The Annals of Statistics
1994, Vol. 22, No. 4, 1896–1916

MAXIMUM LIKELIHOOD ESTIMATION WITH PARTIALLY CENSORED DATA[1]

BY AAD VAN DER VAART

Vrije Universiteit

Suppose one observes independent samples of size n from both the mixture density $\int p(x \mid z)\, d\eta(z)$ and from the distribution η. The kernel $p(x \mid z)$ is known. We show asymptotic normality and efficiency of the maximum likelihood estimator for η.

1. Introduction. If one observes a sample of independent, identically distributed random elements $Z_1, \ldots, Z_n$ from a completely unknown probability distribution η, then the usual estimator for η is the empirical distribution $\widehat{\eta} = n^{-1}\sum_{j=1}^{n} \delta_{Z_j}$. Consider the situation wherein the observed $Z_1, \ldots, Z_n$ 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 $X_1, \ldots, X_m, Z_1, \ldots, Z_n$; all observations are independent and their joint distribution can formally be written as

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

(The first factor in the product is a density with respect to μ^n; the second factor is just formal notation.)

In this situation the set $Z_1, \ldots, Z_n$ clearly contains much more information about η than the set $X_1, \ldots, X_m$. Nevertheless, one would certainly want to take the information available in $X_1, \ldots, X_m$ into account and obtain an improved estimator for η relative to using $\widehat{\eta}$, the empirical distribution of the second sample. Surprisingly enough there may be a considerable gain in using $X_1, \ldots, X_m$ even in situations where the information (in the technical sense of semiparametric theory) in $X_1, \ldots, X_m$ alone is 0 and $\sqrt{n}$-consistent estimators based on the first sample alone do not exist. For $m = n$ use of the additional sample always becomes visible as a cut in the asymptotic variance of the estimator.

It is thus of interest to study estimators for η based on the whole set of observations. In this paper we show under some smoothness conditions that the maximum likelihood estimator for η attains a $\sqrt{n}$-rate and is asymptotically

Received March 1992; revised February 1994.
[1]Research supported by NSF Grant DMS-85-05550.
AMS 1991 *subject classifications*. 62G20, 62F12.
Keywords and phrases. Mixture model, maximum likelihood, efficiency, deconvolution, censoring.

efficient for estimating η in the semiparametric sense. For a definition of this efficiency concept, see, for example the monograph of Bickel, Klaassen, Ritov and Wellner (1993). For general notes on nonparametric maximum likelihood estimation, see Gill (1989).

Actually we obtain this result for a slightly simpler version of the model. While our methods apply with inessential changes to the case that m and n are of comparable magnitude, it is assumed for simplicity of notation that $m = n$. Then the observations can be paired and the total set of observations can be rearranged as an i.i.d. sample $(X_1, Z_1), \ldots, (X_n, Z_n)$ from the distribution given formally as

$$\int p(x \mid y)\, d\eta(y)\, d\eta(z).$$

For definiteness let $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Z}, \mathcal{C})$ 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. In most of our examples the space $\mathcal{Z}$ is an interval in the real line or the plane. The density of X_i is (with abuse of notation) written as $p(x \mid \eta) = \int p(x \mid z)\, d\eta(z)$.

The model as defined here has been studied by Hasminskii and Ibragimov (1983), Bhanja and Ghosh (1988), Vardi (1989), Vardi and Zhang (1992) and Bickel, Klaassen, Ritov and Wellner (1993). Vardi and Zhang obtain asymptotic normality of the maximum likelihood estimator for one particular kernel $p(x \mid z)$, while Bickel, Klaassen, Ritov and Wellner (1993) and Bickel and Ritov (1993) discuss the general model, but not the maximum likelihood estimator. In the literature the type of distribution of each X_i is called a *mixture model* and sometimes a *structural model*. Estimation of η by maximum likelihood based on $X_1, \ldots, X_n$ alone has been considered among others by Kiefer and Wolfowitz (1956), Laird (1978), Jewell (1982), Lindsay (1983), Heckman and Singer (1984), van der Vaart (1988), Pfanzagl (1988) and Groeneboom (1991). This problem is completely different from the present one.

For the present problem van der Vaart and Wellner (1993) obtained the existence and consistency of the maximum likelihood estimator for the weak topology under a smoothness condition on the map $\eta \to p(x \mid \eta)$. Moreover, by the argument of Lindsay (1983) there is always a maximum likelihood estimator that is a discrete distribution with at most $2n$ support points. Although the condition for consistency in van der Vaart and Wellner (1993) is simple and weak, it nevertheless fails in a number of cases of interest. In such cases the maximum likelihood estimator may be undefined and/or inconsistent. We refer to the paper by Bickel and Ritov (1993). We note that roughly the maximum likelihood estimator is consistent in the present model whenever it is consistent in the mixture model. Thus having the "good" observations $Z_1, \ldots, Z_n$ in addition to the "bad" observations $X_1, \ldots, X_n$ causes no trouble for the maximum likelihood estimator if it was not in trouble already. Conversely, the fact that the maximum likelihood estimator based on the "good" observations (the empirical distribution of the Z's) behaves very well, does not guard against bad

behavior of the maximum likelihood estimator in the model with both "good" and "bad" observations.

The organization of the paper is as follows. In Section 2 it is shown that the maximum likelihood estimator is an M-estimator in the sense that it satisfies an (infinite) set of equations. Furthermore, a theorem on asymptotic normality of infinite dimensional M-estimators is formulated in general and next specified to the present problem. Section 3 contains the main results of the paper. It is explained how the general results of Section 2 can be applied and several results give conditions that are relatively easy to check. One condition that appears harder to verify by elementary methods is discussed separately in Section 4. In Section 5 and 6 we discuss concrete examples. Some proofs have been omitted, but can be found in a technical report.

Our main method to prove asymptotic efficiency of the maximum likelihood estimator for functionals of the type $\int h\,d\eta$ (for a fixed h such as an indicator) is to write up and invert a convenient collection of likelihood equations, one of which is indexed by the functional of interest and the others of which are chosen for technical reasons. This approach is explained in Section 3.

2. The likelihood equations. For a measure η write $\eta\{z\}$ for the mass that η gives to the one-point set $\{z\}$. Given fixed (observed) values $(x_1, z_1), \ldots, (x_n, z_n)$ a maximum likelihood estimator $\widetilde{\eta}_n$ is any probability distribution on $\mathcal{Z}$ that maximizes the "likelihood function." For our purpose this is the function

$$\eta \to \operatorname{lik}(\eta) = \prod_{i=1}^{n} p(x_i \mid \eta) \prod_{j=1}^{n} \eta\{z_j\}.$$

Van der Vaart and Wellner (1993) show under some weak conditions that $\widetilde{\eta}_n$ exists (in other words, the supremum is achieved), is consistent for the weak topology and can be taken finitely discrete with no more than $2n$ support points. (Since the maximum likelihood estimator may be nonunique, the last does not mean that the maximum likelihood estimator is necessarily discrete.)

Our proof of asymptotic normality proceeds by showing that any maximum likelihood estimator solves a collection of likelihood equations. Next the totality of equations is treated as a map of the parameter space into a function space, to which a general theorem of infinite-dimensional M-estimation can be applied. The latter is a straightforward extension to infinite dimensions of results due to Huber (1967) and Pakes and Pollard (1989) and results of this type can also be found in Bickel, Klaassen, Ritov and Wellner (1993).

To set up a set of estimating equations consider a class $\mathcal{H}'$ of bounded functions $h\colon \mathcal{Z} \to \mathbb{R}$. (The prime notation is used in this section to separate the present class of functions from a more fundamental class of functions introduced later.) Then for each fixed probability measure $\widetilde{\eta}$, each h and every sufficiently small real number $|t|$, we can define a probability measure $\widetilde{\eta}_t$ by

$$d\widetilde{\eta}_t = \bigl(1 + t(h - E_{\widetilde{\eta}} h)\bigr)\, d\widetilde{\eta}.$$

For $\tilde{\eta}$ equal to the maximum likelihood estimator evaluated at fixed data points, the map $t \to \text{lik}(\tilde{\eta}_t)$ is maximized over a neighborhood of 0 at $t = 0$. Take the derivative with respect to t at $t = 0$ to obtain a likelihood equation. Varying h over the class $\mathcal{H}'$ yields a large class of equations.

In the present case the equations take the following form. Define operators A_η and l_η by

$$\begin{aligned} A_\eta h(x,z) &= l_\eta h(x) + h(z) \\ &= \frac{\int h(y)p(x\,|\,y)\,d\eta(y)}{p(x\,|\,\eta)} + h(z). \end{aligned}$$

Then the maximum likelihood equation for the one-dimensional submodel $t \to \eta_t$ is $W_n(\eta)h = 0$ for W_n given by

$$W_n(\eta)h = \hat{P}_n A_\eta h - P_\eta A_\eta h.$$

Here $\hat{P}_n$ is the empirical distribution of the observations, P_η is the distribution of (X,Z) and we write Pf for $\int f\,dP$. It can be checked directly from the formulas that any maximum likelihood estimator $\tilde{\eta}_n$ indeed satisfies $W_n(\tilde{\eta}_n)h = 0$. Alternatively, it is helpful to note that A_η is the "score operator" in a missing data problem. View each observed data value (X,Z) as arising from a triple (X,Y,Z) in which (X,Y) is independent from Z and X given Y has conditional density $p(x\,|\,y)$. Score functions for the unobserved data (X,Y,Z) take the form

$$h(y) + h(z) = \frac{\partial}{\partial t}\bigg|_{t=0} \log p(x\,|\,y)\,d\eta_t(y)\,d\eta_t(z).$$

The operator A_η turns these into score functions for the model of the observed data through

$$A_\eta h(X,Z) = E_\eta\big(h(Y) + h(Z)\,|\,X,Z\big).$$

Again these formulas can be checked directly for the present case. Alternatively, connections of this type can be deduced very generally when using an L_2-type definition of scores (as derivatives of root densities) and are intimately connected with asymptotic efficiency theory. From the L_2-theory we borrow the notation for the "adjoint" operator

$$l^* g(x) = \int g(x)p(x\,|\,z)\,d\mu(x),$$

which maps functions of x into functions of z. According to semiparametric efficiency theory a best estimator sequence T_n for the "parameter" $\eta \to \int h\,d\eta$ satisfies that $\sqrt{n}(T_n - \int h\,d\eta)$ is asymptotically normally distributed with zero mean and variance

$$\sigma_\eta^2(h) = \int \left[A_\eta(I + l^* l_\eta)^{-1}(h - E_n h)\right]^2 dP_\eta.$$

We do not discuss the efficiency theory in this paper, but show that $T_n = \int h\, d\widetilde{\eta}_n$ is asymptotically mean zero normal with the given variance.

Define W as the expected version of W_n under the true parameter η_0. Thus

$$W(\eta)h = P_0 A_\eta h - P_\eta A_\eta h,$$

where here and in the sequel the subscript 0 is short for η_0. Since A_η is a conditional expectation operator, the map $h \to A_\eta h(x,y)$ is bounded if h ranges over a uniformly bounded set of functions $\mathcal{H}'$. Thus both $\eta \to W_n(\eta)$ and $\eta \to W(\eta)$ can be viewed as maps into the space $l^\infty(\mathcal{H}')$ of all uniformly bounded functions $z\colon \mathcal{H}' \to \mathbb{R}$. The domain of these two maps is the set P of all probability distributions η on $\mathcal{Z}$. These will be identified with maps

$$\eta\colon h \to \int h\, d\eta,$$

evaluating the expectations of a class $\mathcal{H}$ of functions $h\colon \mathcal{Z} \to \mathbb{R}$. This second class $\mathcal{H}$ need not have any relationship to the class $\mathcal{H}'$ introduced earlier. Basically, the class $\mathcal{H}'$ is chosen for convenience, to make the proofs work. On the other hand, the choice of the class $\mathcal{H}$ determines the nature of our limiting results, We shall obtain asymptotic normality of $\sqrt{n}(\int d\widetilde{\eta}_n - \int h\, d\eta)$ for every $h \in \mathcal{H}$ (uniformly in $\mathcal{H}$).

It is assumed that the functions in $\mathcal{H}$ are also uniformly bounded. In that case W_n and W can be viewed as maps

$$W, W_n\colon \mathcal{P} \subset l^\infty(\mathcal{H}) \to l^\infty(\mathcal{H}').$$

We make specific choices for $\mathcal{H}$ and $\mathcal{H}'$ later on. The two l^∞-spaces are both equipped with the uniform norm. In the case of the first this is the norm $\|z\|_{\mathcal{H}} = \sup\{|z(h)|\colon h \in \mathcal{H}\}$. The proof of asymptotic normality of $\sqrt{n}(\widetilde{\eta}_n - \eta)$ is based on the following general result.

Let W_n and W be random and fixed maps as in the last displayed equation that satisfy the following three conditions. For every sequences $\varepsilon_n \downarrow 0$ and $\eta_n \to \eta_0$ within the domain of W_n and W,

$$\sup_{\|\eta - \eta_0\| < \varepsilon_n} \|\sqrt{n}(W_n - W)(\eta) - \sqrt{n}(W_n - W)(\eta_0)\| \to_P 0, \tag{2.1}$$

$$\sqrt{n}(W_n - W)(\eta_0) \rightsquigarrow \mathbb{G}, \tag{2.2}$$

$$W(\eta_n) - W(\eta_0) = \dot{W}_0(\eta_n - \eta_0) + o\big(\|\eta_n - \eta_0\|\big). \tag{2.3}$$

Here in the third condition $\dot{W}_0$ is required to be a linear map with domain the linear span of the domain P of W_n and W. The convergence in the second condition is weak convergence in $l^\infty(\mathcal{H}')$, where we use the definition of "convergence in law without defining laws" due to Hoffmann-Jørgensen (1984) to avoid measurability problems. [See Dudley (1985) for a partial review.] Similarly, in the first condition the convergence is understood to be in outer probability if the variables are not measurable.

THEOREM 2.1. *Let W_n, W: $\mathcal{P} \subset l^\infty(\mathcal{H}) \to l^\infty(\mathcal{H}')$ be maps such that* (2.1)–(2.3) *hold for a linear and one-to-one map $\dot{W}_0$: $\operatorname{lin} \mathcal{P} \to l^\infty(\mathcal{H}')$ whose inverse is continuous on the range of $\dot{W}_0$. Let $\widetilde{\eta}_n$ be random maps into the domain P of W_n and W with $W_n(\widetilde{\eta}_n) - W(\eta_0) = o_P^*(n^{-1/2})$. If $\widetilde{\eta}_n \to_P \eta_0$ in $l^\infty(\mathcal{H})$, then $\sqrt{n}(\widetilde{\eta}_n - \eta_0)$ converges weakly to $-\dot{W}_0^{-1}\mathbb{G}$. In fact, one has $\dot{W}_0\sqrt{n}(\widetilde{\eta}_n - \eta_0) = -\sqrt{n}(W_n - W)(\eta_0) + o_P^*(1)$.*

The previous theorem gives a fair amount of flexibility to choose the classes $\mathcal{H}$ and $\mathcal{H}'$. The three main conditions to be verified are the differentiability (2.3) of W, the continuity of the inverse $\dot{W}_0$ and the tightness properties (2.1) and (2.2) of the processes $\sqrt{n}(W_n - W)(\eta)$. For our special choice of W_n the latter conditions can be expressed in the language of empirical process theory; cf. Dudley (1984). Condition (2.2) is exactly that the class $\{A_0\mathcal{H}'\}$ be P_0-Donsker. Condition (2.1) can be checked with the help of exponential inequalities, but for our purpose it suffices that it is implied by the class of functions $\{A_\eta h\colon \|\eta - \eta_0\| < \varepsilon, h \in \mathcal{H}'\}$ being Donsker for some $\varepsilon > 0$ plus a second moment condition.

Informally the derivative of W can be derived as follows. For $\eta \approx \eta_0$,

$$\begin{aligned}(W(\eta) - W(\eta_0))h &= -\int A_\eta h \, d(P_\eta - P_0) \\ &\approx -\int A_0 h \, d(P_\eta - P_0) \\ &= -\int (I + l^* l_0)h \, d(\eta - \eta_0) = \dot{W}_0(\eta - \eta_0)h,\end{aligned}$$

where we use the definition of A_0 and Fubini's theorem in the third step. (Alternatively, this step is an immediate consequence of the fact that l^* is the adjoint of both l_η and l_0.) These remarks lead to the following reformulation of the previous result.

COROLLARY 2.2. *Let $\mathcal{H}$ and $\mathcal{H}'$ be classes of bounded functions h: $\mathcal{Z} \to \mathbb{R}$. Assume that for some $\varepsilon > 0$,*

$$\{A_\eta h\colon h \in \mathcal{H}', \|\eta - \eta_0\|_{\mathcal{H}} < \varepsilon\} \text{ is } P_0\text{-Donsker}. \tag{2.4}$$

Furthermore, assume that as $\eta \to \eta_0$ within the range of the estimators $\widetilde{\eta}_n$,

$$\sup_{h \in \mathcal{H}'} \mathrm{E}_0\big[A_\eta h(X,Z) - A_0 h(X,Z)\big]^2 \to 0, \tag{2.5}$$

$$\frac{\sup_{h \in \mathcal{H}'} |\int l^*(l_\eta - l_0)h \, d[\eta - \eta_0]|}{\sup_{h \in \mathcal{H}} |\int h \, d[\eta - \eta_0]|} \to 0. \tag{2.6}$$

Moreover, assume that for every probability measure η_1, η_2,

$$\frac{\sup_{h \in \mathcal{H}'} |\int (I + l^* l_0)h \, d(\eta_1 - \eta_2)|}{\sup_{h \in \mathcal{H}} |\int h \, d[\eta_1 - \eta_2]|} \geq \varepsilon > 0. \tag{2.7}$$

Then $\sqrt{n}(\widetilde{\eta}_n - \eta_0)$ converges under η_0 to a tight Gaussian process in $l^\infty(\mathcal{H})$, provided $\widetilde{\eta}_n$ is consistent for η_0 in this space. In particular, the sequence $\sqrt{n}(\int h\, d\widetilde{\eta}_n - \int h\, d\eta_0)$ converges under η_0 in distribution to a zero-mean normal distribution with the optimal variance $\sigma^2_{\eta_0}(h)$ for every $h \in \mathcal{H}$.

3. Main results. This section gives sufficient conditions for asymptotic normality of the maximum likelihood estimator that are more readily verifiable than those of the previous corollary, although they are still sufficiently abstract to apply to a variety of examples. The main idea is to translate the conditions of the previous corollary into simple conditions on the score and information operator. It will be required that the operators l^* and l_0 are themselves continuous or compact with respect to a suitable norm, for which, moreover, the dependence $\eta \to l_\eta$ is continuous. Appropriate norms for our examples will be specified in the next sections and typically involve Lipschitz or variational norms, the uniform norm being too weak. These choices correspond to the fact that l_η and l^* are often smoothing operators: they transform bounded functions $h\colon \mathcal{Z} \to \mathbb{R}$ and $g\colon \mathcal{X} \to \mathbb{R}$, respectively, into smooth functions on the other space. This is immediately clear if the kernel $p(x \mid z)$ is smooth in either z or x, but is even true for such irregular kernels as the uniform ones.

A first result of this type is as follows. In all our examples the total variation distance between the probability measures with densities $p(\cdot \mid z_1)$ and $p(\cdot \mid z_2)$ is proportional to the Euclidean distance $\|z_1 - z_2\|$. As a consequence l^* transforms bounded functions into Lipschitz functions and this operator is compact with respect to a lower-order Lipschitz norm. Given a semimetric space $\mathcal{Z}$, let $C^\alpha(\mathcal{Z})$ be the set of all uniformly bounded functions $h\colon \mathcal{Z} \to \mathbb{R}$ that are Lipschitz of order $\alpha \in (0, 1]$. Equip this space with the norm

$$\|h\|_\alpha = \|h\|_\infty \vee \sup_{d(z_1, z_2) > 0} \frac{|h(z_1) - h(z_2)|}{d(z_1, z_2)^\alpha}.$$

Recall that an operator from one normed space into another is *compact* if it maps bounded sets into precompact sets. Equivalently, $A\colon X \to Y$ is compact if the sequence Ah_n is relatively compact in Y (every subsequence has a further converging subsequence) for every sequence h_n in X with $\|h_n\| \le 1$.

LEMMA 3.1. *Let $\mathcal{Z}$ be a semimetric space and assume that*

$$\int |p(x \mid z_1) - p(x \mid z_2)|\, d\mu(x) \le d^\alpha(z_1, z_2),$$

for some $\alpha > 0$. Then the range of l^* is contained in $C^\alpha(\mathcal{Z})$ and $l^*\colon l^\infty(\mathcal{X}) \to C^\alpha(\mathcal{Z})$ is continuous. Consequently, if $\mathcal{Z}$ is totally bounded, then the operators $l^*\colon l^\infty(\mathcal{X}) \to C^\beta(\mathcal{Z})$ and $l^* l_\eta\colon l^\infty(\mathcal{Z}) \to C^\beta(\mathcal{Z})$ are compact for every $\beta < \alpha$.

PROOF. Since l^* is a conditional expectation operator, we have $\|l^* g\|_\infty \le \|g\|_\infty$ for every g. The continuity assertion is an immediate consequence of this

and the inequality

$$|l^*g(z_1) - l^*g(z_2)| \leq \int |g(x)| \left| p(x \mid z_1) - p(x \mid z_2) \right| d\mu(x).$$

To prove the second assertion, let g_n be a sequence with $\|g_n\|_\infty \leq 1$ for every n. Then $\|l^*g_n\|_\alpha$ is uniformly bounded by continuity of l^*. By an extension of the Arzela–Ascoli theorem, this implies that the sequence l^*g_n is precompact in lower-order Lipschitz norms. Hence l^* is compact for the bounded Lipschitz of order $\beta < \alpha$.

The composition of a compact and a continuous operator is always compact, as is easy to see. Therefore the final assertion follows from the second assertion and the observation that l_η is a conditional expectation operator, hence is continuous for the uniform norm. □

For verification of the condition that $\dot{W}_0$ is continuously invertible, we will need the fact that the operator $I + l^*l_0$ is continuously invertible for a suitable norm. A well-known result from functional analysis asserts that the sum of a continuously invertible operator (such as the identity) and a compact operator from a Banach space into itself is Fredholm. In particular, it is onto and has a continuous inverse if and only if it is one to one. We use this result for the operator $I + l^*l_0$, which is usually one to one. Recall that μ is the dominating measure for the set of densities $x \to p(x \mid z)$.

LEMMA 3.2. *If $\mu \ll P_0$, then the operator $I + l^*l_0$: $l^\infty(\mathcal{Z}) \to l^\infty(\mathcal{Z})$ is one to one.*

PROOF. Suppose that $(I + l^*l_0)h = 0$. Since l^*l_0 is a self-adjoint, positive-definite operator of the Hilbert space $L_2(\eta_0)$ into itself, the spectrum of the operator $I + l^*l_0$ is contained in the interval $[1, \infty)$ and $I + l^*l_0$ is certainly continuously invertible with respect to the Hilbert space norm. In particular, it follows that $h = 0$ in the Hilbert space sense, which means almost surely under η_0. Then $l_0h = 0$ almost surely under P_0 by the definition of l_0 and therefore almost surely under μ by assumption. Finally, $h = -l^*l_0h$ is identically 0 by the definition of l^*. □

It was seen that the operator l^* has good properties with respect to the bounded Lipschitz norm $\|\cdot\|_\alpha$ under fairly weak conditions on the kernel $p(x \mid z)$. Since the norm $\|\cdot\|_\alpha$ is not strong enough for all applications, in particular when $\mathcal{Z}$ is higher dimensional, the main result of this paper will be formulated for a general norm, denoted $\|\cdot\|$. Let $(H, \|\cdot\|)$ be a Banach space of bounded functions h: $\mathcal{Z} \to \mathbb{R}$ and let H_1 be its unit ball $\{h$: $\|h\| \leq 1\}$, Special choices of H are made later on with a minimal requirement being that the operator l^*l_0 maps H into itself. The following theorem assumes in addition that l^*l_0 is compact and that the map $\eta \to l^*l_\eta$ is continuous in the operator norm of H. Together with two weak regularity conditions these conditions suffice for verification of (2.5)–(2.7). Unfortunately, it appears difficult to express condition (2.4) that the class of functions $A_\eta h$ be Donsker in simple conditions on the kernel $p(x \mid z)$. This

condition is simply restated in the following result and is addressed separately in a later section.

THEOREM 3.3. *Let $(H, \|\cdot\|)$ be a Banach space contained in $l^\infty(\mathcal{Z})$ such that $\|\cdot\|_\infty \le \|\cdot\|$ and as $\|\eta - \eta_0\|_{H_1} \to 0$,*

$$\{A_\eta h\colon h \in H_1, \|\eta - \eta_0\|_{H_1} < \varepsilon\} \text{ is } P_0\text{-Donsker}, \tag{3.1}$$

$$\mu \ll P_0, \tag{3.2}$$

$$l^* l_0\colon H \to H \text{ is compact}, \tag{3.3}$$

$$\sup_{h \in H_1} |(l_\eta - l_0)h| \to 0, \quad P_0\text{-a.s.}, \tag{3.4}$$

$$\sup_{h \in H_1} \|l^*(l_\eta - l_0)h\| \to 0. \tag{3.5}$$

Then the conditions of Corollary 2.2 are satisfied for $\mathcal{H}' = \mathcal{H} = H_1$. Consequently, if $\|\widetilde{\eta}_n - \eta_0\|_{H_1} \to 0$ in probability under η_0, then $\sqrt{n}(\widetilde{\eta}_n - \eta_0)$ converges weakly in the space $l^\infty(H_1)$ to a tight Gaussian variable, whose marginals have zero means and variances $\sigma^2_{\eta_0}(h)$.

PROOF. By (3.2) and the preceding lemma, the operator $I + l^* l_0$ is one to one. Combination with the compactness (3.3) shows that $I + l^* l_0\colon H \to H$ is onto and continuously invertible. Thus $CH_1 \subset (I + l^* l_0)H_1$ for the positive constant $C = \|(I + l^* l_0)^{-1}\|^{-1}$. This immediately yields the continuity (2.7) of $\dot{W}_0^{-1}$.

Next (3.5) is a different way of saying that $l^*(l_n - l_0)H_1 \subset \varepsilon_n H_1$ for some sequence ε_η with $\varepsilon_\eta \to 0$ as $\eta \to \eta_0$. Hence the quotient in (2.6) is bounded above by ε_η.

Finally, (2.5) follows from (3.4) and the dominated convergence theorem. □

Clearly, in a particular application the norm $\|\cdot\|$ should be chosen so as to satisfy conditions (3.1) and (3.3)–(3.5) at the same time. It is of interest that the last three conditions can be relaxed considerably if the norm is chosen such that the unit ball H_1 is totally bounded for the uniform norm. In view of the Arzela–Ascoli theorem this would be the case, for instance, for $\|\cdot\|$ stronger than a bounded Lipschitz norm on a compact metric space $\mathcal{Z}$.

LEMMA 3.4. *Suppose that in the situation of the previous theorem the unit ball H_1 is totally bounded for the uniform norm. Then* (3.3)–(3.5) *are implied by: for all h,*

$$l^*\colon l^\infty(\mathcal{X}) \to H \text{ is continuous}, \tag{3.3$'$}$$

$$|(l_\eta - l_0)h| \to 0, \quad P_0\text{-a.s.}, \tag{3.4$'$}$$

$$\|l^*(l_\eta - l_0)h\| \to 0. \tag{3.5$'$}$$

More precisely, condition (3.3′) *implies* (3.3) *and under* (3.3′) *conditions* (3.4) *and* (3.5) *are implied by* (3.4′) *and* (3.5′), *respectively.*

PROOF. The assumption that H_1 is totally bounded in $l^\infty(\mathcal{Z})$ is a different way of saying that the identity I: $H \to l^\infty(\mathcal{Z})$ is compact. Next l_0: $l^\infty(\mathcal{Z}) \to l^\infty(\mathcal{X})$ is continuous, because it is a conditional expectation operator and l^*: $l^\infty(\mathcal{X}) \to H$ is continuous by assumption (3.3′). The composition $l^* l_0 I$: $H \to H$ of one compact and two continuous operators is compact.

Since each of the operators l_η is a conditional expectation operator, the difference $l_\eta - l_0$: $l^\infty(\mathcal{Z}) \to l^\infty(\mathcal{Z})$ is continuous with norm bounded above by 2. Hence

$$\|l^*(l_\eta - l_0)h_1 - l^*(l_\eta - l_0)h_2\| \le \|l^*\| \, \|(l_\eta - l_0)(h_1 - h_2)\|_\infty \le \|l^*\| 2 \|h_1 - h_2\|_\infty.$$

Take an ε-net H_ε over the unit ball H_1 for the uniform norm. Then

$$\sup_{h \in H_1} \|l^*(l_\eta - l_0)h\| \le \sup_{h \in H_\varepsilon} \|l^*(l_\eta - l_0)h\| + 2\varepsilon \|l^*\|.$$

The first term goes to 0 for every fixed ε by (3.5′). The second can be made arbitrarily small by total boundedness of H_1. Thus (3.5) follows.

It can be argued in a similar manner that (3.4) follows from (3.4′). □

A set of functions that is totally bounded for the uniform norm can have discontinuities at at most finitely many points in $\mathcal{Z}$. This makes the assumption of the preceding lemma that H_1 be totally bounded for the uniform norm in certain situations unattractive. For instance, only finitely many indicator functions $h = 1_{(-\infty, t]}$ can be included in H_1. This has as a consequence that the final result asserts asymptotic normality of $\int h \, d\widetilde{\eta}_n$ uniformly over a certain set of smooth functions h and a finite set of indicator functions, but not uniformly over all indicator functions. Since typically the finite set of indicator functions can be varied, asymptotic normality for all such functions can be proved, albeit not uniformly. Thus it is for instance obtained that the (normalized) marginals of the cumulative distribution function $\widetilde{\eta}_n(t)$ converge weakly, but the question as to whether the distribution function converges weakly as a process is left open by the previous lemma. In contrast, Theorem 3.3 can give a positive answer to this question, provided its stronger assumptions can be checked.

A second lemma strengthens the assumption (3.3′) to compactness of l^*. While the continuity (3.3′) turns out to hold in all our examples, compactness (for a sufficiently strong norm) typically requires more smoothness of the kernel $p(x \mid z)$ in z. Compactness allows us to drop the continuity condition (3.5) or (3.5′) on $\eta \to l^* l_\eta$ altogether.

LEMMA 3.5. *In the situation of the preceding theorem or lemma, suppose that the operator* l^*: $l^\infty(\mathcal{X}) \to H$ *is compact and* $\|\cdot\|_\infty \le \|\cdot\|$. *Suppose that* (3.2) *holds. Then* (3.5) *follows from* (3.4) *and* (3.5′) *from* (3.4′).

PROOF. Take any sequences h_n in H_1 and $\eta_n \to \eta_0$. By the assumed compactness of l^*, the sequence $l^*(l_{\eta_n} - l_0)h_n$ is relatively compact for $\|\cdot\|$. Its limit points can be identified from pointwise convergence. Now

$$l^*(l_{\eta_n} - l_0)h_n(z) = \int (l_{\eta_n} - l_0)h(x)p(x \mid z)d\mu(x).$$

Under (3.4) the integrand converges to 0 for P_0-almost all x, hence for μ-almost all x. By the dominated convergence theorem the integral converges to 0 for all fixed z. This concludes the proof that (3.4) implies (3.5).

To prove the same implication, but with primes, apply the same argument with $h_n = h$ fixed. □

4. Donsker classes. Unfortunately, there appears to be no simple way of expressing the condition (3.1) that the class $\{A_\eta H_1 \colon \|\eta - \eta_0\| \leq \varepsilon\}$ be Donsker directly in properties of the kernel $p(x \mid z)$. However, examples suggest that the Donsker property is often satisfied. In this section we discuss two general approaches toward verifying the condition. The first method is applicable if the kernel $x \to p(x \mid z)$ is smooth in x, while the second is applicable in situations where the random variables $(X, Y) \sim p(x \mid y)\, d\eta(y)$ are positively dependent or satisfy a related condition. We note that, while these two approaches cover all our examples, recent advances in empirical process theory have yielded many other potentially useful characterizations of Donsker classes.

4.1. *Smoothness of $x \to l_\eta h(x)$.* It is well known that classes of smooth functions on a (bounded or unbounded) subset of Euclidean space are Donsker classes. To define such classes, let for a given function $g \colon \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ and $\alpha > 0$,

$$\|g\|_\alpha = \max_{k_\cdot \leq \lfloor \alpha \rfloor} \sup_x |D^k g(x)| \vee \max_{k_\cdot = \lfloor \alpha \rfloor} \sup_{x,\, y} \frac{|D^k g(x) - D^k g(y)|}{\|x - y\|^{\alpha - \lfloor \alpha \rfloor}},$$

where the suprema are taken over all x, y in the interior of $\mathcal{X}$ with $x \neq y$, the value $\lfloor \alpha \rfloor$ is the greatest integer strictly smaller than α and for each vector k of d integers D^k is the differential operator

$$D^k = \frac{\partial^{k_\cdot}}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}},$$

where $k_\cdot = \sum k_i$. Note that for $\alpha \leq 1$ the norm $\|\cdot\|_\alpha$ is simply the Lipschitz norm of order α introduced earlier, while for larger values of α the norm involves bounds on the partial derivatives of g together with a Lipschitz norm on the partial derivatives of highest order. Let $C_M^\alpha(\mathcal{X})$ be the set of all continuous functions $g \colon \mathcal{X} \to \mathbb{R}$ with $\|g\|_\alpha \leq M$. Giné and Zinn (1986) show that the class $C_1^1(\mathbb{R})$ is P-Donsker if and only if $\sum_{j=-\infty}^{\infty} P^{1/2}[j, j+1) < \infty$. The following extension of their result is proved by van der Vaart (1993).

LEMMA 4.1. *Let $\mathcal{X} = \bigcup_{j=1}^{\infty} I_j$ be a partition of $\mathbb{R}^d$ into bounded, convex sets whose Lebesgue measure is bounded uniformly away from 0 and ∞. Let $\mathcal{G}$ be a class of functions $g \colon \mathcal{X} \to \mathbb{R}$ such that the restrictions $g_{|I_j}$ belong to $C_{M_j}^\alpha(I_j)$ for every j and some fixed $\alpha > d/2$. Then $\mathcal{G}$ is P-Donsker for every probability measure P on $\mathcal{X}$ such that $\sum_{j=1}^{\infty} M_j P^{1/2}(I_j) < \infty$.*

The given tail condition on P is fairly weak. For instance, for a subset of the real line, a partition in cells $I_j = \{x: j \le |x| < j+1\}$ and $M_j = j^k$, it suffices that $\int |x|^m \, dP(x) < \infty$ for some $m > 2k+1$.

Thus, for the class $\{l_\eta h: h \in H_1, \|\eta - \eta_0\| < \varepsilon\}$ to be Donsker, a uniform α-smoothness condition on the functions $l_\eta h$ for some $\alpha > d/2$ will usually suffice. Under appropriate conditions on the map $x \to p(x \mid z)$, straightforward differentiation yields

$$\frac{\partial}{\partial x_{i'}} l_\eta h(x) = \operatorname{cov}_x\left(h(Z), \frac{\partial}{\partial x_i} \log p(x \mid Z)\right), \tag{4.1}$$

where for each x the covariance is computed for the random variable Z having the (conditional) density $z \to p(x \mid z)\,d\eta(z)/p(x \mid \eta)$. Thus, for a given bounded function h,

$$\left|\frac{\partial}{\partial x_i} l_\eta h(x)\right| \le \|h\|_\infty \frac{\int \left|(\partial/\partial x_i) \log p(x \mid z)\right| p(x \mid z) d\eta(z)}{\int p(x \mid z) d\eta(z)}.$$

Dependent on the function $(\partial/\partial x_i) \log p(x \mid z)$, this leads to a bound on the first derivative and hence on the Lipschitz constant of order 1 of the function $x \to l_\eta h(x)$. If $\mathcal{X}$ is an interval in $\mathbb{R}$, this is sufficient for applicability of the previous lemma.

EXAMPLE 1 (Normal deconvolution). Let $p(x \mid z) = z_2^{-1}\phi(z_2^{-1}(x - z_1))$ be the normal density with mean z_1 and standard deviation z_2. Then $(\partial/\partial x) \log p(x \mid z) = -z_2^{-2}(x - z_1)$. If $\eta(\delta \le z_2 \le \delta^{-1}) = 1$ for some $\delta > 0$, then the preceding argument gives the estimate

$$\left|\frac{\partial}{\partial x_i} l_\eta h(x)\right| \le \frac{\int z_2^{-3}|x - z_1| \phi\left(z_2^{-1}(x - z_1)\right) d\eta(z)}{\int z_2^{-1} \phi\left(z_2^{-1}(x - z_1)\right) d\eta(z)} \le \int \frac{|x - z_1|}{\delta^2 z_2^2} d\eta(z).$$

The right side is bounded by $\delta^{-4}(|x| + \int |z_1| \, d\eta(z))$. One possible conclusion is that the set of functions $l_\eta h$ formed by letting h range over all functions with $\|h\|_\infty \le 1$ and η range over all probability measures on a fixed compact subset of $\mathbb{R} \times (0, \infty)$ is Donsker.

Similar arguments apply if $\mathcal{X}$ is a subset of a higher-dimensional Euclidean space, although it becomes necessary to consider higher-order derivatives. For instance, in dimension 2 any Lipschitz condition on the first-order partial derivatives suffices ($\alpha > 1$), while in dimension 3 we need a Lipschitz condition of order $>1/2$ on these derivatives ($\alpha > 3/2$). Straightforward calculations show that

$$\begin{aligned}\frac{\partial^2}{\partial x_i \, \partial x_j} l_\eta h(x) = {} & \operatorname{cov}_x\left(h(Z), \frac{\partial^2}{\partial x_i \, \partial x_j} \log p(x \mid Z)\right) \\ & - \operatorname{cov}_x\left(h(Z) \frac{\partial}{\partial x_i} \log p(x \mid Z)\right) \mathrm{E}_x \frac{\partial}{\partial x_j} \log p(x \mid Z) \\ & - \operatorname{cov}_x\left(h(Z) \frac{\partial}{\partial x_j} \log p(x \mid Z)\right) \mathrm{E}_x \frac{\partial}{\partial x_i} \log p(x \mid Z).\end{aligned}$$

This expression can be bounded as before.

EXAMPLE 2 (Exponential family). Let $p(x \mid z) = h_0(x)c(z)\exp(z't(x))$ range through a k-dimensional exponential family, where x ranges over a convex subset of either $\mathbb{R}^2$ or $\mathbb{R}^3$ and t is sufficiently smooth. Then

$$\left|\frac{\partial^2}{\partial x_i\,\partial x_j} l_\eta h(x)\right| \le 2\mathrm{E}_x|Z|'\left|\frac{\partial^2}{\partial x_i\,\partial x_j} t(x)\right| + 4\mathrm{E}_x|Z|'\left|\frac{\partial}{\partial x_j} t(x)\right|\mathrm{E}_x|Z|'\left|\frac{\partial}{\partial x_i} t(x)\right|.$$

These functions are appropriately bounded for application of the previous lemma if η ranges over all distributions on a given compact set and the partial derivatives of t behave well. Of course, the smoothness of t can be assured by parameterizing the exponential family in its standard form with $t(x) = x$.

4.2. *Functions of bounded variation.* In this section we assume that $\mathcal{Z}$ and $\mathcal{X}$ are intervals in the real line. It turns out that in many examples the score operator has the following property: if the function $z \to h(z)$ is nondecreasing, then so is $x \to l_\eta h(x)$. In view of the fact that $l_\eta h(X) = E_\eta(h(Y) \mid X)$ where (X, Y) has density $p(x \mid y)\,d\eta(y)$, this property is equivalent to the set of conditional distributions $P^{Y \mid X = x}$ being stochastically increasing in x.

Since l_η also maps bounded functions into bounded functions, it follows that l_η is continuous for the variation norm. For a function $h\colon \mathcal{Z} \to \mathbb{R}$ define the bounded variation norm as

$$\|h\|_{\mathrm{BBV}} = \|h\|_\infty \vee \sup \sum_i |h(t_{i+1}) - h(t_i)|,$$

where the supermum is over all partitions $t_0 < t_1 < \cdots < t_m$ of $\mathcal{Z}$. Let $\mathrm{BBV}(\mathcal{Z})$ and $\mathrm{BBV}_1(\mathcal{Z})$ be the Banach space of all functions with $\|h\|_{\mathrm{BBV}} < \infty$ and its unit ball, respectively. Then for every l_η with the monotonicity property as in the preceding paragraph,

$$l_\eta\,\mathrm{BBV}_1(\mathcal{Z}) \subset 2\,\mathrm{BBV}_2(\mathcal{X}).$$

It is well known that every set of functions that is of uniformly bounded variation is universally Donsker. Consequently, under the monotonicity property as in the first paragraph, condition (3.1) is verified for any subset H_1 of $\mathrm{BBV}_1(\mathcal{Z})$. More generally, the same conclusion is valid in any situation where the operators $l_\eta\colon \mathrm{BBV}(\mathcal{Z}) \to \mathrm{BBV}(\mathcal{X})$ are equicontinuous.

For smooth kernels $x \to p(x \mid z)$ monotonicity of $x \to l_\eta h(x)$ for a given monotone h is most easily checked from formula (4.1) for the derivative. Since two increasing functions $h(Z)$ and $f(Z)$ of the same variable are always positively correlated, it follows that the derivative of $l_\eta h(x)$ is nonnegative whenever the function $z \to (\partial/\partial x)\log p(x \mid z)$ is nondecreasing.

EXAMPLE 3 (Unimodal deconvolution). Let $p(x \mid z) = p(x - z)$ for a smooth, strongly unimodal density p. Then $(\partial/\partial x)\log p(x \mid z) = p'/p(x - z)$ is increasing

in z, because (by definition) $\log p$ is concave. Thus condition (3.1) is valid for $H_1 \subset \mathrm{BBV}_1(\mathcal{Z})$.

The smoothness condition on p can be relaxed by a direct argument, so as to cover also the double exponential density.

EXAMPLE 4 (Exponential family). Let $p(x \mid z) = h_0(x)c(z)\exp(zt(x))$ be a one-dimensional exponential family. Assume that t is differentiable. Then the function $(\partial/\partial x)\log p(x \mid z) = h_0'/h_0(x) + zt'(x)$ is increasing if $t'(x) \geq 0$ and decreasing otherwise. It follows from a slight modification of the preceding argument that condition (3.1) is satisfied for $H_1 \subset \mathrm{BBV}_1(\mathcal{Z})$ provided t' has only finitely many sign changes. The latter is certainly true if the exponential family is taken in its standard form with $t(x) = x$.

For such kernels $p(\cdot \mid z)$ as corresponding to the uniform distribution on $[0, z]$ or the uniform distribution on $[z, z+1]$, the function $x \to l_\eta h(x)$ is not differentiable for every η, but the basic monotonicity property is still valid. More generally, we have the following result.

LEMMA 4.2. *Let $p(x \mid z) = c(z)h_0(x)1\{\phi(x) < z < \psi(x)]\}$ for fixed functions $\phi, \psi\colon \mathbb{R} \to \overline{\mathbb{R}}$ that are either strictly increasing or else identically $-\infty$ or $+\infty$ and strictly positive h_0. Suppose the map $z \to h(z)$ is nondecreasing. Then the map $x \to l_\eta h(x)$ is nondecreasing on every interval where $p(x \mid \eta) > 0$ for every x. The same is true if one or both of the inequalities in the definition of $p(x \mid z)$ is replaced by a less than or equal sign.*

PROOF. Fix $x_1 \leq x_2$. Each value $l_\eta h(x)$ is a weighted average of h over the interval $(\phi(x), \psi(x))$. If $\psi(x_1) \leq \phi(x_2)$, then the intervals are disjoint and $l_\eta h(x_1) \leq l_\eta h(x_2)$ trivially. Otherwise write

$$l_\eta h(x_2) = \frac{\int_{(\phi(x_2),\,\psi(x_1)]} hc\,d\eta \int_{(\phi(x_2),\,\psi(x_1)]} c\,d\eta}{\int_{(\phi(x_2),\,\psi(x_1)]} c\,d\eta \int_{(\phi(x_2),\,\psi(x_2)]} c\,d\eta} + \frac{\int_{(\psi(x_1),\,\psi(x_2)]} hc\,d\eta \int_{(\psi(x_1),\,\psi(x_2)]} c\,d\eta}{\int_{(\psi(x_1),\,\psi(x_2)]} c\,d\eta \int_{(\phi(x_2),\,\psi(x_2)]} c\,d\eta}.$$

This expression has the form $\overline{h}_2\lambda + \overline{h}_3(1-\lambda)$, where $\overline{h}_2$ and $\overline{h}_3$ are weighted averages of h over the intervals $(\phi(x_2), \psi(x_1)]$ and $(\psi(x_1), \psi(x_2)]$, respectively. The expression is not smaller than $\overline{h}_2$, which is bounded below by $\overline{h}_1\mu + \overline{h}_2(1-\mu)$ for every $\mu \in [0, 1]$ and weighted average $\overline{h}_1$ over the interval $(\phi(x_1), \phi(x_2)]$. This can be made equal to $l_\eta h(x_1)$ by appropriate choice of μ and $\overline{h}_1$.

In the above argument the possibility that one of the denominators is 0 is neglected. By the assumption that $p(x_i \mid \eta) > 0$, only one of the two denominators in each $l_\eta h(x_i)$ can be 0. In that case λ and μ may be 0 or 1 and the equality $l_\eta h(x_1) \leq l_\eta h(x_2)$ can be checked easily. □

EXAMPLE 5 (Uniform scale). For $p(x \mid z) = (1/z)1\{[0, z]\}(x)$ the set of values x where $p(x \mid \eta) = \int_{[x,\,\infty)}(1/z)\,d\eta(z)$ is positive is an interval of the form $(0, x_0)]$

which may or may not include its right endpoint. Thus l_η maps an increasing function h into a function that increases on $(0, x_0)]$ and is identically 0 afterwards. If $\|h_\infty\| \leq 1$, then the variation of $l_\eta h$ is at most 2. Hence the maps l_η: $\mathrm{BBV}(\mathcal{Z}) \to \mathrm{BBV}(\mathcal{X})$ are equicontinuous, and (3.1) is satisfied for every $H_1 \subset \mathrm{BBV}_1(\mathcal{Z})$.

EXAMPLE 6 (Shifted uniform). For $p(x \mid z) = 1\{(z, z+1)\}(x)$ the set on values x where $p(x \mid \eta) = \eta(x-) - \eta(x-1)$ is positive is a union of intervals. For all η that are sufficiently close to some η_0 that charges every interval of length 1, we have $p(x \mid \eta) > 0$ for all x. For such η the operator l_η: $\mathrm{BBV}(\mathcal{Z}) \to \mathrm{BBV}(\mathcal{X})$ is norm bounded by 2, and (3.1) is satisfied for every $H_1 \subset \mathrm{BBV}_1(\mathcal{Z})$.

5. Smooth kernels. Given sufficient smoothness of the kernel $p(x \mid z)$ in z, the conclusion of Lemma 3.1 can be strengthened. The following lemma will be suitable for dimensions 1 to 3. For higher dimensions additional smoothness would be helpful, but we omit a discussion.

LEMMA 5.1. *Let $\mathcal{Z}$ be a convex subset of $\mathbb{R}^d$ and assume that the maps $z \to p(x \mid z)$ are differentiable for each x with partial derivatives $(\partial/\partial z_i)p(x \mid z)$ satisfying*

$$\int \left| \frac{\partial}{\partial z_i} p(x \mid z) - \frac{\partial}{\partial z_i} p(x \mid z') \right| d\mu(x) \leq K \|z - z'\|^\alpha,$$

$$\int \left| \frac{\partial}{\partial z_i} p(x \mid z) \right| d\mu(x) \leq K$$

for all z, z' in $\mathcal{Z}$ and fixed constants K and $\alpha > 0$. Then l^*: $l^\infty(\mathcal{X}) \to C^{1+\alpha}(\mathcal{Z})$ is continuous. Consequently, if $\mathcal{Z}$ is totally bounded, then the operator l^*: $l^\infty(\mathcal{X}) \to C^{1+\beta}(\mathcal{Z})$ is compact for every $\beta < \alpha$.

5.1. *One dimensional $\mathcal{Z}$.* In this subsection suppose that $\mathcal{Z} = [a, b]$ is a compact interval in the real line. Our aim is to verify the conditions of Theorem 3.3 for H equal to the Banach space $\mathrm{BBV}(\mathcal{Z})$ of all uniformly bounded functions of bounded variation equipped with the bounded variation norm $\|\cdot\|_{\mathrm{BBV}}$. This norm satisfies

$$\|h\|_{\mathrm{BBV}} \leq \big((b-a) \vee 1\big) \|h\|_1,$$

where $\|h\|_1$ is the bounded Lipschitz norm of order 1. If the kernel $p(x \mid z)$ satisfies the conditions of the previous lemma, then l^*: $l^\infty(\mathcal{X}) \to C^1(\mathcal{Z})$ is compact for the Lipschitz norm, which is stronger than the bounded variation norm. Therefore, certainly l^*: $l^\infty(\mathcal{X}) \to \mathrm{BBV}(\mathcal{Z})$ is a compact operator. Basically, this leaves only the Donsker condition (3.1) of Theorem 3.3 to be checked. We have the following result.

THEOREM 5.2. *Let $\mathcal{Z} = [a, b]$ be a compact interval in the real line and let the kernel $p(x \mid z)$ satisfy the conditions of the previous lemma. Furthermore, assume*

that $\mu \ll P_0$, that the true underlying distribution η_0 has no discrete component and that $\{l_\eta \mathrm{BBV}_1(\mathcal{Z})\colon \|\eta - \eta_0\| < \varepsilon\}$ is P_0-Donsker for some $\varepsilon > 0$. Then the conditions of Theorem 3.3 are satisfied for $H_1 = \mathrm{BBV}_1(\mathcal{Z})$. Consequently, the maximum likelihood process $\sqrt{n}(\widetilde{\eta}_n - \eta_0)$ converges weakly to the given Gaussian distribution in $l^\infty(H_1)$, provided $\widetilde{\eta}_n$ is consistent for the weak topology.

PROOF. In view of Lemmas 3.5 and 5.1 it suffices to verify (3.4). For fixed x let F_η be the distribution function $F_\eta(z) = \int_{[a,z]} p(x \mid s)\, d\eta(s)$. By partial integration

$$\left| \int h(z) p(x \mid z)\, d\eta(z) - \int h(z) p(x \mid z)\, d\eta_0(z) \right| \leq |h(b)|\, |F_\eta(b) - F_{\eta_0}(b)| + |h(a)|\, |F_\eta\{a\}| + \|h\|_{\mathrm{BBV}} \|F_\eta - F_{\eta_0}\|_\infty.$$

Here $\|F_\eta - F_{\eta_0}\|_\infty \to 0$, because $z \to p(x \mid z)$ is continuous (and hence bounded) and η_0 has no point masses. Thus the numerator of $l_\eta h(x)$ converges to the numerator of $l_0 h(x)$ uniformly in $h \in \mathrm{BBV}_1(\mathcal{Z})$ for each fixed x. Deduce that $l_\eta h(x) \to l_0 h(x)$ uniformly in h for each fixed x. □

EXAMPLE 7 (Deconvolution). Let $p(x \mid z) = p(x - z)$, where p is a bounded, positive, strongly unimodal density with respect to Lebesgue measure on $\mathbb{R}$. Assume that p possesses two bounded, continuous, Lebesgue integrable derivatives. For instance, let p be the normal or the logistic density.

Let the maximization of the likelihood be carried out over all probability distributions η that are supported on some fixed interval $\mathcal{Z}$, that also supports the true underlying distribution η_0. Then the cumulative distribution function of the maximum likelihood estimator $\widetilde{\eta}_n$ is asymptotically normal in $l^\infty(\mathbb{R})$, provided the true distribution is identifiable and has no point masses. [Identifiability means that the Lebesgue measure of the set $\{x\colon p(x \mid \eta_0) \neq p(x \mid \eta)\}$ is positive for every $\eta \neq \eta_0$.]

For the validity of this statement the smoothness conditions on p can be somewhat relaxed. Furthermore, the unimodality is used only to ensure the Donsker condition (through Example 3) and can be replaced by other conditions.

As a concrete example of alternative conditions, consider a bimodal density of the form $p(x) = \lambda\phi(x) + (1 - \lambda)\phi(x - \mu)$ with λ and μ fixed known numbers and ϕ the normal density. In this case the Donsker condition follows from Lemma 4.1 combined with the estimate

$$\left| \frac{\partial}{\partial x} l_\eta h(x) \right| \leq \|h\|_\infty \frac{\int |\lambda\phi'(x - z) + (1 - \lambda)\phi'(x - \mu - z)|\, d\eta(z)}{\int \lambda\phi(x - z) + (1 - \lambda)\phi(x - \mu - z)\, d\eta(z)} \leq \|h\|_\infty \left(\frac{\int |\phi'(x - z)|\, d\eta(z)}{\int \phi(x - z)\, d\eta(z)} + \frac{\int |\phi'(x - z - \mu)\, d\eta(z)}{\int \phi(x - z - \mu)\, d\eta(z)} \right).$$

This is bounded by a constant (depending only on the interval $\mathcal{Z}$) times $|x|$ uniformly in uniformly bounded sets of h. Since the mixture of normal distributions has very thin tails, the Donsker condition of Lemma 4.1 is certainly

satisfied. Hence the maximum likelihood estimator is asymptotically normal by the previous theorem.

5.2. *Two and three-dimensional* $\mathcal{Z}$. In principle, the argument using functions of bounded variation given in the previous subsection can be extended to higher dimensions. This would lead to a proof of uniform asymptotic normality of, for instance, the cumulative distribution function of the maximum likelihood estimator. Here we restrict ourselves to proving asymptotic normality of the maximum likelihood estimator indexed by finitely many indicator functions. For the dimension d of $\mathcal{Z}$ equal to 2 or 3, this is an easy corollary of the results obtained so far. For higher dimensions the same approach can be used, but since $C_1^\alpha(\mathcal{Z})$ is Donsker only if $\alpha > d/2$ Lemma 5.1 needs to be replaced by a stronger result involving higher-order partial derivatives of the map $z \to p(x \mid z)$. We restrict ourselves to dimensions 2 and 3.

Given finitely many points $z_1, \ldots, z_k$ in $\mathcal{Z}$, let $\mathcal{Z} = \bigcup_j \mathcal{Z}_j$ be a partition of $\mathcal{Z}$ in finitely many rectangles, such that no z_i is in the interior of some partitioning set. (For instance, the intersection of all quadrants or octants defined by the z_i.) Let H_1^β be the set of all functions $h\colon \mathcal{Z} \to \mathbb{R}$ whose restrictions $h_{|\mathcal{Z}_j}$ belong to $C_1^\beta(\mathcal{Z}_j)$ for each j. Here β is a fixed number with $\beta > d/2$. Note that H_1^β is precisely the unit ball in the Banach space of all functions $h\colon \mathcal{Z} \to \mathbb{R}$ for which the norm

$$\|h\| = \sup_{j=1,\ldots,k} \left\|h_{|\mathcal{Z}_j}\right\|_\beta$$

is finite. By Lemma 4.1 this is a Donsker class of functions for every probability measure on the compact set $\mathcal{Z}$.

THEOREM 5.3. *Let $\mathcal{Z}$ be a bounded, convex subset of $\mathbb{R}^d$ with d equal to 2 or 3 and let the kernel $p(x \mid z)$ satisfy the conditions of Lemma 5.1 for some $\alpha > d/2 - 1$. Furthermore, assume that $\mu \ll P_0$, that the true underlying distribution η_0 does not charge the boundaries of the partitioning sets $\mathcal{Z}_j$ and that $\{l_\eta H_1^\beta\colon \|\eta - \eta_0\|_{H_1^\beta} \leq \varepsilon\}$ is P_0-Donsker for some $\varepsilon > 0$ and $\beta > d/2$. Then the conditions of Theorem 3.3 are satisfied for H_1^β. Consequently, the vector $(\sqrt{n}(\widetilde{\eta}_n(z_1) - \eta_0(z_1)), \ldots, \sqrt{n}(\widetilde{\eta}_n(z_k) - \eta_0(z_k)))$ is asymptotically normal, provided $\widetilde{\eta}_n$ is consistent for the weak topology.*

PROOF. By Lemma 5.1 the operator $l^*\colon l^\infty(\mathcal{X}) \to C^\beta(\mathcal{Z})$ is compact for all $\beta < 1 + \alpha$, in particular for some β with $d/2 < \beta < 1 + \alpha$. Thus it is certainly compact with respect to the weaker norm $\|\cdot\|$ of H_1^β. By the Ascoli–Arzela theorem H_1^β is precompact for the uniform norm. Combination of Theorem 3.3 and Lemmas 3.4 and 3.5 shows that it suffices to check that $l_\eta h(x) \to l_0 h(x)$ for P_0-almost all x and every $h \in H_1^\beta$. This is immediate from the definitions, the continuity of $z \to p(x \mid z)$ and the condition on the support of η_0. □

EXAMPLE 8 (Normal deconvolution). Let $p(x \mid z) = z_2^{-1}\phi(z_2^{-1}(x - z_1))$ be the normal density with mean z_1 and standard deviation z_2. Take $\mathcal{Z}$ a compact sub-

set of $\mathbb{R} \times (0, \infty)$. Assume that the true underlying distribution η_0 is Lebesgue absolutely continuous. Then the cumulative distribution function of the maximum likelihood estimator is marginally asymptotically normal.

The normal density in this example can be replaced by any other smooth density, provided the Donsker condition can be checked. For the normal density the Donsker condition is verified in Example 1.

6. Some nonsmooth kernels. In this section we discuss three examples, where the support of the density $p(\cdot \mid z)$ depends on z. In each of the three examples the maximum likelihood estimator is asymptotically normal. This is obtained directly from Theorem 3.3 by approximately the same method for each of them. We discuss the method in detail for the first example and are brief for the second and the third.

6.1. *Shifted uniform.* Let $p(x \mid z) = 1\{[z, z+1]\}(x)$ be the density of the uniform distribution on $[z, z+1]$. Then

$$p(x \mid \eta) = \eta(x) - \eta(x-1-),$$

$$l_\eta h(x) = \frac{\int_{[x-1,x]} h(z)\, d\eta(z)}{p(x \mid \eta)},$$

$$l^* g(z) = \int_{[z,z+1]} g(x)\, dx.$$

Inspection of the likelihood function shows that the maximum likelihood estimator is not uniquely determined in this problem. For instance, the likelihood remains unchanged if mass is moved left in any interval with right endpoint x_i, provided the mass is not moved farther left than some other observation or some $x_j - 1$. Since the likelihood is nondecreasing if mass is moved to the closest x_i or z_j to the right, there always exists a maximum likelihood estimator that is supported on the observations x_i, z_j. A similar argument shows that there always exists a maximum likelihood estimator that is supported on the points $x_i - 1, z_j$. For consistency the nonuniqueness does not make any difference: van der Vaart and Wellner (1993) show that any sequence of maximum likelihood estimators is consistent under every η_0.

There always exists a maximum likelihood estimator that is supported on the interval $[\min x_i \wedge \min z_j, \max(x_i - 1) \vee \max z_j]$. This has the pleasant property that it is supported on the convex hull of the support of the true distribution η_0. Here we prove asymptotic normality of maximum likelihood estimators with this additional property, under the assumption that η_0 has compact support (which need not be known). Maximum likelihood estimators without this additional property are shown to be asymptotically normal under every compactly supported η_0 whose support contains the far left and right ends of $\mathbb{Z}$. (The point is that the last, unnatural condition on η_0 is not necessary if the maximum likelihood estimators are constructed to have support within the convex hull of the support of η_0.)

The proof is based on Theorem 3.3 applied with the norm

$$\|h\| = \sup_{z_1 \neq z_2 < a+\varepsilon} \frac{|h(z_1) - h(z_2)|}{|z_1 - z_2|^{1/2}} \vee \|h\|_{\mathrm{BBV}} \vee \sup_{z_1 \neq z_2 > b-\varepsilon} \frac{|h(z_1) - h(z_2)|}{|z_1 - z_2|^{1/2}}.$$

Here the interval $[a, b]$ is the convex hull of the support of η_0 and $\varepsilon > 0$ is fixed. Note that the norm is made to depend on the true underlying parameter. The two Lipschitz parts in this norm are motivated by the technical problem that the denominator $p(x \mid \eta_0)$ of $l_0 h(x)$ converges to 0 as $x \downarrow a$ or $x \uparrow b$. The extra control over h provided by the Lipschitz norms counterbalance this.

Let H be the set of all functions $h\colon [a, b] \to \mathbb{R}$ with $\|h\| < \infty$. A key observation is the following lemma.

LEMMA 6.1. *Let η_0 have no discrete component and compact support. Moreover, assume that η_0 charges every interval of length 1 contained in the convex hull $[a, b] \subset \mathbb{R}$ of its support. Then $l_0\colon H \to l^\infty(\mathcal{X})$ is compact.*

PROOF. Take any sequence h_n in the unit ball of H. Since $\|\cdot\|$ includes a Lipschitz norm on $(a, a+\varepsilon)$ and $(b-\varepsilon, b)$, there exists by the Arzela–Ascoli theorem a subsequence $h_{n'}$ whose restrictions to these intervals converge uniformly to a limit. By the definition of l_0 it follows that $l_0 h_{n'}$ converges uniformly on $(a, a+\varepsilon)$ and $(b+1-\varepsilon, b+1)$.

Since the sequence h_n is of uniformly bounded variation, there exists by Helly's theorem a further subsequence along which the positive and negative variations $h_{n''}^+$ and $h_{n''}^-$ converge pointwise to monotone functions h_1 and h_2 at every continuity point of h_1 and h_2. The set of discontinuity points consists of at most countably many points, hence is a null set for η_0. By the definition of l_0 we have $l_0 h_{n''}^+ \to l_0 h_1$ and $l_0 h_{n''}^- \to l_0 h_2$ for every $x \in (a, b+1)$. Since every function of the form $x \to l_0 h(x)$ is continuous and the functions $l_0 h_n^+$ and $l_0 h_n^-$ are monotone, the convergence is uniform on every compact subinterval $[a+\varepsilon, b+1-\varepsilon]$.

Combination of the two previous paragraphs yields that $l_0 h_{n''}$ converges uniformly on $(a, b+1)$. □

By Lemma 3.1 the operator $l^*\colon l^\infty(\mathcal{X}) \to C^1[a, b]$ is continuous. Combination with the previous lemma shows that $l^* l_0\colon H \to C^1[a, b]$ is compact. For a compact interval the Lipschitz norm of order 1 is stronger than the bounded variation norm. Thus $l^* l_0\colon H \to H$ is compact, verifying a key condition of Theorem 3.3. It is of interest that in this example the compactness results from l_0, rather than from l^* as in the case of smooth kernels.

THEOREM 6.2. *Let η_0 have no discrete component and compact support that has no holes of length 1. Then the conditions of Theorem 3.3 are satisfied for the norm $\|\cdot\|$ as specified in the present subsection. Consequently, for any sequence of maximum likelihood estimators $\widetilde{\eta}_n$ that are supported on the convex hull of the support of η_0, the process $\sqrt{n}(\widetilde{\eta}_n(t) - \eta_0(t))$ converges weakly in $l^\infty(c, d)$ for every interval $[c, d]$ that is strictly within the convex hull of the support of η_0.*

6.2. *Uniform scale.* Let $p(x \mid z) = (1/z)1\{[0,z]\}$ be the density of the uniform distribution on $[0,z]$. By a careful analysis of this special kernel, Vardi and Zhang (1992) show that the maximum likelihood estimator is consistent and its cumulative distribution function is asymptotically normal in $l^\infty(0,\infty)$. Our general method does not yield a result as strong as this, but it comes close under the condition that the true parameter has compact support.

Precisely, assume that η_0 has compact support in $(0,\infty)$ and no discrete component. Then it can be shown by our method that the maximum likelihood estimator is asymptotically normal in $l^\infty(0,d)$ for every d such that $\eta_0(d,\infty) > 0$. The method of proof is almost identical to the method used for the shifted uniforms in the previous subsection and will only be indicated briefly. We have

$$l_\eta h(x) = \frac{\int_{[x,\infty)} h(z) z^{-1}\, d\eta(z)}{\int_{[x,\infty)} z^{-1}\, d\eta(z)},$$

$$l^* g(z) = \frac{1}{z}\int_0^z g(x)\, dx.$$

In this model the denominator of $l_0 h(x)$ is ill-behaved as x increases to the upper endpoint of the support of η_0. This motivated the use of the norm

$$\|h\| = \|h\|_{\mathrm{BBV}} \vee \sup_{z_1 \neq z_2 > b-\varepsilon} \frac{|h(z_1) - h(z_2)|}{|z_1 - z_2|^{1/2}}.$$

Here $b - \varepsilon$ is strictly less than the upper endpoint of the support of η_0, so that the norm is dependent on the true underlying parameter. Let $[a,b]$ be the convex hull of the support of the true parameter η_0 and H be the set of all functions $h\colon [a,b] \to \mathbb{R}$ with $\|h\| < \infty$. In analogy with the previous subsection, $l_0\colon H \to l^\infty(\mathcal{X})$ is compact, provided η_0 has no discrete component. This verifies a key condition of Theorem 3.3. For completeness we formulate the corollary that can be drawn from this general theorem for the special example. However, as noted previously, Vardi and Zhang's (1992) result is stronger than ours.

THEOREM 6.3. *Let η_0 have compact support in $(0,\infty)$ and no discrete component. Then the condition of Theorem 3.3 are satisfied for the norm $\|\cdot\|$ as specified in the present subsection. Consequently, any sequence of maximum likelihood estimators $\widetilde{\eta}_n$ satisfies that $\sqrt{n}(\widetilde{\eta}_n(t) - \eta_0(t))$ converges weakly in $l^\infty(0,d)$ for every d with $\eta_0(0,d) < 1$.*

REFERENCES

BHANJA, J. and GHOSH, J. K. (1988). Efficient estimation with many nuisance parameters. Technical Report, Indian Statistical Inst., Calcutta.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins Univ. Press.

BICKEL, P. J. and RITOV, Y. (1993). Efficient estimation using both direct and indirect observations. *Theory Probab. Appl.* **38** 194–213.

DUDLEY, R. M. (1984). *A Course on Empirical Processes. École d'Éte de Probabilités de Saint-Flour XII. Lecture Notes in Math.* **1097** 1–142. Springer, Berlin.

DUDLEY, R. M. (1985). An extended Wichura theorem, definitions of Donsker classes, and weighted empirical processes. *Probability in Banach Spaces V. Lecture Notes in Math.* **1153** 1306–1326. Springer, New York.

GILL, R. D. (1989). Non- and semiparametric maximum likelihood estimation and the von Mises method. *Scand. J. Statist.* **16** 97–128.

GINÉ, E. and ZINN, J. (1986). *Lectures on the Central Limit Theorem for Empirical Processes. Lecture Notes in Math.* **1221** 50–113. Springer, Berlin.

GROENEBOOM, P. (1991). Nonparametric maximum likelihood estimators for interval censoring and deconvolution. Technical Report 91-53, Technische Univ. Delft.

HASMINSKII, R. Z. and IBRAGIMOV, I. A. (1983). On asymptotic efficiency in the presence of an infinite dimensional nuisance parameter. *Probability Theory and Mathematical Statistics. Lecture Notes in Math.* (K. Itô and J. V. Prohorov, eds.) **1021** 195–229. Springer, New York.

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.

HOFFMAN-JØRGENSEN, J. (1984). Stochastic processes on polish spaces. Unpublished manuscript.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 221–233. Univ. California Press, Berkeley.

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

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.

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

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

PAKES, A. and POLLARD, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica* **57** 1027–1057.

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

VAN DER VAART, A. W. (1988). Estimating a parameter in incidental and structural models by approximate maximum likelihood. Technical Report 139, Dept. Statistics, Univ. Washington.

VAN DER VAART, A. W. and WELLNER, J. A. (1993). Existence and consistency of maximum likelihood in upgraded mixture models. *J. Multivariate Anal.* **43** 133–147.

VAN DER VAART, A. (1993). New Donsker classes. Preprint.

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

VARDI, Y. and ZHANG, C.-H. (1992). Large sample study of empirical distributions in a random-multiplicative censoring model. *Ann. Statist.* **20** 1022–1040.

FACULTEIT DER WISKUNDE EN INFORMATICA
VRIJE UNIVERSITEIT
DE BOELELAAN 1081A
1081 HV AMSTERDAM
THE NETHERLANDS