The Annals of Statistics
1990, Vol. 18, No. 4, 1868–1877

LARGE DEVIATION PROBABILITIES FOR CERTAIN NONPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATORS

BY J. PFANZAGL

University of Cologne

Let $(X, \mathscr{A})$ be a measurable space and $\{P_{\vartheta,\tau}|\mathscr{A}: \vartheta \in \Theta, \tau \in \mathrm{T}\}$ a family of probability measures. Given an appropriate estimator sequence for ϑ, we define a sequence of asymptotic maximum likelihood estimators for τ and give bounds for its large deviation probabilities under conditions which are natural for the application to the estimation of mixing distributions.

This paper generalizes earlier results of Pfanzagl to the following cases: (i) estimator sequences restricted to a sieve; (ii) estimator sequences using a given estimator sequence for a nuisance parameter; (iii) convergence under the "wrong model;" (iv) large deviation probabilities instead of consistency.

1. The results. For $(\vartheta, \tau) \in \Theta \times \mathrm{T}$ let $P_{\vartheta,\tau}|\mathscr{A}$ be a p-measure. For $(\vartheta, \alpha) \in \Theta \times \mathrm{A}$, let $m(\cdot, \vartheta, \alpha)$: $X \to (0, \infty)$ be a measurable function. Assume that for every $(\vartheta_0, \tau_0) \in \Theta \times \mathrm{T}$ there exists $\alpha_0 \in \mathrm{A}$ such that

$$(1.1)\quad P_{\vartheta_0,\tau_0}(\log[m(\cdot, \vartheta_0, \alpha_0)/m(\cdot, \vartheta_0, \alpha)]) > 0 \quad \text{for all } \alpha \in \mathrm{A}, \alpha \neq \alpha_0.$$

Our problem is to estimate α_0, based on an i.i.d. sample from $P^n_{\vartheta_0,\tau_0}$, if a preliminary estimator sequence $\vartheta^{(n)}$: $X^n \to \Theta$ for ϑ is available.

The natural idea is to use some sort of "maximum" estimators for α, e.g., to choose $\alpha^{(n)}(\mathbf{x})$ such that

$$(1.2)\quad \sum_1^n \log\left[m\left(x_\nu, \vartheta^{(n)}(\mathbf{x}), \alpha^{(n)}(\mathbf{x})\right)/m\left(x_\nu, \vartheta^{(n)}(\mathbf{x}), \alpha\right)\right] \geq 0 \quad \text{for all } \alpha \in \mathrm{A}.$$

In the following we give conditions under which estimator sequences of this kind are consistent.

The family $\{P_{\vartheta,\tau}: (\vartheta, \tau) \in \Theta \times \mathrm{T}\}$ was introduced only for indicating the framework within which the following results can be applied. The consistency theorem itself deals with a fixed p-measure $P_0|\mathscr{A}$ (which is P_{ϑ_0,τ_0} in the applications, and the assumptions on P_0 have to be fulfilled for any P_{ϑ_0,τ_0} in the family).

We assume that Θ is a first countable Hausdorff space and A a first countable compact Hausdorff space which is also a convex subset of a linear space. (No connection between the Hausdorff topology and linear operation is required.)

Received January 1989; revised October 1989.
AMS 1980 *subject classifications*. Primary 62F10, 62F12.
Key words and phrases. Maximum likelihood estimators, consistency, nonparametric models, mixtures.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics*.
www.jstor.org

The following conditions refer to a fixed element ϑ_0.

(1.3) For every $\hat{\alpha} \in \mathrm{A}$ the map $(\vartheta, \alpha) \to m(x, \vartheta, \alpha)$ is continuous at $(\vartheta, \alpha) = (\vartheta_0, \hat{\alpha})$, for P_0-a.a. $x \in X$.

(1.4) For P_0-a.a. $x \in X$ and all $\vartheta \in \Theta$ the map $\alpha \to m(x, \vartheta, \alpha)$ is concave.

(1.5) There exists $\alpha_0 \in \mathrm{A}$ such that

$$P_0(\log[m(\cdot, \vartheta_0, \alpha_0)/m(\cdot, \vartheta_0, \alpha)]) > 0 \quad \text{for } \alpha \in \mathrm{A},\ \alpha \neq \alpha_0.$$

The following proposition shows that (1.5) follows from (1.3) and (1.4) under mild conditions. Observe that no extra effort is needed if $\mathrm{A} \subset \mathbb{R}^m$. Since $\alpha \to P_0(\log m(\cdot, \vartheta_0, \alpha))$ is concave, continuity follows in this case.

PROPOSITION. *If* $\sup_{\alpha \in \mathrm{A}} P_0(\log m(\cdot, \vartheta_0, \alpha))$ *is finite, then* (1.3) *and* (1.4) *imply the existence of* $\alpha_0 \in \mathrm{A}$ *such that*

$$P_0(\log[m(\cdot, \vartheta_0, \alpha_0)/m(\cdot, \vartheta_0, \alpha)]) \geq 0 \quad \textit{for all } \alpha \in \mathrm{A}.$$

If α_0 *is identifiable in the sense that* "$m(x, \vartheta_0, \alpha) = m(x, \vartheta_0, \alpha_0)$ *for* P_0-*a.a.* $x \in X$ *implies* $\alpha = \alpha_0$," *then this inequality is strict for* $\alpha \in \mathrm{A}$, $\alpha \neq \alpha_0$.

For the *preliminary estimator sequence* $\vartheta^{(n)}$, $n \in \mathbb{N}$, we assume the following.

For every neighborhood U of ϑ_0 there exist $a > 0$ and $\delta \in (0, 1)$ such that for all $n \in \mathbb{N}$,

$$P_0^n\{\mathbf{x} \in X^n : \vartheta^{(n)}(\mathbf{x}) \notin U\} \leq a\delta^n. \tag{1.6}$$

To have the result as versatile as possible, the definition of the estimator sequence for α_0 will be based on a "sieve."

(1.7) A_n, $n \in \mathbb{N}$, is a nondecreasing sequence of convex and compact subsets of A; $\bigcup_1^\infty \mathrm{A}_n$ is dense in A.

In the following theorem we obtain large deviation probabilities without any restriction on the rate at which the sequence A_n, $n \in \mathbb{N}$, increases. [Compare with Geman and Hwang (1982), who introduce such restrictions in connection with a more general case.]

The estimator $\alpha^{(n)}|X^n$ maps into A_n. To obtain consistency, a condition weaker than (1.2) suffices:

(1.8) There exists $\gamma \in (0, 1]$ such that for all $n \in \mathbb{N}$, $\mathbf{x} \in X^n$ and $\alpha \in \mathrm{A}_n$,

$$\prod_1^n m(x_\nu, \vartheta^{(n)}(\mathbf{x}), \alpha^{(n)}(\mathbf{x})) \geq \gamma \prod_1^n m(x_\nu, \vartheta^{(n)}(\mathbf{x}), \alpha).$$

THEOREM. *Assume conditions* (1.3)–(1.7). *Then any estimator sequence* $\alpha^{(n)}$, $n \in \mathbb{N}$, *fulfilling* (1.8) *has the following property*: *For any neighborhood* V *of* α_0 *there exist* $a > 0$ *and* $\delta \in (0, 1)$ *such that for all* $n \in \mathbb{N}$,

$$P_0^n\{\mathbf{x} \in X^n: \alpha^{(n)}(\mathbf{x}) \notin V\} \le a\delta^n. \tag{1.9}$$

REMARK 1. Relation (1.9) implies

$$P_0^{\mathbb{N}} \bigcup_{m=n}^{\infty} \{\mathbf{x} \in X^{\mathbb{N}}: \alpha^{(m)}(\mathbf{x}) \notin V\} \le \frac{a}{1-\delta}\delta^n$$

[if we now interpret $\alpha^{(m)}$ as a map on $X^{\mathbb{N}}$, depending on $(x_1, \dots, x_m)$ only].

Hence, estimator sequences fulfilling (1.9) are, in particular, strongly consistent, i.e., $\lim_{n\to\infty} \alpha^{(n)}(\mathbf{x}) = \alpha_0$ for $P_0^{\mathbb{N}}$-a.a. $\mathbf{x} \in X^{\mathbb{N}}$. A slight modification of the proof shows that this also holds true if condition (1.6) is replaced by $\lim_{n\to\infty} \vartheta^{(n)}(\mathbf{x}) = \vartheta_0$ for $P_0^{\mathbb{N}}$-a.a. $\mathbf{x} \in X^{\mathbb{N}}$.

REMARK 2. The specific feature which distinguishes this theorem from theorems on the consistency of maximum likelihood estimators in general is the concavity of $\alpha \to m(x, \vartheta, \alpha)$. The intended application is to the estimation of mixing distributions. Such applications have been discussed extensively in Pfanzagl [(1988), Sections 5 and 6]. Hence, we restrict ourselves to a few additional remarks.

Let H be a locally compact Hausdorff space with countable base, endowed with the Borel algebra. Let $\{P_{\vartheta,\eta}|\mathscr{A}: \vartheta \in \Theta, \eta \in \mathrm{H}\}$ be a family of mutually absolutely continuous p-measures. Let $\mu|\mathscr{A}$ be a dominating σ-finite measure. If Γ is a finite measure over H, the Γ-mixture is defined as $P_{\vartheta,\Gamma}(A) := \int P_{\vartheta,\eta}(A)\Gamma(d\eta)$, $A \in \mathscr{A}$. Without loss of generality, we assume that the densities are positive and finite on X. Throughout the following we assume that $(x, \eta) \to p(x, \vartheta, \eta)$ is measurable for every $\vartheta \in \Theta$. Then the function $x \to p(x, \vartheta, \Gamma) := \int p(x, \vartheta, \eta)\Gamma(d\eta)$ is a μ-density of $P_{\vartheta,\Gamma}$. (Observe that such a product measurable version of the densities always exists if $\mathscr{A}$ is countably generated [see, e.g. Strasser (1985), page 17 ff.]. Here we need continuity of $\eta \to p(x, \vartheta, \eta)$ anyway. Then *any* version of these densities is product measurable [see Pfanzagl and Wefelmeyer (1985), page 451, Lemma 3.1.12]).

The basic family $\{P_{\vartheta,\tau}: \vartheta \in \Theta, \tau \in \mathrm{T}\}$ is now $\{P_{\vartheta,\Gamma}: \vartheta \in \Theta, \Gamma \in \mathscr{G}_0\}$, if $\mathscr{G}_0$ denotes the class of all probability measures over H. Our intention is to apply the theorem with $p(\cdot, \vartheta, \Gamma)$ taking the role of $m(\cdot, \vartheta, \alpha)$. The map $\Gamma \to p(\cdot, \vartheta, \Gamma)$ is linear, hence, in particular, concave. The theorem requires, moreover, conditions of a topological nature: continuity of $(\vartheta, \alpha) \to m(x, \vartheta, \alpha)$ and compactness of A. Hence we use for A the class of all sub-probability measures over H, say $\mathscr{G}$. $\mathscr{G}$ is convex, and, endowed with the vague topology, a compact metrizable space [see Bauer (1981), page 243, Corollary 7.8.3 and Theorem 7.8.4].

One could think, of course, of using other topologies. If, for every $x \in X$, $\vartheta \in \Theta$, the function $\eta \to p(x, \vartheta, \eta)$ is in $\mathscr{C}_0(\mathrm{H})$, the class of continuous functions which vanish at infinity, the vague topology is, however, the natural one.

If Γ is identifiable in $\mathscr{G}$, then the linear closure of $\{p(x, \vartheta, \cdot): x \in X\}$ (with $\vartheta \in \Theta$ fixed) is dense in $\mathscr{C}_0(\mathrm{H})$ with respect to the sup-norm. [See Blum and Susarla (1977), page 201, Theorem 2.1.] Hence, continuity of $\Gamma \to \int p(x, \vartheta, \eta)\Gamma(d\eta)$ for all $x \in X$ implies continuity of $\Gamma \to \int h(\eta)\Gamma(d\eta)$ for all $h \in \mathscr{C}_0(\mathrm{H})$. Therefore, the vague topology is the smallest possible in this case.

Conditions for the identifiability of Γ in the case of exponential families $\{P_{\vartheta,\eta}: \eta \in \mathrm{H}\}$ with $\mathrm{H} \subset \mathbb{R}^m$ can be found in Pfanzagl [(1988), page 152, Proposition 6.2].

A is not necessarily to be taken as all of $\mathscr{G}$. One can use for A any convex and closed subset of $\mathscr{G}$ (which may or may not contain the true Γ_0). Since $P_{\vartheta_0,\Gamma_0}(\log p(\cdot, \vartheta_0, \Gamma)) \le P_{\vartheta_0,\Gamma_0}(\log p(\cdot, \vartheta_0, \Gamma_0)) < \infty$ for all $\Gamma \in \mathscr{G}$, the Proposition guarantees that the function $\Gamma \to P_{\vartheta_0,\Gamma_0}(\log p(\cdot, \vartheta_0, \Gamma))$ attains its supremum on any convex and closed subset of $\mathscr{G}$.

The definition of the estimator sequence $\Gamma^{(n)}$, $n \in \mathbb{N}$, is based on a sieve. In the particular case of mixtures such a sieve consists of convex and compact subsets of sub-probability measures $\mathscr{G}_n \subset \mathscr{G}$. In addition to these general properties (see (1.7)) we require that $\Gamma \in \mathscr{G}_n$ implies $\Gamma/\Gamma(\mathrm{H}) \in \mathscr{G}_n$. This guarantees that one can always take the estimator fulfilling (1.8) to be a probability measure. As a particular example of such a sieve, consider the case where H is a separable metric space with a countably dense subset $\{\eta_k: k \in \mathbb{N}\}$. Then one can take for $\mathscr{G}_n$ the class of all sub-probability measures with support $\{\eta_1, \dots, \eta_n\}$. It follows from Parthasarathy [(1967), page 44, Theorem 6.3] that $\bigcup_1^\infty \mathscr{G}_n$ is dense in $\mathscr{G}$ with respect to the vague topology.

REMARK 3. Whereas condition (1.8) suffices for consistency of the resulting estimator sequence, in applications it appears natural to determine estimators fulfilling the stronger condition (1.2), with A replaced by A_n. Under a slightly stronger version of conditions (1.3) and (1.4) [both conditions for *all* $x \in X$, and (1.3) for *all* $\vartheta \in \Theta$), condition (1.2), with A replaced by A_n, is equivalent to the following condition:

$$n^{-1} \sum_1^n m(x_\nu, \vartheta^{(n)}(\mathbf{x}), \alpha)/m(x_\nu, \vartheta^{(n)}(\mathbf{x}), \alpha^{(n)}(\mathbf{x})) \le 1 \quad \text{for all } \alpha \in \mathrm{A}_n. \tag{1.10}$$

Condition (1.2) requires that $\alpha^{(n)}(\mathbf{x})$ maximize $\alpha \to \sum_1^n \log m(x_\nu, \vartheta^{(n)}(\mathbf{x}), \alpha)$ for $\alpha \in \mathrm{A}_n$. Apart from the technical question whether the supremum is attained, this poses no problem. As against that, the existence of an $\alpha^{(n)}(\mathbf{x})$ fulfilling (1.10), and its identity with the $\alpha^{(n)}(\mathbf{x})$ fulfilling (1.2), is anything but obvious. It depends on the concavity of $\alpha \to m(x, \vartheta, \alpha)$ and can be established by some sort of minimax theorem. [See, e.g., Lindsay (1983), Part I, for the mixture model.]

With the lemma below at our disposal, this can easily be seen as follows. Let n and $\mathbf{x} = (x_1, \dots, x_n) \in X^n$ be fixed. Let $\alpha^{(n)}(\mathbf{x}) \in \mathrm{A}_n$ be a value fulfilling (1.2) with A replaced by A_n (which exists by the proposition stated previously).

For $\nu \in \{1, \dots, n\}$ let $h(\nu, \alpha) := m(x_\nu, \vartheta^{(n)}(\mathbf{x}), \alpha)/m(x_\nu, \vartheta^{(n)}(\mathbf{x}), \alpha^{(n)}(\mathbf{x}))$. Applying the lemma with $L(t) = \log t$, $\alpha_0 = \alpha^{(n)}(\mathbf{x})$ and $Q\{\nu\} = 1/n$ for $\nu \in \{1, \dots, n\} = Z$, we obtain

$$\sum_1^n L\left(\frac{m\left(x_\nu, \vartheta^{(n)}(\mathbf{x}), \alpha\right)}{m\left(x_\nu, \vartheta^{(n)}(\mathbf{x}), \alpha^{(n)}(\mathbf{x})\right)}\right) \le 0 \tag{1.11}$$

for all $\alpha \in \mathrm{A}_n$, and all functions L which are strictly increasing, concave and fulfill $L(1) = 0$. Relation (1.10) follows with $L(t) = t - 1$. Since the function "log" is concave, (1.2) follows immediately from (1.10).

The fact that we may switch in (1.11) from one function L to another one offers the possibility of using one function L for the numerical computation of $\alpha^{(n)}(\mathbf{x})$, and another function L for studying the asymptotic behavior of the estimator sequence $\alpha^{(n)}$, $n \in \mathbb{N}$. In the case of the mixture model, $\Gamma^{(n)}$ can be determined by the E.M. algorithm using (1.10), based on $L(t) = t - 1$, whereas the proof of the theorem uses $L(t) = (1 + t^{-1})^{-1} - \frac{1}{2}$.

2. Proofs.

LEMMA. *Let $(Z, \mathscr{C})$ be a measurable space, and $Q|\mathscr{C}$ a probability measure. Let A be a convex set, and $h: Z \times \mathrm{A} \to [0, \infty)$ a function such that*

(i) *$z \to h(z, \alpha)$ is measurable for every $\alpha \in \mathrm{A}$,*
(ii) *$\alpha \to h(z, \alpha)$ is concave for Q-a.a. $z \in Z$,*
(iii) *there exists $\alpha_0 \in \mathrm{A}$ such that $h(z, \alpha_0) \equiv 1$ for Q-a.a. $z \in Z$.*

Let $\mathscr{L}$ denote the class of all strictly increasing and concave functions L: $[0, \infty) \to [-\infty, \infty)$ with $L(1) = 0$, and $\mathscr{L}_0 \subset \mathscr{L}$ the subclass of functions which are differentiable in a neighborhood of 1, with a derivative which is continuous at 1.

If there exists $L \in \mathscr{L}_0$ such that

$$\int L(h(z, \alpha))Q(dz) \le 0 \quad \text{for all } \alpha \in \mathrm{A}, \tag{2.1}$$

then this relation holds for all $L \in \mathscr{L}$.

If L is strictly concave, then (2.1) implies strict inequality for all $\alpha \in \mathrm{A}$, except for the case of $h(\cdot, \alpha) \equiv 1$ Q-a.e.

PROOF. (i) We shall show that relation (2.1) for some $L \in \mathscr{L}_0$ implies

$$\int h(z, \alpha)Q(dz) \le 1 \quad \text{for all } \alpha \in \mathrm{A}. \tag{2.2}$$

Since $h(z, \cdot\,)$ is concave, we obtain for $\alpha \in \mathrm{A}$, $u \in [0, 1]$,

$$h(z, u\alpha + (1 - u)\alpha_0) \ge 1 + u(h(z, \alpha) - 1).$$

Hence (2.1) implies for $\alpha \in \mathrm{A}$, $u \in (0, 1]$,

$$\int L[1 + u(h(z, \alpha) - 1)]Q(dz) \le 0.$$

For any $c \geq 1$, $h(z, \alpha) > c$ implies $L[1 + u(h(z, \alpha) - 1)] > L(1) = 0$, hence also

$$\int_{h(\cdot, \alpha) \leq c} L[1 + u(h(z, \alpha) - 1)] Q(dz) \leq 0. \tag{2.3}$$

In the following, $\alpha \in \mathrm{A}$ is fixed. For $\varepsilon > 0$ there exists $\delta_\varepsilon > 0$ such that

$$|L'(1 + y) - L'(1)| < \varepsilon \quad \text{for } |y| < \delta_\varepsilon.$$

Since, for some $\eta \in (0, 1)$,

$$\begin{aligned} L(1 + y) &= L(1) + yL'(1 + \eta y) \\ &= L'(1)y + y[L'(1 + \eta y) - L'(1)], \end{aligned}$$

$|y| < \delta_\varepsilon$ implies

$$L'(1)y \leq L(1 + y) + |y|\varepsilon.$$

Let $c > 1$ be fixed. For all $u \in (0, \delta_\varepsilon / c)$ the relation $h(z, \alpha) \leq c$ implies $u|h(z, \alpha) - 1| < \delta_\varepsilon$, hence

$$L'(1)u(h(z, \alpha) - 1) \leq L[1 + u(h(z, \alpha) - 1)] + uc\varepsilon. \tag{2.4}$$

From (2.3) and (2.4) we obtain for all $u \in (0, \delta_\varepsilon / c)$,

$$L'(1)u \int_{h(\cdot, \alpha) \leq c} (h(z, \alpha) - 1) Q(dz) \leq uc\varepsilon,$$

and therefore,

$$L'(1) \int_{h(\cdot, \alpha) \leq c} (h(z, \alpha) - 1) Q(dz) \leq c\varepsilon. \tag{2.5}$$

Since $\varepsilon > 0$ and $c > 1$ are arbitrary and $L'(1) > 0$, relation (2.2) follows.

(ii) Since L is concave and increasing, we obtain from Jensen's inequality and (2.2) that

$$\begin{aligned} \int L(h(z, \alpha)) Q(dz) &\leq L\left(\int h(z, \alpha) Q(dz)\right) \\ &\leq L(1) = 0, \end{aligned} \tag{2.6}$$

which is (2.1).

(iii) It remains to be shown that the inequality in (2.1) is strict if L is strictly concave, and not $h(\cdot, \alpha) \equiv 1$ Q-a.e.

If L is strictly concave, the first inequality in (2.6) is strict unless $h(z, \alpha) = \int h(\xi, \alpha) Q(d\xi)$ for Q-a.a. $z \in Z$. The second inequality in (2.6) is strict unless $\int h(\xi, \alpha) Q(d\xi) = 1$. Hence, equality in (2.1) implies $h(z, \alpha) = 1$ for Q-a.a. $z \in Z$. □

PROOF OF THE PROPOSITION. Since ϑ_0 remains fixed, it will be omitted throughout this proof.

(i) Let $\varepsilon_n \downarrow 0$. For $n \in \mathbb{N}$ choose $\alpha_n \in \mathrm{A}$ such that

$$P_0(\log m(\cdot, \alpha_n)) > s - \varepsilon_n, \quad \text{with } s := \sup_{\alpha \in \mathrm{A}} P_0(\log m(\cdot, \alpha)).$$

Without loss of generality we may assume that $\alpha_0 := \lim_{n\to\infty} \alpha_n$ exists. We shall show that

$$s = P_0(\log m(\cdot, \alpha_0)).$$

For any $\alpha \in \mathrm{A}$, $u \in [0,1]$,

$$\begin{aligned} P_0(\log m(\cdot, \alpha_n)) > s - \varepsilon_n &\geq P_0(\log m(\cdot, (1-u)\alpha_n + u\alpha)) - \varepsilon_n \\ &\geq P_0(\log[(1-u)m(\cdot, \alpha_n) + um(\cdot, \alpha)]) - \varepsilon_n. \end{aligned}$$

Hence,

$$(2.7) \quad P_0\left(\log\left[1 + u\left(\frac{m(\cdot, \alpha)}{m(\cdot, \alpha_n)} - 1\right)\right]\right) \leq \varepsilon_n \quad \text{for all } \alpha \in \mathrm{A},\ u \in [0,1].$$

To simplify our notations, let

$$h_n(x) := \frac{m(x, \alpha)}{m(x, \alpha_n)} - 1, \qquad n = 0, 1, 2, \dots .$$

Let $t_n \uparrow \infty$, $t_n > 1$. For all $u \in [0,1]$,

$$1_{\{h_n \leq t_n\}} \log(1 + uh_n) \leq \log(1 + uh_n),$$

hence,

$$(2.8) \qquad P_0\big(1_{\{h_n \leq t_n\}} \log(1 + uh_n)\big) \leq \varepsilon_n.$$

Since $\log(1+z) \geq z - z^2/(1+z)$ for $z > -1$, we obtain from $h_n(x) > -1$ for $u \in (0, \frac{1}{2})$,

$$\log(1 + uh_n(x)) \geq uh_n(x) - 2u^2 h_n(x)^2,$$

and therefore,

$$(2.9) \qquad 1_{\{h_n \leq t_n\}}(x) \log(1 + uh_n(x)) \geq uh_n(x) 1_{\{h_n \leq t_n\}}(x) - 2u^2 t_n^2.$$

From (2.8) and (2.9) we obtain for $u \in (0, \frac{1}{2})$,

$$uP_0\big(h_n 1_{\{h_n \leq t_n\}}\big) \leq \varepsilon_n + 2u^2 t_n^2.$$

Taking $u = \varepsilon_n^{2/3}$, $t_n = \varepsilon_n^{-1/6}$, we obtain

$$(2.10) \qquad \limsup_{n\to\infty} P_0\big(h_n 1_{\{h_n \leq t_n\}}\big) \leq 0.$$

Since $h_n \to h_0$ P_0-a.e., we have

$$h_n 1_{\{h_n \leq t_n\}} \to h_0 \qquad P_0\text{-a.e.}$$

Since $h_n 1_{\{h_n \leq t_n\}} \geq -1$, this implies by Fatou's lemma

$$(2.11) \qquad P_0(h_0) \leq \liminf_{n\to\infty} P_0\big(h_n 1_{\{h_n \leq t_n\}}\big).$$

Together with (2.10) this implies

$$P_0(h_0) \le 0.$$

Expressed in the original notation,

$$P_0\left(\frac{m(\cdot,\alpha)}{m(\cdot,\alpha_0)}\right) \le 1 \quad \text{for all } \alpha \in \mathrm{A}.$$

By the lemma this implies

$$P_0\left(\log\left[\frac{m(\cdot,\alpha_0)}{m(\cdot,\alpha)}\right]\right) \ge 0 \quad \text{for all } \alpha \in \mathrm{A}.$$

(ii) It remains to be shown that this inequality is strict. Since $\alpha \to m(x,\alpha)$ is concave for P_0-a.a. $x \in X$, so is $\alpha \to \log m(x,\alpha)$. Hence,

$$\begin{aligned} P_0\left(\log\left[\frac{m(\cdot,\alpha_0)}{m(\cdot,\alpha)}\right]\right) &\ge P_0\left(\log\left[\frac{m(\cdot,\frac{1}{2}\alpha_0+\frac{1}{2}\alpha)}{m(\cdot,\alpha)}\right]\right) \\ &\ge \frac{1}{2}P_0\left(\log\left[\frac{m(\cdot,\alpha_0)}{m(\cdot,\alpha)}\right]\right), \end{aligned}$$

where the last inequality is strict unless $m(\cdot,\alpha_0) = m(\cdot,\alpha)$ P_0-a.e. Hence, $\alpha \ne \alpha_0$ implies strict inequality, and $P_0(\log[m(\cdot,\alpha_0)/m(\cdot,\alpha)]) > 0$ follows. □

Proof of the theorem. (i) The conditions of the lemma are fulfilled for $h(x,\alpha) = m(x,\vartheta_0,\alpha)/m(x,\vartheta_0,\alpha_0)$ [see (1.4) and (1.5)], $L(t) = \log t$ (the function in $\mathscr{L}_0$) and $L(t) = (1+t^{-1})^{-1} - \frac{1}{2}$ (a strictly concave function in $\mathscr{L}$). Since α_0 is identifiable, we obtain

$$\int\left(1+\frac{m(x,\vartheta_0,\alpha_0)}{m(x,\vartheta_0,\alpha)}\right)^{-1} P_0(dx) < \frac{1}{2}. \tag{2.12}$$

Since $(1+m(x,\vartheta_0,\alpha_0)/m(x,\vartheta_0,\alpha))^{-1} \le 1$, Fatou's lemma can be used to infer from condition (1.3) the existence of open neighborhoods $U_\alpha \ni \vartheta_0$, $V_\alpha \ni \alpha_0$ and $W_\alpha \ni \alpha$ such that

$$\int\left(1+\frac{\underline{m}(x,U_\alpha,V_\alpha)}{\overline{m}(x,U_\alpha,W_\alpha)}\right)^{-1} P_0(dx) < \frac{1}{2}, \tag{2.13}$$

where

$$\begin{aligned} \underline{m}(x,U,V) &:= \inf\{m(x,\vartheta,\alpha): \vartheta \in U, \alpha \in V\} \\ \overline{m}(x,U,V) &:= \sup\{m(x,\vartheta,\alpha): \vartheta \in U, \alpha \in V\}. \end{aligned}$$

(ii) From (1.8),

$$\sum_1^n \log\left[\frac{m(x_\nu,\vartheta^{(n)}(\mathbf{x}),\alpha^{(n)}(\mathbf{x}))}{m(x_\nu,\vartheta^{(n)}(\mathbf{x}),\hat{\alpha})}\right] \ge \log\gamma \quad \text{for all } \hat{\alpha} \in \mathrm{A}_n. \tag{2.14}$$

From (1.4) we obtain for all $\hat{\alpha} \in \mathrm{A}_n$ and P_0^n-a.a. $\mathbf{x} \in X^n$,

$$\tfrac{1}{2}\left(m\left(x_\nu, \vartheta^{(n)}(\mathbf{x}), \hat{\alpha}\right) + m\left(x_\nu, \vartheta^{(n)}(\mathbf{x}), \alpha^{(n)}(\mathbf{x})\right)\right) \\ \leq m\left(x_\nu, \vartheta^{(n)}(\mathbf{x}), \tfrac{1}{2}\hat{\alpha} + \tfrac{1}{2}\alpha^{(n)}(\mathbf{x})\right).$$

Since A_n is convex, $\hat{\alpha} \in \mathrm{A}_n$ implies $\frac{1}{2}\hat{\alpha} + \frac{1}{2}\alpha^{(n)}(\mathbf{x}) \in \mathrm{A}_n$. Hence, (2.14) implies for $\hat{\alpha} \in \mathrm{A}_n$ and P_0^n-a.a. $\mathbf{x} \in X^n$,

$$\sum_1^n \log\left[2\left(1 + \frac{m\left(x_\nu, \vartheta^{(n)}(\mathbf{x}), \hat{\alpha}\right)}{m\left(x_\nu, \vartheta^{(n)}(\mathbf{x}), \alpha^{(n)}(\mathbf{x})\right)}\right)^{-1}\right] \geq \log \gamma. \tag{2.15}$$

(iii) Let V_0 be an arbitrary open neighborhood of α_0 in A. Since A is compact, so is $\mathrm{A} - V_0$, and the open cover $\{W_\alpha\colon \alpha \in \mathrm{A} - V_0\}$ contains a finite subcover, say $\{W_{\alpha_1}, \dots, W_{\alpha_r}\}$. Let $U := \bigcap_{i=1}^r U_{\alpha_i}$, $V := \bigcap_{i=1}^r V_{\alpha_i}$.

The following relations hold for $i = 1, \dots, r$:

$$\underline{m}(x, U, V) \geq \underline{m}\left(x, U_{\alpha_i}, V_{\alpha_i}\right)$$

$$\overline{m}\left(x, U, W_{\alpha_i}\right) \leq \overline{m}\left(x, U_{\alpha_i}, W_{\alpha_i}\right),$$

hence also

$$\int\left(1 + \frac{\underline{m}(x, U, V)}{\overline{m}\left(x, U, W_{\alpha_i}\right)}\right)^{-1} P_0(dx) \\ \leq \int\left(1 + \frac{\underline{m}\left(x, U_{\alpha_i}, V_{\alpha_i}\right)}{\overline{m}\left(x, U_{\alpha_i}, W_{\alpha_i}\right)}\right)^{-1} P_0(dx). \tag{2.16}$$

Let

$$\delta := \max_{i=1,\dots,r} 2\int\left(1 + \frac{\underline{m}(x, U, V)}{\overline{m}\left(x, U, W_{\alpha_i}\right)}\right)^{-1} P_0(dx). \tag{2.17}$$

From (2.13), (2.16) and (2.17)

$$\delta \in [0, 1). \tag{2.18}$$

(iv) We have

$$P_0^n\{\mathbf{x} \in X^n\colon \alpha^{(n)}(\mathbf{x}) \notin V_0\} \leq P_0^n\{\mathbf{x} \in X^n\colon \vartheta^{(n)}(\mathbf{x}) \notin U\} \\ + \sum_{i=1}^r P_0^n\left\{\mathbf{x} \in X^n\colon \vartheta^{(n)}(\mathbf{x}) \in U,\ \alpha^{(n)}(\mathbf{x}) \in W_{\alpha_i}\right\}. \tag{2.19}$$

Now, $\vartheta^{(n)}(\mathbf{x}) \in U$ and $\alpha^{(n)}(\mathbf{x}) \in W_{\alpha_i}$ imply

$$m\left(x_\nu, \vartheta^{(n)}(\mathbf{x}), \alpha^{(n)}(\mathbf{x})\right) \leq \overline{m}\left(x_\nu, U, W_{\alpha_i}\right).$$

If n is large enough, we have $\mathrm{A}_n \cap V \neq \varnothing$, hence,

$$m\left(x_\nu, \vartheta^{(n)}(\mathbf{x}), \hat{\alpha}\right) \geq \underline{m}\left(x_\nu, U, V\right) \quad \text{for some } \hat{\alpha} \in \mathrm{A}_n.$$

Therefore, $\vartheta^{(n)}(\mathbf{x}) \in U$ and $\alpha^{(n)}(\mathbf{x}) \in W_{\alpha_i}$ imply by (2.15),

$$\sum_{\nu=1}^{n} \log\left[2\left(1 + \frac{\underline{m}(x_\nu, U, V)}{\overline{m}(x_\nu, U, W_{\alpha_i})}\right)^{-1}\right] \geq \log \gamma.$$

Hence, by Markov's inequality, the following relation holds with δ defined by (2.17):

$$\begin{aligned} &P_0^n\{\mathbf{x} \in X^n : \vartheta^{(n)}(\mathbf{x}) \in U, \alpha^{(n)}(\mathbf{x}) \in W_{\alpha_i}\} \\ (2.20) \qquad &\leq P_0^n\left\{\mathbf{x} \in X^n : \sum_{\nu=1}^{n} \log\left[2\left(1 + \frac{\underline{m}(x_\nu, U, V)}{\overline{m}(x_\nu, U, W_{\alpha_i})}\right)^{-1}\right] \geq \log \gamma\right\} \\ &\leq \gamma^{-1}\delta^n. \end{aligned}$$

From (2.19) and (2.20),

$$P_0^n\{\mathbf{x} \in X^n : \alpha^{(n)}(\mathbf{x}) \notin V_0\} \leq P_0^n\{\mathbf{x} \in X^n : \vartheta^{(n)}(\mathbf{x}) \notin U\} + r\gamma^{-1}\delta^n,$$

for all sufficiently large $n \in \mathbb{N}$. Since $\delta \in [0, 1)$ by (2.18), this proves the assertion. □

Acknowledgments. The author wishes to thank Mr. R. Hamböker for carefully checking the proofs. The comments of the referees contributed to making the paper more readable and to eliminating a slip.

REFERENCES

BAUER, H. (1981). *Probability Theory and Elements of Measure Theory*. Academic, London.

BLUM, J. R. and SUSARLA, V. (1977). Estimation of a mixing distribution function. *Ann. Probab.* **5** 200–209.

GEMAN, S. and HWANG, C. R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10** 401–414.

LINDSAY, B. G. (1983). The geometry of mixture likelihoods. Part I: A general theory. Part II: The exponential family. *Ann. Statist.* **11** 86–94 and 783–792.

PARTHASARATHY, K. R. (1967). *Probability Measures on Metric Spaces*. Academic, London.

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

PFANZAGL, J. and WEFELMEYER, W. (1985). *Asymptotic Expansions for General Statistical Models. Lecture Notes in Statist.* **31**. Springer, New York.

STRASSER, H. (1985). *Mathematical Theory of Statistics: Statistical Experiments and Asymptotic Theory. Studies in Math.* **7**. de Gruyter, Berlin.

MATHEMATICAL INSTITUTE
UNIVERSITY OF COLOGNE
WEYERTAL 86
5000 COLOGNE 41
WEST GERMANY