL, D. (2000). *Finite Mixture Models*. Wiley, New York.
- PEARSON, K. (1894). Contribution to the theory of mathematical evolution. *Philos. Trans. R. Soc. Lond. Ser. A* **186** 71–110.
- PETERS, B. C. and WALKER, H. F. (1978). An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. *SIAM J. Appl. Math.* **35** 362–378.
- PFANZAGL, J. (1988). Consistency of maximum likelihood estimators for certain nonparametric families, in particular: Mixtures. *J. Statist. Plann. Inference* **19** 137–158. [MR0944202](#)
- QUINN, B. G., MCLACHLAN, G. J. and HJORT, N. (1987). A note on the Aitkin–Rubin approach to hypothesis testing in mixture models. *J. R. Stat. Soc., B* **49** 311–314.
- REDNER, R. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Ann. Statist.* **9** 225–228. [MR0600553](#)
- REDNER, R. A., HATHAWAY, R. J. and BEZDEK, J. C. (1987). Estimating the parameters of mixture models with modal estimators. *Comm. Statist. Theory Methods* **16** 2639–2660.
- REDNER, R. A. and WALKER, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239. [MR0738930](#)
- 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.
- RUBIN, H. (1956). Uniform convergence of random functions with applications to statistics. *Ann. Math. Stat.* **27** 200–203. [MR0076246](#)
- SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, Inc., New York. [MR0595165](#)
- STEPHENS, M. (2000). Dealing with label switching in mixture models. *J. R. Stat. Soc., B* **62** 795–809.
- TEICHER, H. (1961). Identifiability of mixtures. *Ann. Math. Stat.* **32** 244–248.
- TEICHER, H. (1963). Identifiability of finite mixtures. *Ann. Math. Stat.* **34** 1265–1269. [MR0155376](#)

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, Chichester. [MR0838090](#)

WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Stat.* **20** 595–601.

WU, C.-F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11** 95–103. [MR0684867](#)