JOURNAL OF MULTIVARIATE ANALYSIS 43, 133–146 (1992)

Existence and Consistency of Maximum Likelihood in Upgraded Mixture Models

A. W. VAN DER VAART*

Faculteit Wiskunde en Informatica, Vrije Universiteit,
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

AND

JON A. WELLNER

Department of Statistics GN-22, University of Washington,
Seattle, Washington 98195

Communicated by the Editors

Suppose one observes a sample of size m from the mixture density $\int p(x \mid z)\, d\eta(z)$ and a sample of size n from the distribution η. The kernel $p(x \mid z)$ is known. We show existence of the maximum likelihood estimator for η, characterize its support, and prove consistency as $m, n \to \infty$. © 1992 Academic Press, Inc.

1. INTRODUCTION

If one observes a sample of independent, identically distributed variables $Z_1, ..., Z_n$ from a completely unknown distribution η, then the usual estimator for η is the empirical distribution $\hat{\eta} = n^{-1} \sum_{j=1}^{n} \delta_{Z_j}$. Consider the situation wherein the observed $Z_1, ..., Z_n$ are actually part of a larger number $m+n$ of replications of some experiment. Unfortunately, m out of the $m+n$ times the Z value is not observed, but instead one gets to see X which conditionally on $Z=z$ has a known density $p(x \mid z)$ with respect to a fixed measure μ. Hence the total set of observations is $X_1, ..., X_m$, $Z_1, ..., Z_n$; all observations are independent and their joint distribution can formally be written as

$$\prod_{i=1}^{m} \int p(x_i \mid z)\, d\eta(z) \prod_{j=1}^{n} d\eta(z_j).$$

Received February 2, 1991; revised February 14, 1992.

AMS 1980 subject classifications: 62G20, 62F12.

Key words and phrases: mixture model, maximum likelihood, consistency.

* Research partially supported by NSF Grant DMS-8723011.

0047-259X/92 $5.00

(The first factor in the product is a density with respect to μ^n; the second factor is just formal notation.) For definiteness let $(\mathscr{X}, \mathscr{A})$ and $(\mathscr{Z}, \mathscr{C})$ be the sample spaces for each X_i and Z_j, respectively. It is assumed throughout that the function $(x, z) \to p(x \mid z)$ is (jointly) measurable and also that $\mathscr{C}$ contains the one-point sets in $\mathscr{Z}$. The density of X_i is (with abuse of notation) written as $p(x \mid \eta) = \int p(x \mid z)\, d\eta(z)$; it is assumed to be finite for every x.

In this situation the set $Z_1, ..., Z_n$ clearly contains much more information about η than the set $X_1, ..., X_m$ if m and n are of comparable magnitude. Nevertheless, one would certainly want to take all information available in $X_1, ..., X_m$ into account and obtain improved estimator for η relative to using $\hat{\eta}$, the empirical distribution of the second sample. Formal calculations in the manner of semi-parametric theory (cf. Bickel, Klaassen, Ritov, and Wellner [3]) show that a decrease in asymptotic variance of as much as $m/(m+n)$ percent is possible, depending of course on the aspect of η one is interested in. Surprisingly enough there may even be considerable gain in using $X_1, ..., X_m$ in situations where the information (in the technical sense of semi-parametric theory) in $X_1, ..., X_m$ alone is zero and $\sqrt{m}$ consistent estimators based on the first sample do not exist.

It is thus of interest to study estimators for η based on the whole set of observations $X_1, ..., X_m$, $Z_1, ..., Z_n$. In this paper we limit ourselves to showing that maximum likelihood estimators exist and are asymptotically consistent under weak conditions on the "kernel" $p(x \mid z)$. We intend to study asymptotic efficiency of the maximum likelihood estimator in a later paper, using different methods.

The model as defined here, or special cases thereof, has been studied by Has'minskii and Ibragimov [6], Bhanja and Ghosh [2], Vardi [14], and Vardi and Zhang [15]. In the literature the type of distribution of each X_i is called a *mixture model* and sometimes a *structural model*. Estimation of η based on $X_1, ..., X_m$ alone has been considered by among others Kiefer and Wolfowitz [9], Laird [10], Jewell [8], Heckman and Singer [7], and van der Vaart [13]. The problem of existence of the maximum likelihood estimator in mixture models is solved by Lindsay [11] and the problem of consistency of Pfanzagl [12]. Roughly, the proofs in the present paper are carried through by conditioning on the "good" observations $Z_1, ..., Z_n$ and next extending the arguments as developed for mixture models by these authors. We do not discuss computation of the maximum likelihood estimator. However, modifications of methods developed for pure mixture models, for instance those motivated by the EM-algorithm or the methods of Groeneboom [5], apply.

2. Existence and Support

For a measure η write $\eta\{z\}$ for the mass that η gives to the one-point set $\{z\}$. In this section $x_1, ..., x_m, z_1, ..., z_n$ are fixed (observed) values. For our purposes the *likelihood function* is the map

$$\eta \rightarrow \prod_{i=1}^{m} p(x_i | \eta) \prod_{j=1}^{n} \eta\{z_j\}.$$

A maximum likelihood estimator $\tilde{\eta}$ is a probability distribution that maximizes the likelihood function. In this section it is shown that $\tilde{\eta}$ always exists (in other words the supremum is achieved) and can be taken finitely discrete with no more than $m+n$ support points. (We do not address uniqueness, but note that in some examples the maximum likelihood estimator is clearly nonunique and also distributions with a continuous component may maximize the likelihood.)

The main conditions are expressed in terms of the following subsets of $\mathbb{R}^m$,

$$U = \{(p(x_1 | z), ..., p(x_m | z)) : z \in \mathscr{Z}\}$$
$$V = \{\alpha u : 0 \leqslant \alpha \leqslant 1, u \in U\}$$
$$W = \{(p(x_1 | \eta), ..., p(x_m | \eta)) : \eta \in \mathscr{H}\},$$

where $\mathscr{H}$ is the set of all subprobability measures on $(\mathscr{Z}, \mathscr{C})$. (The positive measures with total mass less than or equal to 1.) It is clear that $U \subset V \subset W$. Furthermore, for every finitely discrete subprobability distribution η the corresponding element in W is a convex linear combination of elements of V. If we write conv(V) for the convex hull of V and $\overline{\text{conv}}(V)$ for its closure, then it is also true that

$$\text{conv}(V) \subset W \subset \overline{\text{conv}}(V).$$

Here the last inclusion is a consequence of the geometric form of Jensen's inequality: the random variable $(p(x_1 | Z), ..., p(x_m | Z))$ takes its values in the closed convex set $\overline{\text{conv}}(V)$, so its expectation under η is in this set too. It is well known that the convex hull of a compact subset of $\mathbb{R}^m$ is automatically compact, hence closed. Consequently, if V is compact, then so is W and the inclusions in the last display are equalities. It may be noted also that if U is compact, then so is V.

THEOREM 2.1. *If W is compact in $\mathbb{R}^m$, then there exists a probability distribution $\tilde{\eta}$ which maximizes the likelihood function. Moreover, If V is compact in $\mathbb{R}^m$, then $\tilde{\eta}$ can be taken a discrete distribution with between n and $m+n$ support points.*

Proof. Maximization of the likelihood can be carried out in two steps. First fix $p_1, ..., p_n$ for the point masses $\eta\{z_j\}$ and maximize

$$\prod_{i=1}^{m} p(x_i \mid \eta)$$

over all subprobability distributions η with $\eta\{z_j\} = p_j$ for every j. Suppose the maximum value is $m_{\mathbf{p}}$ and taken for $\eta_{\mathbf{p}}$. Then in the second step maximize

$$m_{\mathbf{p}} \prod_{j=1}^{n} p_j$$

over nonnegative subprobability vectors $\mathbf{p} = (p_1, ..., p_n)$. If the maximum value is taken for $\tilde{p}$, then $\eta_{\tilde{p}}$ is a maximum likelihood estimator.

For fixed $p_1, ..., p_n$ the first problem is equivalent to maximization of the function $v \to \prod_{i=1}^{m} v_i$ over all vectors v in the set $W_{\mathbf{p}}$ given by

$$W_{\mathbf{p}} = \left(\sum_{j=1}^{n} p_j p(x_1 \mid z_j), ..., \sum_{j=1}^{n} p_j p(x_m \mid z_j) \right) + \left(1 - \sum_{j=1}^{n} p_j \right) W.$$

(Here $u + \alpha W$ is the set of all vectors of the form $u + \alpha W$ with $w \in W$.) If W is compact, then so is $W_{\mathbf{p}}$. Consequently, the maximum is taken for some $v \in W_{\mathbf{p}}$. Also, the set $W_{\mathbf{p}}$ depends continuously on $\mathbf{p}$, so that the maximum value of the everywhere continuous function $v \to \prod_{i=1}^{m} v_i$ depends continuously on $\mathbf{p}$. This implies that the second maximization problem, consisting of maximizing $\mathbf{p} \to m_{\mathbf{p}} \prod_{j=1}^{n} p_j$ over the compact set of all subprobability vectors $\mathbf{p}$, has a solution too. This concludes the proof of existence of the maximum likelihood estimator.

For the proof of the second part of the theorem let $\tilde{p}$ be the vector of probabilities $(\tilde{\eta}\{z_1\}, ..., \tilde{\eta}\{z_n\})$ and let $\tilde{w} \in W$ be such that $v \to \prod_{i=1}^{m} v_i$ is maximized at

$$v = \left(\sum_{j=1}^{n} \tilde{p}_j p(x_1 \mid z_j), ..., \sum_{j=1}^{n} \tilde{p}_j p(x_m \mid z_j) \right) + \left(1 - \sum_{j=1}^{n} \tilde{p}_j \right) \tilde{w},$$

in the first maximization problem. Thus the function

$$w \to \prod_{i=1}^{m} \left(\sum_{j=1}^{n} \tilde{p}_j p(x_i \mid z_j) + \left(1 - \sum_{j=1}^{n} \tilde{p}_j \right) w_i \right)$$

is maximized over W at $w = \tilde{w}$. Since this function is convex and W is convex, the point $\tilde{w}$ must be on the boundary of W. Since W is the closed convex hull of the compact set V, every point on its boundary is expressible as a convex linear combination of at most m elements of V. Hence $\tilde{w}$ can be written

$$\tilde{w} = \sum_{i=1}^{m} q_i(p(x_1 \mid y_i), ..., p(x_m \mid y_i))$$

for a subprobability vector $\mathbf{q}$ and suitable $y_1, ..., y_m$. Then the discrete measure $\tilde{\eta}$ with $\tilde{\eta}\{z_j\} = \tilde{p}_j$ and $\tilde{\eta}\{y_i\} = (1 - \sum_{j=1}^{n} \tilde{p}_j) q_i$ for each j and i maximizes the likelihood function. Renorming the vector $\mathbf{q}$ so that it is a probability vector will lead to a likelihood that is certainly not smaller. Hence $\tilde{\eta}$ may be assumed to be a probability measure. ∎

The conditions on V and W in the existence theorem are usually satisfied and can easily be checked directly. (The pictures one can draw for $m = 2$ often give a good indication of how to approach the problem.) Alternatively, a large class of examples can be handled through continuity of the functions $z \to p(x \mid z)$. Recall that a metric space is called locally compact if every point has a compact neighbourhood (a compact set containing a ball around the point). Any such space has a one-point compactification, written $\mathscr{Z} \cup \{\infty\}$. A function $f\colon \mathscr{Z} \to \mathbb{R}$ is said to *vanish at infinity* if $\lim_{z \to \infty} f(z) = 0$. More explicitly f vanishes at infinity if for every $\varepsilon > 0$ there is a compact $K \subset \mathscr{Z}$ with $|f(z)| < \varepsilon$ if $z \notin K$. The set of all continuous functions that vanish at infinity is denoted $C_0(\mathscr{Z})$. Examples of locally compact metric spaces are $\mathbb{R}^k$, closed or open subsets of $\mathbb{R}^k$ and cells $(c, d]$. Each of these examples is also separable and a function vanishes at infinity if its value converges to zero as the argument approaches an open boundary. (The open boundary as a whole, if there is one, may be considered the point ∞.)

LEMMA 2.2. *Let $\mathscr{Z}$ be a locally compact separable metric space with Borel σ-field $\mathscr{C}$. Suppose that for each fixed x the function $z \to p(x \mid z)$ is continuous and vanishes at infinity. Then the set $U \cup \{0\}$ and consequently the sets V and W are compact in $\mathbb{R}^m$.*

Proof. Under the stated conditions the one-point compactification $\mathscr{Z} \cup \{\infty\}$ is a metrizable compact space. Thus both the set of all probability measures on $\mathscr{Z}$ and the set of all one-point probability measures on $\mathscr{Z}$ are compact for the weak topology. Set $p(x \mid \infty) = 0$ for each x. Then $z \to p(x \mid z)$ is continuous on the one-point compactification. Hence so is the map

$$\eta \to (p(x_1 \mid \eta), ..., p(x_m \mid \eta))$$

from the Borel measures on Z $\cup \{\infty\}$ to $\mathbb{R}^m$. The set $U \cup \{0\}$ is the image of all one-point probability measures under this map. The set W is the image of all probability measures. The set V is the image of $(U \cup \{0\}) \times [0, 1]$ under the map $(u, \alpha) \to \alpha u$, so is compact. ∎

3. Consistency

Throughout this section let $\mathscr{Z}$ be a locally compact, separable metric space and let $\mathscr{C}$ be its Borel σ-field. The set $\mathscr{H}$ of all Borel subprobability measures on $\mathscr{Z}$ can be equipped with the *vague topology*. This can be determined by defining that $\eta_n \Rightarrow \eta$, or η_n converges vaguely to η, if and only if

$$\int f \, d\eta_n \to \int f \, d\eta, \qquad \text{every} \quad f \in C_0(\mathscr{Z}).$$

(The set $\mathscr{C}_0(\mathscr{Z})$ of continuous functions that vanish at infinity was defined at the end of Section 2.) It is well known that under the stated on $\mathscr{Z}$ the vague topology is metrizable; the set of all subprobability measures is vaguely compact; and for probability measures η_n and η vague convergence $\eta_n \Rightarrow \eta$ is equivalent to the more usual weak convergence. (See, e.g., Bauer [1].)

Assume that the kernel $p(x \mid z)$ satisfies the weak smoothness condition

$$\lim_{\eta_n \Rightarrow \eta} p(x \mid \eta_n) = p(x \mid \eta), \qquad \text{for } \mu\text{-almost all } x. \tag{3.1}$$

(The exceptional set of x may depend on η, but not on the sequence η_n.) Moreover, assume that

$$\text{the map } x \to \sup_{\gamma \in U} p(x \mid \gamma) \text{ is measurable} \tag{3.2}$$

for every sufficiently small open ball $U \subset \mathscr{H}$. These conditions are certainly satisfied if $z \to p(x \mid z)$ is in $C_0(\mathscr{Z})$ for every x. Since convergence need only hold almost everywhere, the conditions actually cover a much larger class of examples.

Secondly, assume that η is identifiable in the pure mixture model in the sense that

$$\mu(x\colon p(x \mid \eta') \neq p(x \mid \eta)) > 0 \qquad \text{for every} \quad \eta' \neq \eta. \tag{3.3}$$

These conditions suffice for consistency.

Theorem 3.1. *Let $\mathscr{Z}$ be a locally compact, separable metric space and let the kernel $p(x \mid z)$ satisfy* (3.1) *for every $\eta \in \mathscr{H}$,* (3.2), *and* (3.3) *for the true η. Then any sequence of maximum likelihood estimators $\tilde{\eta}_{m,n}$ satisfies $\tilde{\eta}_{m,n} \Rightarrow \eta$ almost surely under η if $m \to \infty$.*

Proof. *Case* 1. Both $m \to \infty$ and $n \to \infty$. Let $\hat{\eta}$ be the empirical distribution of $Z_1, ..., Z_n$. It is well known that $\hat{\eta} \Rightarrow \eta$ for almost every realization $z_1, z_2, ...$. (Varadarajan's theorem, see Dudley [4, p. 313].) Fix such a sequence throughout the remainder of this part of the proof. It will be shown that $\tilde{\eta} = \tilde{\eta}_{m,n}(X_1, ..., X_m, z_1, ..., z_n) \Rightarrow \eta$ conditionally on the sequence $z_1, z_2, ...$.

Since $\hat{\eta}$ is the maximum likelihood estimator for η based on $z_1, ..., z_n$ alone,

$$\prod_{j=1}^{n} \tilde{\eta}\{z_j\} \leqslant \prod_{j=1}^{n} \hat{\eta}\{z_j\}.$$

From this and concavity of the function $u \to \log u$ we obtain that

$$\sum_{j=1}^{n} \log\left[1 + \alpha\left(\frac{\hat{\eta}\{z_j\}}{\tilde{\eta}\{z_j\}} - 1\right)\right]\Bigg) \geqslant 0, \tag{3.4}$$

for every $\alpha \in (0, 1)$. Fix such an α throughout the remainder of the proof.

The true parameter η is also identifiable in the sense that $P_\eta(p(X | \gamma) \neq p(X | \eta)) > 0$ for every subprobability $\partial \neq \eta$. Indeed, if this probability were zero, then it would follow that $\mu(x\colon p(x | \gamma) \neq p(x | \eta), p(x | \eta) > 0) = 0$. So for μ-almost x with $p(x | \eta) > 0$ the densities $p(x | \eta)$ and $p(x | \gamma)$ are equal. Since the total mass of the second density is not larger than 1, the total mass of the first density, it must be that $p(x | \gamma) = 0$ at almost all x where $p(x | \eta) = 0$. Combination yields that $\mu(x : p(x | \gamma) \neq p(x | \eta)) = 0$, in contradiction to the identifiability condition (3.3).

Fix a subprobability measure $\gamma \neq \eta$. By convexity of the function $u \to u \log(1 + \alpha(u - 1))$, identifiability of η and Jensen's inequality

$$E_\eta \log\left[1 + \alpha\left(\frac{p(X | \eta)}{p(X | \gamma)} - 1\right)\right] > 0. \tag{3.5}$$

(Let the quotient of a positive number and zero be infinity.) By (3.1) one has $p(x | \tilde{\eta}) \to p(x | \eta)$ for almost all x. For an open ball U around γ define the expression $\bar{p}(x | U)$ as $\sup_{\gamma' \in U} p(x | \gamma')$. Then for a sequence of open balls U_n decreasing to γ we have $\bar{p}(x | U_n) \to p(x | \gamma)$ by (3.1). Thus by Fatou's lemma and (3.5) we have for every such $U_n \downarrow \gamma$ and $M_n \uparrow \infty$, no matter how slowly,

$$\liminf_{n \to \infty} \int_{\{x: p(x | \eta) > 0\}} \left\{\log\left[1 + \alpha\left(\frac{p(x | \tilde{\eta})}{\bar{p}(x | U_n)} - 1\right)\right] \wedge M_n\right\}$$
$$\times\, p(x | \eta)\, d\mu(x) > 0.$$

(Note that $\log(1+\alpha(u-1))$ is bounded below by $\log(1-\alpha)$ if $u \geqslant 0$.) This implies that there is an open neighbourhood U_γ of γ and a constant M_γ with

$$\liminf_{n\to\infty} E_\eta \log\left[1+\alpha\left(\frac{p(X\mid\hat\eta)}{\bar p(X\mid U_\gamma)}-1\right)\right] \wedge M_\gamma > 0, \tag{3.6}$$

where the expectation is to be understood as conditional on the fixed sequence $z_1, z_2, \ldots$. This has been obtained for an arbitrary $\gamma \neq \eta$.

The likelihood function is at $\tilde\eta$ not smaller than at $\alpha\hat\eta + (1-\alpha)\tilde\eta$. Using the linearity of the map $\eta \to p(x\mid\eta)$ this can be expressed as

$$\sum_{i=1}^{m} \log\left[1+\alpha\left(\frac{p(X_i\mid\hat\eta)}{p(X_i\mid\tilde\eta)}-1\right)\right] + \sum_{j=1}^{n} \log\left[1+\alpha\left(\frac{\hat\eta\{z_j\}}{\tilde\eta\{z_j\}}-1\right)\right] \leqslant 0. \tag{3.7}$$

Combination with (3.4) yields that

$$\sum_{i=1}^{m} \log\left[1+\alpha\left(\frac{p(X_i\mid\hat\eta)}{p(X_i\mid\tilde\eta)}-1\right)\right] \leqslant 0 \tag{3.8}$$

Fix a vaguely open neighbourhood U of the true η. The complement of U in the set of subprobability measures is a vaguely closed subset of a compact set, so vaguely compact. The open cover $\{U_\gamma : \gamma \notin U\}$ of this complement has a finite subcover $U_{\gamma_1}, \ldots, U_{\gamma_p}$. If $\tilde\eta$ is not in U, then it is in one of the U_{γ_k}, in which case $\bar p(x\mid U_{\gamma_k}) \geqslant p(x\mid\tilde\eta)$ for every x. Thus by (3.8)

$$\{\tilde\eta \notin U\} \subset \bigcup_{k=1}^{p} \left\{ m^{-1} \sum_{i=1}^{m} \log\left[1+\alpha\left(\frac{p(X_i\mid\hat\eta)}{\bar p(X_i\mid U_{\gamma_k})}-1\right)\right] \wedge M_{\gamma_k} \leqslant 0 \right\}.$$

The conditional probability (given $z_1, z_2, \ldots$) of each of the sets in the union is the probability that an average of m uniformly bounded random variables is nonnegative. For a fixed, sufficiently large n these variables have a positive expectation under η by (3.6). By Hoeffding's inequality each of the probabilities is of order $e^{-\varepsilon m}$ for some $\varepsilon > 0$. More precisely, ε can be chosen equal to

$$\frac{2\mu^2}{(M_\gamma - \log(1-\alpha))^2},$$

and the upper bound $e^{-\varepsilon m}$ holds for every n such that the expectation in (3.6) is larger than μ, say $n \geqslant N$. Consequently,

$$\sum_{m=1}^{\infty} \sup_{n \geqslant N} P(\tilde\eta_{m,n} \notin U) < \infty.$$

By a minor modification of the Borel–Cantelli lemma it follows that $\tilde{\eta}_{m,n} \in U$, eventually, almost surely.

Case 2. n fixed and $m \to \infty$. The likelihood function is at η not smaller than at $\alpha\eta + (1-\alpha)\tilde{\eta}$. Rewrite this as in (3.7), but with η substituted for $\hat{\eta}$. The second term on the left is bounded below by $n \log(1-\alpha)$. Hence we obtain

$$\sum_{i=1}^{m} \log\left[1+\alpha\left(\frac{p(X_i \mid \eta)}{p(X_i \mid \tilde{\eta})}-1\right)\right] \leqslant -n \log(1-\alpha).$$

Though the right side of this inequality is now positive, the proof can be finished as before, where (3.5) is now used instead of (3.6). ▌

Note. The proof of the theorem shows that condition (3.1) may be relaxed to the two conditions:

— $\eta \to p(x \mid \eta)$ is vaguely continuous at the true η for almost all x;

— $\eta \to p(x \mid \eta)$ is vaguely upper semi-continuous at every η for almost all x, where the set of exceptional set of x may depend on η.

Note. The measurability condition (3.2) requires some "separability" of the stochastic process $\eta \to p(x \mid \eta)$; it is satisfied if there is a countable set $\mathscr{H}'$ of subprobability measures such that for every x the supremum of $p(x \mid \gamma)$ over $\gamma \in U$ is the same as the supremum over all $\gamma \in U \cap \mathscr{H}'$. One sufficient condition for this is lower semi-continuity of the map $\gamma \to p(x \mid \gamma)$ for every x and at every γ. (Then any countable dense $\mathscr{H}'$ qualifies.) This is in turn true if the map $z \to p(x \mid z)$ is lower semi-continuous and vanishes at infinity for every x. Other situations wherein the process $\eta \to p(x \mid \eta)$ is separable occur when the map $z \to p(x \mid z)$ is left- or right-continuous for every x. We exploit this in the examples rather than write up a general lemma.

To compute a maximum likelihood estimate the hardest problem is to find the location of the support points of $\tilde{\eta}$. One way to avoid this problem is to fix a grid of support points from the beginning and maximize the likelihood over all distributions with support in this grid. If the number of grid points is chosen larger and larger as the number of observations increases, this procedure is known as the "method of sieves."

In the present problem the likelihood at η is positive only if η gives positive mass to every of the points $z_1, ..., z_n$. Therefore assume that our sieves $\mathscr{H}_{m,n}$ are stochastic subsets of $\mathscr{H}$, possibly depending on $z_1, ..., z_n$, but not on $X_1, ..., X_m$, that contain $\hat{\eta}_n$ for every m, n. This already suffices to render the sequence of maximum likelihood estimators over $\mathscr{H}_{m,n}$ consistent.

THEOREM 3.2. *Let $\mathscr{X}$ be a locally compact, separable, metric space and let the kernel $p(x \mid z)$ satisfy* (3.1) *for every $\eta \in \mathscr{H}$,* (3.2), *and* (3.3) *for the true η. Let $\mathscr{H}_{m,n}$ be subsets of $\mathscr{H}$ that depend on $z_1, ..., z_n$, m, and n only and contain $\hat{\eta}_n$. Let $\tilde{\eta}$ satisfy*

$$\prod_{i=1}^{m} p(X_i \mid \tilde{\eta}) \prod_{j=1}^{n} \tilde{\eta}\{z_j\} \geqslant c \sup_{\eta \in \mathscr{H}_{m,n}} \prod_{i=1}^{m} p(X_i \mid \eta) \prod_{j=1}^{n} \eta\{z_j\}$$

for some $c > 0$. Then $\tilde{\eta} \Rightarrow \eta$ almost surely if both m, $n \to \infty$.

Proof. This is almost identical to the proof of Case 1 in the previous theorem. ▌

4. EXAMPLES

EXAMPLE 1 (Shifted Uniform). Let $p(x \mid z) = 1_{(z, z+1)}(x)$ be the density of the uniform distribution on $(z, z+1)$ and let $\mathscr{X}$ equal $\mathbb{R}$ or $(0, \infty)$. Then U and V are compact, the maximum likelihood estimator exists, has finite discrete support, and is consistent.

First note that every point in the set U defined in Section 2 is a vector of zeros and ones. Hence U is a finite set and certainly compact. Next the map $\eta \to p(x \mid \eta) = \eta(x-1, x)$ is vaguely continuous at every η that does not charge the points x and $x-1$. Hence (3.1) is satisfied for all η. The map $\eta \to \eta(x-1, x)$ is also lower semi-continuous for every x. This implies that

$$\sup_{\gamma \in U} p(x \mid \gamma) = \sup_{\gamma \in U \cap \mathscr{H}'} p(x \mid \gamma)$$

for every vaguely dense subset of $\mathscr{H}'$. Take a countable dense $\mathscr{H}'$ to verify (3.2). Finally, suppose $p(x \mid \eta') = p(x \mid \eta)$ for Lebesgue almost all x. In terms of the cumulative distribution function this equality becomes $\eta(x-) - \eta(x-1) = \eta'(x-) - \eta'(x-1)$. Approach an arbitrary y from above by a suitable sequence x_n to find that $\eta(y-1, y] = \eta'(y-1, y]$ for every y, whence $\eta(-\infty, y] = \eta'(-\infty, y]$ for every y, and $\eta = \eta'$.

EXAMPLE 2 (Shifted Exponential). Let $\mathscr{X} = [0, \infty)$ and let the kernel be the shifted exponential density $p(x \mid z) = e^{-(x-z)} 1_{x \geqslant z}$. Then the set V is compact and the conditions of the consistency theorem satisfied.

To see the first, define for $1 \leqslant i \leqslant m$

$$U_i = \{(p(x_1 \mid z), ..., p(x_m \mid z)) \colon x_{(i-1)} < z \leqslant x_{(i)}\}.$$

Here for $i = 0$ read $x_{(0)} = 0-$ so that z ranges over $[0, x_{(1)}]$. Then $U = \bigcup_{i=1}^{m} U_i \cup \{0\}$ and to show that V is compact it suffices to show that each $V_i = \{\alpha u: 0 \leqslant a \leqslant 1, u \in U_i\}$ is compact. Each element of U_i has $(i-1)$ coordinates equal to zero. Assume for simplicity that $x_1 < x_2 < \cdots < x_m$. Then

$$U_i = \{e^z(0, ..., 0, e^{-x_i}, ..., e^{-x_m}): x_{i-1} < z \leqslant x_i\}$$

and

$$V_i = \{\alpha(0, ..., 0, e^{-x_i}, ..., e^{-x_m}): 0 \leqslant \alpha \leqslant e^{x_i}\}.$$

This completes the proof of compactness of V.

Since $z \to p(x \mid z)$ has only one discontinuity point for every fixed x and vanishes at infinity, one has $p(x \mid \eta_n) \to p(x \mid \eta)$ as $\eta_n \Rightarrow \eta$ for every x where η does not have a jump; hence, for fixed η certainly for Lebesgue almost all x. This verifies (3.1). If $p(x \mid \eta') = p(x \mid \eta)$ for Lebesgue almost all x, then

$$\int_{[0,x]} e^z \, d\eta'(z) = \int_{[0,x]} e^z \, d\eta(z)$$

for almost all x. By right continuity of these functions equality must hold for every x. Hence $\eta' = \eta$. This verifies (3.3).

Finally for verification of (3.2) let $\mathscr{H}'$ be the set of all discrete probability measures with finite support and point masses in the rationals. Since $z \to p(x \mid z)$ is right continuous and bounded for every x, it follows that for every $\eta \in \mathscr{H}$ and x there is a sequence η_n in $\mathscr{H}'$ with $\eta_n \Rightarrow \eta$ and $p(x \mid \eta_n) \to p(x \mid \eta)$. (Discretize η on a grid $z_1 < z_2 < \cdots < z_n$ putting the mass of η in $(z_{i-1}, z_i]$ at z_i for every i.) This implies that

$$\sup_{\gamma \in U} p(x \mid \gamma) = \sup_{\gamma \in U \cap \mathscr{H}'} p(x \mid \gamma)$$

for every open set U. Since $\mathscr{H}'$ is countable this function is measurable.

EXAMPLE 3 (Uniform Scale). Let $\mathscr{Z} = (0, \infty)$ and $p(x \mid z) = (1/z) 1_{(0,z]}(x)$. This example is treated in detail in Vardi [14] and Vardi and Zhang [15]; the latter paper also derives the asymptotic distribution of the maximum likelihood estimator. Here we show briefly that this example also falls under the present set-up.

Again V is compact and the conditions of the consistency theorem are satisfied. To see the first, set $x_{(m+1)} = \infty$ and for $1 \leqslant i \leqslant m$

$$U_i = \{(p(x_1 \mid z), ..., p(x_m \mid z)): x_{(i)} \leqslant z < x_{(i+1)}\}.$$

Then $U = \{0\} \cup \bigcup_{i=1}^m U_i$. Every element of U_i has i nonzero coordinates. Assume for simplicity that $x_1 < x_2 < \cdots < x_m$. Then

$$U_i = \{(1/z)(1, ..., 1, 1, 0, 0, ..., 0)): x_{(i-1)} < z \leqslant x_{(i)}\},$$

where the first zero occurs at the $(i+1)$th spot. The set $V_i=\{\alpha u: 0\leqslant\alpha\leqslant 1, u\in U_i\}$ satisfies

$$V_i=\{\alpha(1, ..., 1, 1, 0, 0, ..., 0): 0\leqslant\alpha\leqslant(1/x_i)\}.$$

Hence each V_i, and consequently V, is compact.

Validity of (3.1) follows from the fact that $z\rightarrow p(x\mid z)$ has only one discontinuity point and vanishes at infinity, as in the shifted exponential example. By the same method as in that example, (3.2) follows from right continuity of $z\rightarrow p(x\mid z)$ and identifiability (3.3) follows from right continuity of $x\rightarrow p(x\mid\eta)$.

A closer look reveals that the support points of the maximum likelihood estimator can be taken equal to the totality of observed values $x_1, ..., x_m, z_1, ..., z_n$.

EXAMPLE 4 (Exponential Family). Let $p(x\mid z)$ be the density of a one-dimensional exponential family of the form

$$p(x\mid z)=c(z)\,h(x)\,e^{z\tau(x)}.$$

The function h can always be absorbed into the dominating measure μ, so it is not restrictive to assume that it is strictly positive. Let $\mathscr{Z}=\{z: \int h(x)\,e^{z\tau(x)}\,d\mu(x)<\infty\}$ be the natural parameter space of the family. This is an interval that may or may not be closed at its endpoints.

In many examples the function $z\rightarrow p(x\mid z)$ is contained in $C_0(\mathscr{Z})$ for every x, but this is not necessarily the case if $\mathscr{Z}$ is unbounded.

LEMMA 4.1. *Let h be strictly positive. The function $z\rightarrow p(x\mid z)$ is contained in $C_0(\mathscr{Z})$ for every x if and only if both of the following statements are true*:

— *$\mathscr{Z}$ is bounded below or $\mu\{t: \tau(t)<\tau(x)\}>0$ for every x*;

— *$\mathscr{Z}$ is bounded above or $\mu\{t: \tau(t)>\tau(x)\}>0$ for every x.*

Proof. It is well known that the function $z\rightarrow c(z)^{-1}=\int h(x)\,e^{z\tau(x)}\,d\mu(x)$ is continuous where it is defined and finite, hence so are $z\rightarrow c(z)$ and $z\rightarrow p(x\mid z)$. The latter vanishes at infinity if and only if it converges to zero as z converges to an open boundarty of $\mathscr{Z}$. If $\mathscr{Z}$ is bounded from above and is open on the right, then it must be that $\int h(x)\,e^{z\tau(x)}\,d\mu(x)\rightarrow\infty$ as z increases to the boundary of $\mathscr{Z}$. Consequently, $c(z)\rightarrow 0$ and $p(x\mid z)\rightarrow 0$. If $\mathscr{Z}$ is unbounded from above, then

$$p(x\mid z)=h(x)\left(\int e^{z(\tau(t)-\tau(x))}\,h(t)\,d\mu(t)\right)^{-1}$$

converges to zero as $z \to \infty$ if and only if $\mu\{t: \tau(t) > \tau(x)\} > 0$. This takes care of the right boundary of $\mathscr{X}$. The argument for the left boundary is analogous. ∎

In the case that $\tau(x) = x$ the condition for $z \to p(x \mid z)$ to be in $C_0(\mathscr{X})$ may be summarized as: on both ends either $\mathscr{X}$ is bounded or the support of μ is unbounded.

If the conclusion of the lemma holds, then the set V is compact, so the maximum likelihood estimator exists and may be taken to have finite support. Furthermore, of the conditions for consistency only (3.3) remains to be checked.

LEMMA 4.2. *Suppose that for every μ null set A the set $\{\tau(x): x \notin A\}$ contains a converging sequence with limit not equal to* $\inf \tau(x)$ *or* $\sup \tau(x)$ *where x runs through the set of all such that $p(x \mid \eta)$ is finite. Then η is identifiable in the sense of* (3.3). *In particular, if $\tau(x) = x$ then η is identifiable if μ is equivalent to Lebesgue measure on an open interval or is discrete with a limit point in the interior of its support.*

Proof. The equality $p(x \mid \eta') = p(x \mid \eta)$ leads immediately to

$$\int e^{z\tau(x)}\, d\eta'(z) = \int e^{z\tau(x)}\, d\eta(z).$$

If $p(x \mid \eta') = p(x \mid \eta)$ almost everywhere, then η' and η have the same Laplace transform at almost every $\tau(x)$. If the Laplace transforms are equal on a converging sequence with limit in the interior of the interval where they are finite, then the two measures are equal. This interval includes all values $\tau(x)$ for which $p(x \mid \eta)$ is finite, whence the result. ∎

REFERENCES

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

[2] BHANJA, J., AND GHOSH, J. K. (1988). *Efficient Estimation with Many Nuisance Parameters.* Technical Report, Indian Statist. Institute.

[3] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y., AND WELLNER, J. A. (1992). *Efficient and Adaptive Inference in Semiparametric Models.* Johns Hopkins, Press, Baltimore.

[4] DUDLEY, R. M. (1989). *Real Analysis and Probability.* Wadsworth & Brooks/Cole, Pacific Grove, California.

[5] GROENEBOOM, P. (1988). Untitled preprint.

[6] HASMINSKII, R. Z., AND IBRAGIMOV, I. A. (1983). On asymptotic efficiency in the presence of an infinite dimensional nuisance parameter. (ed.: Ito and Prohorov). Lecture Notes in Mathematics 1021, Springer, New York. 195–229.

[7] HECKMAN J., AND SINGER, B. (1984). A method for minimizing the impact of distributional assumptions in economic studies for duration data. *Econometrica* **52** 271–320.

[8] JEWELL, N. P. (1982). Mixtures of exponential distributions. *Ann. Statist.* **10** 479–484.

[9] 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.

[10] LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.* **73** 805–811.

[11] LINDSAY, B. G. (1983). The geometry of mixture likelihoods, I and II. *Ann. Statist.* **11** 86–94 and 783–792.

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

[13] VAN DER VAART, A. W. (1988). *Estimating a Parameter in Incidental and Structural Models by Approximate Maximum Likelihood.* Technical Report, No. 139, Department of Statistics, University of Washington.

[14] VARDI, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika* **76** 751–761.

[15] VARDI, Y., AND ZHANG, C.-H. (1989). Large sample study of empirical distributions in a random-multiplicative censoring model. Preprint.