The Annals of Statistics
1995, Vol. 23, No. 1, 221–233

OPTIMAL RATE OF CONVERGENCE FOR FINITE MIXTURE MODELS[1]

BY JIAHUA CHEN

University of Waterloo

In finite mixture models, we establish the best possible rate of convergence for estimating the mixing distribution. We find that the key for estimating the mixing distribution is the knowledge of the number of components in the mixture. While a $\sqrt{n}$-consistent rate is achievable when the exact number of components is known, the best possible rate is only $n^{-1/4}$ when it is unknown. Under a strong identifiability condition, it is shown that this rate is reached by some minimum distance estimators. Most commonly used models are found to satisfy the strong identifiability condition.

1. Introduction. Let $\{f(x,\theta)\colon \theta \in \Theta\}$ be a family of densities with respect to a measure μ, and let $\mathscr{G}_m$ be the class of all m-point mixing distributions whose support points lie in the compact set Θ. A finite mixture model with m-mixing components is given by

$$f(x,G) = \int f(x,\theta)\,dG(\theta), \tag{1.1}$$

with $G \in \mathscr{G}_m$. Suppose we are doing inference based on the model assumption that $G \in \bigcup_{j=1}^{m} \mathscr{G}_j$ but the true G, call it G_0, is in $\bigcup_{j=1}^{m-1} \mathscr{G}_j$. Let $\hat{G}$ be a consistent estimator of G in the model $G \in \bigcup_{j=1}^{m} \mathscr{G}_j$. We will show that this estimator, viewed as a distribution function, cannot converge to G_0 in the $\mathscr{L}_1$-metric any faster than $n^{-1/4}$, where n is the sample size. This should be compared with the rate $n^{-1/2}$ that is possible if $G_0 \in \mathscr{G}_m$. Here the $\mathscr{L}_1$-metric is defined to be

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

The optimal rate of estimating the mixing distribution G has recently attracted much discussion when $f(x,\theta) = f(x-\theta)$. It is known that the optimal rate ranges from $(\log n)^{-1/2}$ to $n^{-1/4}$ for various distribution families of $f(x-\theta)$. See Carroll and Hall (1988), Zhang (1990) and Fan (1991) for details. The optimal rate of convergence for other mixture models remains largely unknown. The nonparametric maximum likelihood estimator of G is known to be consistent under general conditions; however, its rate of convergence remains

Received April 1991; revised March 1994.

[1]The research was supported by the Natural Sciences and Engineering Research Council of Canada.

AMS 1991 *subject classifications*. Primary 62G05; secondary 62G20.

Key words and phrases. Local asymptotic normality, maximum likelihood estimate, minimum distance, mixing distribution, mixture model, rate of convergence, strong identifiability.

almost untouched. See Kiefer and Wolfowitz (1956) and Pfanzagl (1988). For more complete references, see Titterington, Smith and Makov (1985) and Titterington (1990).

In Section 2, we will consider a simple one-parameter model, with parameter h, that is embedded in the general problem. In this parametric model, the mixing distribution has one component if $h = 0$, and has two components otherwise. We will show that the optimal rate of convergence for estimating h, when $h = 0$, is $n^{-1/4}$ using a result of Hájek. Moreover, the $\mathcal{L}_1$-distance between G_h and G_0 is proportional to h. This one-parameter problem is contained in our original problem, which proves our claim about the best possible rate. In Section 3, we will show that this rate is optimal, since it is attainable by a minimum distance estimator. In Section 4, several commonly used distribution families are shown to satisfy required conditions.

2. The best possible convergence rate. We begin by constructing a one-parameter model that captures the essential features of our problem. Let

$$G_h(\theta) = \tfrac{2}{3}\delta_{-h}(\theta) + \tfrac{1}{3}\delta_{2h}(\theta), \tag{2.1}$$

where

$$\delta_h(\theta) = \begin{cases} 0, & \theta < h, \\ 1, & \theta \geq h, \end{cases}$$

and $h \in R$. We show that the maximum likelihood estimator (MLE) of h has convergence rate $n^{-1/4}$ at $h = 0$. Note that

$$\|G_h - G_0\| = \int |G_h(\theta) - G_0(\theta)|\, d\theta = \tfrac{4}{3}h.$$

Hence, $G_{\hat{h}}$ estimates G at the rate $n^{-1/4}$ too.

PROPOSITION 1. *Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (iid) random variables with density function $f(x, G_h)$, and let*

$$l(X_1, X_2, \ldots, X_n, h) = \sum_{i=1}^{n} \log\left(\tfrac{2}{3}f(X_i, -h) + \tfrac{1}{3}f(X_i, 2h)\right)$$

be the log-likelihood function. Assume the density function $f(x, \theta)$ satisfies regularity conditions

$$E\left|\frac{f^{(i)}(X, \theta)}{f(X, \theta)}\right|^2 < \infty \quad \text{for } i = 2, 3, 4,$$

and there exists a function $g(x)$ such that

$$\left|\frac{f^{(4)}(X, \theta_1)}{f(X, \theta_1)} - \frac{f^{(4)}(X, \theta_2)}{f(X, \theta_2)}\right| \leq g(X)|\theta_1 - \theta_2|^{\varepsilon},$$

for some $\varepsilon > 0$, *and*

$$Eg^2(X) < \infty.$$

Then the MLE of h *has convergence rate* $n^{-1/4}$ *at* $h = 0$.

The proof is given in the Appendix.

The slower rate of convergence is due to the fact that G has only one support point at $h = 0$ and $\int \theta \, dG = 0$. As a consequence, the model has zero Fisher information at $h = 0$. The distribution family $\{G_h,\ h \in R\}$ is otherwise well defined in the sense that different h correspond to different G and it is smooth around $h = 0$.

Even though the MLE fails to achieve a $\sqrt{n}$ rate of convergence at $h = 0$, this does not mean it is impossible, due to the phenomenon of superefficiency. However, it will be shown that if one considers the performance of an estimator in a neighborhood of $h = 0$, then no estimator can uniformly over that neighborhood achieve a rate better than $n^{-1/4}$. For this purpose, let us introduce the following simplified definition of local asymptotic normality (LAN) from Ibragimov and Has'minski (1981) and a result from Hájek (1972).

DEFINITION 1. Let $X_1, X_2, \ldots, X_n, \ldots$ be iid random variables with the density function belonging to $\{f(x,\theta),\ \theta \in \Theta\}$. The latter is called locally asymptotically normal at the point $\theta_0 \in \Theta$ as $n \to \infty$ if, for some $\varphi(n) = \varphi(n, \theta_0)$ and any $u \in R$, the representation

$$\begin{aligned} Z_{n,\theta_0}(u) &= \frac{\prod_{i=1}^n f(X_i, \theta_0 + \varphi(n)u)}{\prod_{i=1}^n f(X_i, \theta_0)} \\ &= \exp\left\{ u Z_n - \frac{1}{2}u^2 + \psi_n(u, \theta_0) \right\} \end{aligned} \tag{2.2}$$

is valid, where

$$Z_n \to_d N(0,1) \quad \text{as } n \to \infty$$

under $\theta = \theta_0$, and, moreover, for any $u \in R$ we have

$$\psi_n(u, \theta) \to 0 \quad \text{when } \theta = \theta_0$$

in probability as $n \to \infty$.

Distribution families which satisfy the LAN condition have special properties. Let $w(t)$ be a nonnegative symmetric function of t which is continuous at 0, not identically 0 and $w(0) = 0$. Further, assume the sets $\{t\colon\ w(t) < c\}$ are convex for all $c > 0$ and are bounded for all sufficiently small $c > 0$. The following result is attributed to Hájek (1972).

THEOREM 0. *Let* $\{f(x,\theta), \theta \in \Theta\}$ *satisfy the LAN condition at the point* $\theta_0 \in \Theta$ *with the normalizing value* $\varphi(n)$ *and let* $\varphi(n) \to 0$ *as* $n \to \infty$. *Then, for any family of estimators* T_n *of* θ *and any* $\varepsilon > 0$, *the inequality*

$$\liminf_{n\to\infty} \sup_{|\theta-\theta_0|<\varepsilon} E_\theta\{w[\varphi^{-1}(n)(T_n-\theta)]\} \geq \frac{1}{\sqrt{2\pi}} \int w(x) \exp\left\{-\frac{1}{2}x^2\right\} dx$$

is valid.

The theorem we have just described shows that when an arbitrarily small neighborhood of a parameter point is considered, no estimator can do better than what is described above. The $\varphi(n)$ is the best possible rate of convergence at the point. Under this consideration, we will show that the best possible rate for the mixture model (1.1) is $n^{-1/4}$ at some parameter points under some regularity conditions.

LEMMA 1. *Suppose in the model* $f(x, G_h)$ *with* G_h *given by* (2.1), *the density function* $f(x,\theta)$ *satisfies the same regularity conditions given in Proposition* 1 *with the moment requirements increased to the third order. Then* $\{f(x, Q_t),\ Q_t = G_{t/\sqrt{|t|}}\}$ *satisfies the LAN condition at* $t = 0$ *with*

$$\varphi = \varphi(n) = n^{-1/2}\left[E\left(\frac{f''(X_i,0)}{f(X_i,0)}\right)^2\right]^{-1/2}.$$

The proof is given in the Appendix.

The fundamental way that this LAN analysis differs from the standard parametric theory is that here the first derivative of the log-likelihood in h is identically zero, regardless of the data, at $h = 0$, so that higher-order derivatives come into play in the asymptotic analysis and one must switch from parameter h to parameter t to obtain a $\sqrt{n}$ rate.

With this result, we obtain the following theorem.

THEOREM 1. *Suppose that in model* (1.1), *the true mixing distribution* $G \in \bigcup_{j=1}^{m-1} \mathscr{G}_i$. *Then the optimal rate of convergence* $\varphi(n)$ *for estimating* G *is at most* $n^{-1/4}$.

PROOF. The general model contains the submodel given in Lemma 1. The best possible rate for estimating t in the submodel is $\varphi(n) = n^{-1/2}$. Since $\|Q_t - Q_0\| = O(t^{1/2})$, the best possible rate for estimating Q_t or, in general, G is $O(n^{-1/4})$. □

3. The optimal rate. We have shown in a finite mixture model when the number of components is known up to an upper bound, the best possible convergence rate is $n^{-1/4}$. We show in this section that the convergence rate $n^{-1/4}$ is achievable.

Before stating our main results, we need to introduce some concepts of identifiability. The mixture model (1.1) is identifiable if $F(x, G_1) = F(x, G_2)$ implies $G_1 = G_2$. When (1.1) is restricted to finite mixtures, this property is called finitely identifiable.

DEFINITION 2. The family $\{F(x, \theta), \theta \in \Theta\}$ is called strongly identifiable if F is twice differentiable in θ and for any m and m different $\theta_1, \ldots, \theta_m$, the equality

$$\sup_x \left| \sum_{j=1}^{m} [\alpha_j F(x, \theta_j) + \beta_j F'(x, \theta_j) + \gamma_j F''(x, \theta_j)] \right| = 0.$$

implies that $\alpha_j = \beta_j = \gamma_j = 0$ for $j = 1, 2, \ldots, m$.

Clearly the strong identifiability implies finite identifiability if F is twice differentiable. We will show that when a distribution family is strongly identifiable the sup norm distance between two finite mixtures is at least proportional to the size of the squared $\mathscr{L}_1$-distance between their mixing distributions. This enables us to estimate the mixing distribution with the desired convergence rate.

Also, we define

$$\psi(G_1, G_2) = \begin{cases} \sup_x |F(x, G_1) - F(x, G_2)| / \|G_1 - G_2\|^2, & \text{if } G_1 \neq G_2, \\ \infty, & \text{if } G_1 = G_2. \end{cases}$$

Then we have the following lemma.

LEMMA 2. *If, for all x, $F(x, \theta)$ is twice differentiable with respect to θ with second derivative $F''(x, \theta)$ satisfying a uniform Lipschitz condition*

$$|F''(x, \theta_1) - F''(x, \theta_2)| \leq c|\theta_1 - \theta_2|^{\varepsilon}, \tag{3.1}$$

for all x, θ_1, θ_2 and some fixed c and $\varepsilon > 0$, and if $\{F(x, \theta), \theta \in \Theta\}$ is strongly identifiable and Θ is compact, then, for fixed G_0,

$$\lim_{\delta_1 \to 0} \lim_{\delta_2 \to 0} \inf \{\psi(G_1, G_2): \|G_1 - G_0\| \leq \delta_1, \ \|G_2 - G_1\| \leq \delta_2\} > 0, \tag{3.2}$$

where G_0, G_1 and G_2 have at most $m < \infty$ components.

The proof is given in the Appendix.

In terms of this lemma, the minimum distance type estimators can be constructed to achieve the best rate of convergence. Let $F_n(x)$ be the empirical distribution function constructed from iid samples $X_1, X_2, \ldots, X_n$ with distribution function given by (1.1). Let $\hat{G}_n$ be a distribution function on Θ such

that

$$\sup_x |F(x, \hat{G}_n) - F_n(x)| \le \inf_G \sup_x |F(x, G) - F_n(x)| + \frac{1}{n}, \tag{3.3}$$

where $\inf_G$ is taken over all the distribution functions with at most m support points. The term n^{-1} on the right-hand side is used to ensure the existence of $\hat{G}_n$.

THEOREM 2. *Under the conditions of Lemma 2,*

$$\|\hat{G}_n - G_1\| = O_p(n^{-1/4}),$$

in probability under $F(x, G_1)$ uniformly for G_1 such that $\|G_1 - G_o\| \le \varepsilon$, where G_o is a fixed mixing distribution with at most m components.

PROOF. By (3.3), we have

$$\begin{aligned} &\sup_x |F(x, \hat{G}_n) - F(x, G_1)| \\ &\quad \le \sup_x |F(x, \hat{G}_n) - F_n(x)| + \sup_x |F_n(x) - F(x, G_1)| \\ &\quad \le 2 \sup_x |F_n(x) - F(x, G_1)| + \frac{1}{n} = O_p(n^{-1/2}). \end{aligned} \tag{3.4}$$

Note that $\sup_x |F_n(x) - F(x, G_1)| = O_p(n^{-1/2})$ is the Kolmogorov–Smirnov distance. Clearly, $\hat{G}_n$ has to converge to G_1. Otherwise, from the compactness of Θ, there would be a subsequence of $\hat{G}_n$ which converges to $G_2 \ne G_1$. This would lead to $\sup_x |F(x, G_1) - F(x, G_2)| = 0$, which contradicts the identifiability. Lemma 2 then implies

$$\|\hat{G}_n - G_1\|^2 \le c \sup_x |F_n(x) - F(x, G_1)| = O_p(n^{-1/2}). \qquad \square$$

The above minimum distance type estimator of G was also discussed in Deely and Kruse (1968). They found that the estimator is consistent and that efficient linear programming algorithms can be used to construct estimators. They did not, however, discuss the rate of convergence.

4. Conditions on moments and identifiability. In previous sections, a best possible rate of convergence for estimating G in model (1.1) is found and it was shown that this rate is achievable. In this section, we present some results on the strong identifiability of mixture models. Some commonly used models are shown to satisfy the conditions required by Theorems 1 and 2.

THEOREM 3. *Suppose that $f(x)$ is a differentiable density function and $F(x, \theta) = \int_{-\infty}^{x} f(t - \theta)\, dt$. If $\lim_{x \to \pm\infty} f(x) = \lim_{x \to \pm\infty} f'(x) = 0$, then $F(x, \theta)$ satisfies the strong identifiability condition.*

PROOF. We need to show that if

$$\sum_{j=1}^{m}[\alpha_j F(x,\theta_j) + \beta_j F'(x,\theta_j) + \gamma_j F''(x,\theta_j)] = 0, \tag{4.1}$$

for any x, then α_j, β_j and γ_j are all 0. If (4.1) is true, we have

$$\int \exp\{itx\}\, d\left\{\sum_{j=1}^{m}[\alpha_j F(x,\theta_j) + \beta_j F'(x,\theta_j) + \gamma_j F''(x,\theta_j)]\right\} = 0,$$

where $i^2 = -1$. Hence

$$\sum_{j=1}^{m}[\alpha_j - \beta_j(it) + \gamma_j(it)^2]\exp\{it\theta_j\}\int \exp\{itx\}\, dF(x) = 0.$$

Since $\int \exp\{itx\}\, dF(x)$ equals 1 when $t = 0$ and is continuous at the point, we have

$$\sum_{j=1}^{m}[\alpha_j - \beta_j(it) + \gamma_j(it)^2]\exp\{it\theta_j\} = 0,$$

for t in a neighborhood of 0. Since this is an analytic function of it, it must be 0 for all t. Multiplying it by $\exp\{-\frac{1}{2}t^2\}$ and taking the inverse Fourier transformation, we obtain

$$\sum_{j=1}^{m}[\alpha_j - \beta_j H_1(x-\theta_j) + \gamma_j H_2(x-\theta_j)]\exp\{-\tfrac{1}{2}(x-\theta_j)^2\} = 0,$$

for all x, where $H_1(x)$ and $H_2(x)$ are Hermite polynomials. Observe that, when $x \to \infty$, one of $\exp\{-\frac{1}{2}(x-\theta_j)^2\}$ tends to 0 with the slowest rate, hence its corresponding polynomial $\alpha_j - \beta_j H_1(x-\theta_j) + \gamma_j H_2(x-\theta_j)$ must be 0. This implies that all α_j, β_j and γ_j equal 0. Hence the theorem is proved. □

COROLLARY. *The conclusion of Theorem* 3 *remains true when*

$$F(x,\theta) = \frac{1}{\theta}\int_{-\infty}^{x} f\left(\frac{t}{\theta}\right) dt,$$

where θ *is in* $(0,\infty)$.

PROOF. If a random variable X has distribution $F(x)$, then $Y = |X|$ has distribution $F(x) - F(-x)$. Further, the distribution of $\log Y$ belongs to a location model [see also Teicher (1961)]. Hence Theorem 3 applies and it proves the corollary. □

By Theorem 3 and straightforward calculations, the location and scale families of normal and Cauchy distributions satisfy conditions of Theorem 2.

The following calculations illustrate that the Poisson distribution family also satisfies these conditions: since

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

we have

$$\frac{f''(x,\theta)}{f(x,\theta)} = \frac{x(x-1)}{\theta^2} - \frac{2x}{\theta} + 1,$$

$$\frac{f^{(3)}(x,\theta)}{f(x,\theta)} = \frac{x(x-1)(x-2)}{\theta^3} - \frac{3x(x-1)}{\theta^2} + \frac{3x}{\theta} - 1$$

and

$$\frac{f^{(4)}(x,\theta)}{f(x,\theta)} = \frac{x(x-1)(x-2)(x-3)}{\theta^4} - \frac{4x(x-1)(x-2)}{\theta^3} + \frac{6x(x-1)}{\theta^2} - \frac{4x}{\theta} + 1.$$

Clearly, the moment and Lipschitz conditions of Theorems 1 and 2 are satisfied. Now let us examine the strong identifiability. In this model, we have

$$F(x,\theta) = \sum_{i=0}^{x} \frac{\theta^i}{i!}\exp\{-\theta\}.$$

Thus, if

$$\sum_{j=1}^{m}[\alpha_j F(x,\theta_j) + \beta_j F'(x,\theta_j) + \gamma_j F''(x,\theta_j)] = 0, \tag{4.2}$$

for any x, we have to show that all α_j, β_j and γ_j are 0. By calculating the moment generating function of (4.2), we obtain

$$\sum_{j=1}^{m}[(\alpha_j - \beta_j + \gamma_j) + (\beta_j - 2\gamma_j)\exp\{t\} + \gamma_j \exp\{2t\}]\exp\{\theta_j(e^t - 1)\} = 0, \tag{4.3}$$

for any t. Suppose θ_m is the largest among θ_j's. Then $\exp\{\theta_m(e^t - 1) + 2t\}$ goes to ∞ with the fastest speed as t goes to ∞. So γ_m must be 0 because of (4.3). When this is the case, $\exp\{\theta_m(e^t - 1) + t\}$ becomes the fastest one, hence β_m must be 0 and so on. Repeating this procedure, we conclude that if (4.2) and hence (4.3) hold, all the α_j, β_j and γ_j are 0. This shows that the Poisson mixture model is strongly identifiable.

APPENDIX

PROOF OF PROPOSITION 1. The first two derivatives of $l(X_1, X_2, \ldots, X_n, h)$ (with respect to h, the same for f) are

$$l'(h) = \sum_{i=1}^{n} \frac{-2f'(X_i, -h) + 2f'(X_i, 2h)}{2f(X_i, -h) + f(X_i, 2h)}$$

and

$$l''(h) = \sum_{i=1}^{n} \frac{2f''(X_i, -h) + 4f''(X_i, 2h)}{2f(X_i, -h) + f(X_i, 2h)} - \sum_{i=1}^{n} \left(\frac{2f'(X_i, 2h) - 2f'(X_i, -h)}{2f(X_i, -h) + f(X_i, 2h)} \right)^2.$$

Let us try to find the local maximum of the likelihood for which h is closest to 0. Note that no other local maxima, if any, can converge faster. Since $l'(0) = 0$ for any set of observations, $\hat{h} = 0$ is a local maximum when $l''(0) < 0$. Note that

$$l''(0) = 2 \sum_{i=1}^{n} \frac{f''(X_i, 0)}{f(X_i, 0)}.$$

Let E_0 denote the expectation under the distribution corresponding to $h = 0$. We have

$$E_0 l''(0) = 0.$$

Under regularity conditions, we can apply the central limit theorem and hence

$$P_0\{l''(0) \leq 0\} \to \tfrac{1}{2} \quad \text{as } n \to \infty.$$

Thus, the MLE $\hat{h} = 0$ when $l''(0) < 0$, which has probability $\frac{1}{2}$ if $h = 0$. However, when $l''(0) > 0$, we need to calculate $l^{(3)}(0)$ and $l^{(4)}(0)$ to locate the local maximum point of $l(h)$.

By some simple calculations, we obtain

$$l^{(3)}(0) = 2 \sum_{i=1}^{n} \frac{f^{(3)}(X_i, 0)}{f(X_i, 0)}$$

and

$$l^{(4)}(0) = 6 \sum_{i=1}^{n} \frac{f^{(4)}(X_i, 0)}{f(X_i, 0)} - 12 \sum_{i=1}^{n} \left(\frac{f''(X_i, 0)}{f(X_i, 0)} \right)^2.$$

Let

$$A_i = \frac{f''(X_i, 0)}{f(X_i, 0)}, \qquad B_i = \frac{f^{(3)}(X_i, 0)}{f(X_i, 0)}, \qquad C_i = \frac{f^{(4)}(X_i, 0)}{f(X_i, 0)}.$$

Under regularity conditions,

$$E_0 A_i = 0, \qquad E_0 B_i = 0, \qquad E_0 C_i = 0;$$

$$E_0 A_i^2 < \infty, \qquad E_0 B_i^2 < \infty, \qquad E_0 C_i^2 < \infty;$$

so we have

$$\sum_{i=1}^{n} A_i = O_p(n^{1/2}), \qquad \sum_{i=1}^{n} B_i = O_p(n^{1/2}), \qquad \sum_{i=1}^{n} C_i = O_p(n^{1/2}).$$

With this fact and Taylor's expansion, we find

$$l(h) = l(0) + \sum_{i=1}^{n} A_i h^2 + \frac{1}{3}\sum_{i=1}^{n} B_i h^3 - \frac{1}{2}\sum_{i=1}^{n} A_i^2 h^4 + O_p(n^{1/2}h^4),$$

as $h \to 0$. Differentiating and setting equal to zero, we find an approximating cubic equation, one of whose roots is zero. Since we have assumed $\sum_{i=1}^{n} A_i = \frac{1}{2}l''(0) > 0$, the nearest local maximum to $h = 0$ must be one of the other roots

$$\left[\sum_{i=1}^{n} B_i \pm \left(\left(\sum_{i=1}^{n} B_i\right)^2 + 16\sum_{i=1}^{n} A_i^2 \sum_{i=1}^{n} A_i\right)^{1/2}\right] \times \left[4\sum_{i=1}^{n} A_i^2\right]^{-1} (1 + o_p(1)).$$

Unless $E_0[f''(x,\theta)/f(x,\theta)]^2 = 0$, the average $n^{-1}\sum_{i=1}^{n} A_i^2$ tends to a positive constant and, hence,

$$\hat{h} = \delta_0\left(\sum_{i=1}^{n} A_i\right)\left[\sum_{i=1}^{n} A_i^2\right]^{-1/2}\left[\sum_{i=1}^{n} A_i\right]^{1/2}[1 + o_p(1)] = O_p(n^{-1/4}). \qquad \square$$

PROOF OF LEMMA 1. Denote $\sigma^2 = E[f''(x,\theta)/f(x,\theta)]^2$. Using the same notation as in Definition 1, when $u \geq 0$ ($u < 0$ is similar), we have

$$\log Z_{n,0}(u) = \sum_{i=1}^{n} \log\left\{1 + \frac{2f(X_i, -\sqrt{\varphi u}) + f(X_i, 2\sqrt{\varphi u}) - 3f(X_i, 0)}{3f(X_i, 0)}\right\}.$$

Let

$$Y_i = \frac{2f(X_i, -\sqrt{\varphi u}) + f(X_i, 2\sqrt{\varphi u}) - 3f(X_i, 0)}{3f(X_i, 0)}.$$

Then we have the following expansion:

$$Y_i = \frac{f''(X_i, 0)}{f(X_i, 0)}(\varphi u) + \frac{1}{3}\frac{f^{(3)}(X_i, 0)}{f(X_i, 0)}(\varphi u)^{3/2} + \frac{1}{4}\frac{f^{(4)}(X_i, 0)}{f(X_i, 0)}(\varphi u)^2 + O_p[(\varphi u)^{2+\varepsilon/2}]g(X_i).$$

By simple calculation, using $\varphi = n^{-1/2}\sigma^{-1}$ and $E_0[f^{(i)}/f] = 0$, we find

$$\sum_{i=1}^{n} Y_i = n^{-1/2}\sigma^{-1}u\sum_{i=1}^{n} \frac{f''(X_i, 0)}{f(X_i, 0)} + o_p(1),$$

$$\sum_{i=1}^{n} Y_i^2 = n^{-1}\sigma^{-2}u^2\sum_{i=1}^{n}\left(\frac{f''(X_i, 0)}{f(X_i, 0)}\right)^2 + o_p(1)$$

and

$$\sum_{i=1}^{n} |Y_i|^3 = O_p(n^{-1/2}) = o_p(1).$$

Note that, when $\min\{Y_i\} > -\frac{1}{2}$,

$$\left|\log(1+Y_i) - Y_i + \tfrac{1}{2}Y_i^2\right| \le |Y_i|^3.$$

Hence, under the same condition,

$$\begin{aligned}\log Z_{n,0} &= \sum_{i=1}^{n} \log(1+Y_i) = \sum_{i=1}^{n} Y_i - \frac{1}{2}\sum Y_i^2 + C\sum_{i=1}^{n} |Y_i|^3 \\ &= n^{-1/2}\sigma^{-1}u \sum_{i=1}^{n} \frac{f''(X_i,0)}{f(X_i,0)} - \frac{1}{2}n^{-1}\sigma^{-2}u^2 \sum_{i=1}^{n}\left(\frac{f''(X_i,0)}{f(X_i,0)}\right)^2 + o_p(1),\end{aligned} \tag{A.1}$$

where C is a bounded random variable. At the same time, by the Markov inequality, the finite third moment conditions and the expansion of Y_i, we get

$$P\left(\min\{Y_i\} < -\tfrac{1}{2}\right) \le \sum_{i=1}^{n} P\left(Y_i < -\tfrac{1}{2}\right) \le 8\sum_{i=1}^{n} E|Y_i|^3 = O(n^{-1/2}).$$

Hence (A.1) holds with probability tending to 1, which implies that the LAN condition is satisfied. □

PROOF OF LEMMA 2. If (3.2) is not true, there will be sequences of G_{n1} and G_{n2} tending to G_0 and making $\psi(G_{n1}, G_{n2})$ converge to 0. Note that

$$\begin{aligned}\psi(G_{n1}, G_{n2}) &= \sup_x \left| \int_0^1 \left\{F\left(x, G_{n1}^{-1}(u)\right) - F\left(x, G_{n2}^{-1}(u)\right)\right\} du \right| / \|G_{n1} - G_{n2}\|^2 \\ &= \sup_x \Bigg| \int_{O_n^c} \left\{F\left(x, G_{n1}^{-1}\right) - F\left(x, G_{n2}^{-1}\right)\right\} du \\ &\qquad + \int_{O_n} F'\left(x, G_{n2}^{-1}\right)\left\{G_{n1}^{-1} - G_{n2}^{-1}\right\} du \\ &\qquad + \tfrac{1}{2}\int_{O_n} F''\left(x, G_{n2}^{-1}\right)\left\{G_{n1}^{-1} - G_{n2}^{-1}\right\}^2 du \\ &\qquad\qquad + R_n(x) \Bigg| / \|G_{n1} - G_{n2}\|^2 \\ &= \sup_x |A_n(x) + B_n(x) + C_n(x) + R_n(x)| / D_n,\end{aligned} \tag{A.2}$$

where

$$O_n = \left\{u:\ 0 \le u \le 1;\ \left|G_{n1}^{-1}(u) - G_{n2}^{-1}(u)\right| \le \|G_{n1} - G_{n2}\|^{1/2}\right\},$$

O_n^c is its complement and

$$R_n(x) = o\left(\int_{O_n} \{G_{n1}^{-1}(u) - G_{n2}^{-1}(u)\}^2 \, du\right) = o\left(\int_0^1 \{G_{n1}^{-1}(u) - G_{n2}^{-1}(u)\}^2 \, du\right)$$

because of (3.1) and the definition of set O_n. The set O_n is needed because $|G_{n1}^{-1}(u) - G_{n2}^{-1}(u)|$ may not converge to zero for some u. The terms $A_n(x)$, $B_n(x)$ and $C_n(x)$ are linear combinations of $F(x,\theta)$, $F'(x,\theta)$ and $F''(x,\theta)$ for different θ's, respectively. Since Θ is bounded, we can select a subsequence of G_{n1} and G_{n2} further such that they have fixed numbers of components and each of their support points converges to a fixed point in Θ. Hence, after being properly rescaled, the limits of $A_n(x)$, $B_n(x)$ and $C_n(x)$ are still linear combinations of these functions with constant coefficients (not depending on x).

This implies $C_n(x)/D_n \to \sum_{j=1}^m \gamma_j F''(x,\theta_j)$ for some γ_j and not all of them vanishing and with $\theta_j \in \Theta$. The coefficients in $A_n(x)/D_n$ and $B_n(x)/D_n$ can go either to infinity or to a constant by further selecting a subsequence of G's. If they go to infinity, a sequence $d_n = O(1)$ can then be found such that $d_n A_n(x)/D_n$ converges to $\sum_{j=1}^m \alpha_j F(x,\theta_j)$ and $d_n B_n(x)/D_n$ converges to $\sum_{j=1}^m \beta_j F'(x,\theta_j)$ for some noninfinite α_j and β_j and not all of them vanishing. Hence, in any case, we have d_n and $\alpha_j, \beta_j, \gamma_j$ not all zero (although γ_j may have been changed if multiplied by d_n), such that

$$d_n \frac{|\int_0^1 \{F(x, G_{n1}^{-1}(u)) - F(x, G_{n2}^{-1}(u))\} \, du|}{\|G_{n1} - G_{n2}\|^2}$$

$$\to \left|\sum_{j=1}^{m'} [\alpha_j F(x,\theta_j) + \beta_j F'(x,\theta_j) + \gamma_j F''(x,\theta_j)]\right|,$$

for some integer m'. By the strong identifiability, the supremum of the right-hand side of the above equation is nonzero, which contradicts $\psi(G_{n1}, G_{n2}) \to 0$. □

Acknowledgments. I would like to thank Professor J. D. Kalbfleisch, who motivated this research, and Professor C. F. J. Wu for many helpful discussions. Special thanks go to a referee who suggested a new approach in proving Theorem 2 which strengthened the theorem and simplified the proof, and to an Associate Editor for the constructive criticisms which substantially improved the presentation of this paper.

REFERENCES

CARROLL, R. J. and HALL, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* **83** 1184–1186.

DEELY, J. J. and KRUSE, R. L. (1968). Construction of sequences estimating the mixing distribution. *Ann. Math. Statist.* **39** 286–288.

FAN, J. Q. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19** 1257–1272.

HÁJEK, J. (1972). Local asymptotic minimax and admissibility in estimation. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **1** 175–194. Univ. California Press, Berkeley.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation, Asymptotic Theory*. Springer, New York.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Ann. Math. Statist.* **27** 887–906.

PFANZAGL, J. (1988). Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures. *J. Statist. Plann. Inference* **19** 137–158.

TEICHER, H. (1961). Identifiability of mixtures. *Ann. Math. Statist.* **32** 244–248.

TITTERINGTON, D. M. (1990). Some recent research in the analysis of mixture distributions. *Statistics.* **21** 619–641.

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

ZHANG, C. H. (1990). Fourier methods for estimating mixing densities and distributions. *Ann. Statist.* **18** 806–831.

DEPARTMENT OF STATISTICS
AND ACTUARIAL SCIENCE
UNIVERSITY OF WATERLOO
WATERLOO N2L 3G1
CANADA