ASYMPTOTIC NORMALITY IN MIXTURE MODELS

SARA VAN DE GEER

.

Abstract. We study the estimation of a linear function $\theta_0 = \int a dF_0$ of a distribution F_0, using i.i.d. observations of the mixture $p_{F_0} = \int k(\cdot, y) dF_0(y)$. Let $\hat{F}_n$ be the maximum likelihood estimator of F_0 and $\hat{\theta}_n = \int a d\hat{F}_n$. We examine the asymptotic distribution of $\hat{\theta}_n$. A problem here is that usually, $\hat{F}_n$ does not dominate F_0. Our main aim is to show that this can be overcome by considering the convex combination $\alpha \hat{F}_n + (1-\alpha) F_0$, with $\alpha < 1$.

1. Introduction

Let $X_1, \ldots, X_n$ be independent identically distributed random variables on $(\mathcal{X}, \mathcal{A})$, with distribution P. Suppose that for some σ-finite measure μ,

$$p = \frac{dP}{d\mu} \in \{p_F = \int k(\cdot, y) dF(y) : \ F \in \Lambda\},$$

where Λ is the class of all probability measures on a measurable space $(\mathcal{Y}, \mathcal{B})$, and where $k : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ is a given kernel with $\int k(x, y) d\mu(x) = 1$ for all $y \in \mathcal{Y}$. So

$$p = p_{F_0} = \int k(\cdot, y) dF_0(y),$$

for some $F_0 \in \Lambda$.

In this paper, we assume that F_0 is unknown, and that the maximum likelihood estimator $\hat{F}_n$ of F_0 exists. The latter is (not necessarily uniquely) defined by

$$\sum_{i=1}^n \log p_{\hat{F}_n}(X_i) = \max_{F \in \Lambda} \sum_{i=1}^n \log p_F(X_i).$$

Consider now functionals of the form

$$\theta(F) = \int a dF, \ F \in \Lambda,$$

with a a given function on $(\mathcal{Y}, \mathcal{B})$. Write $\theta_0 = \int a dF_0$ and $\hat{\theta}_n = \int a d\hat{F}_n$. We shall investigate the asymptotic behaviour (and efficiency) of the estimator of $\hat{\theta}_n$ of θ_0. An important issue in this context is the appropriate differentiability of $\theta(F)$. Let us briefly sketch the main idea.

ESAIM: Probability and Statistics is an electronic journal with URL address **http://www.emath.fr/ps**.

Received by the journal 8 March 1995. Revised 2 June 1995. Accepted for publication 28 September 1995

Consider a pair of random variables (X, Y) with values in $\mathcal{X} \times \mathcal{Y}$, let $k(\cdot, y)$ be the density of X given $Y = y$, and F_0 the distribution of Y. For a function $b : \mathcal{X} \to \mathbf{R}$, write

$$E(b(X)|Y = y) = A^* b(y).$$

Thus

$$A^* b(y) = \int k(x, y) b(x) d\mu(x).$$

Write for $h : \mathcal{Y} \to \mathbf{R}$,

$$E(h(Y)|X = x) = A_{F_0} h(x),$$

where

$$A_F h(x) = \frac{\int k(x, y) h(y) dF(y)}{p_F(x)}.$$

If for some b,

$$E(b(X)|Y = y) = a(y), \text{ for } F_0\text{−almost all } y, \tag{1.1}$$

then clearly

$$E(b(X)) = \theta_0.$$

So if (1.1) holds for some $b \in L_2(P)$ not depending on F_0, then $(1/n) \sum_{i=1}^n b(X_i)$ is a $\sqrt{n}$-consistent and asymptotically normal estimator of θ_0. We say that $\theta(F_0)$ is *differentiable* at F_0 if a solution of (1.1) exists. Note that if $k(\cdot, y)$ is a complete family for y in the support of F_0, there is at most one solution of (1.1). In general, there may also be several solutions, in which case we would like to take the one with the smallest variance. But such a solution possibly depends on F_0. The arguments below indicate that perhaps the maximum likelihood procedure automatically picks the best solution, with the estimator $\hat{F}_n$ plugged in for F_0.

We shall now discuss the solution with the smallest variance. First, we center the functions. Instead of a in (1.1), we consider the *gradient* of $\theta(F)$, which is defined as

$$\psi_F = a - \int a dF.$$

If for some $h_F \in L_\infty(F)$ with $\int h_F dF = 0$,

$$A^* A_F h_F = \psi_F, \ F\text{−a.s.}, \tag{1.2}$$

we call

$$b_F = A_F h_F \tag{1.3}$$

the efficient influence curve at $\theta(F)$. Note that b_F is also centered now: $\int b_F dP_F = 0$. It follows from Van der Vaart (1991) that $(1/n) \int b_{F_0}^2 dP$ is a lower bound for the asymptotic variance of an estimator of θ_0. He considers parametric submodels with Hilbert space structure to arrive at results of this type in a very general context. In our case, the parametric submodel would be indexed by $\Lambda_F = \{F_t : (dF_t)^{1/2} = (1+\frac{1}{2}th_F)(dF)^{1/2},\ |t| \text{ small}\}$. For this reason, we call h_F a *direction* (along which one can consider a submodel). The assumptions $h_F \in L_\infty(F)$ and $\int h_F dF = 0$ ensure that indeed $\Lambda_F \subset \Lambda$.

Suppose now that the efficient influence curves $b_{\hat{F}_n}$ and b_{F_0} exist. So $b_{\hat{F}_n} = A_{\hat{F}_n} h_{\hat{F}_n}$ for some $h_{\hat{F}_n} \in L_\infty(\hat{F}_n)$, and

$$A^* b_{\hat{F}_n} = \psi_{\hat{F}_n},\ \hat{F}_n\text{–a.s..} \tag{1.4}$$

Observe that $\hat{F}_n$ is an interior point of $\{\hat{F}_{n,t} : d\hat{F}_{n,t} = (1+th_{\hat{F}_n})d\hat{F}_n,\ |t| \text{ small}\}$. So we have

$$\frac{d}{dt}\sum_{i=1}^n \log p_{\hat{F}_{n,t}}(X_i)|_{t=0} = 0,$$

or

$$\sum_{i=1}^n b_{\hat{F}_n}(X_i) = 0.$$

Write this as

$$\int b_{\hat{F}_n} dP_n = 0,$$

with P_n the empirical distribution based on $X_1, \ldots, X_n$ (see (2.1)). Now, let us compare this with $\int b_{\hat{F}_n} dP$. Changing the order of integration gives

$$\int b_{\hat{F}_n} dP = \int A^* b_{\hat{F}_n} dF_0.$$

Moreover, if $\hat{F}_n$ dominates F_0, then (1.4) yields

$$\int A^* b_{\hat{F}_n} dF_0 = \int \psi_{\hat{F}_n} dF_0 = \theta_0 - \hat{\theta}_n.$$

So then we have the identity of van der Laan (1993, 1994):

$$\hat{\theta}_n - \theta_0 = \int b_{\hat{F}_n} d(P_n - P). \tag{1.5}$$

Finally, if $b_{\hat{F}_n}$ converges to b_{F_0} in an appropriate sense, one obtains asymptotic normality (and efficiency) of $\hat{\theta}_n$.

The problem is now that the assumption that $\hat{F}_n$ dominates F_0 is often not valid. Nevertheless, van der Laan (1993) presents some examples that show that the identity (1.5) can hold even when $\hat{F}_n$ does not dominate F_0. We shall however be concerned with the situation where the identity (1.5) is not necessarily true.

Our approach is to consider for each $0 \le \alpha < 1$ and $F \in \Lambda$, the convex combination

$$F_\alpha = \alpha F + (1-\alpha)F_0.$$

Then indeed, $\hat{F}_{n,\alpha}$ dominates F_0. We shall obtain asymptotic normality of $\hat{\theta}_n$ by choosing $\alpha = \hat{\alpha}_n$ in such a way that it tends to one at an appropriate speed.

The paper is organized as follows. In Section 2, we derive a linear approximation for $\hat{\theta}_n$. Asymptotic normality and efficiency follow from this. The conditions we need to arrive at the result are consistency of the maximum likelihood estimator (obtained by separate means), differentiability of $\theta(F)$ at $F = F_\alpha$, $0 \leq \alpha < 1$, bounded directions and suitable continuity conditions on the influence curves. A discussion of these conditions can be found in Section 3. Section 4 presents some examples.

2. Asymptotic normality

As pseudo-metric on Λ, we take the Hellinger distance between the mixtures:

$$d(F, \tilde{F}) = h(p_F, p_{\tilde{F}}) = (\frac{1}{2}\int(\sqrt{p_F} - \sqrt{p_{\tilde{F}}})^2 d\mu)^{1/2},\ \ F, \tilde{F} \in \Lambda.$$

Consistency of $\hat{F}_n$ in this metric holds in fairly general situations. It is closely related to the further assumptions as stated in Condition 1 (see Section 3.1 for more details).

We use the notation

$$P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}, \tag{2.1}$$

i.e. P_n is the empirical distribution that puts mass $(1/n)$ at each of the observations $X_1, \ldots, X_n$. Moreover, we shall make frequent use of stochastic order symbols. If $\{Z_n\}$ is a sequence of real-valued random variables, and $\{k_n\}$ a sequence of positive numbers, then we say that $Z_n = O_{\mathbf{P}}(k_n)$ if

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbf{P}(|Z_n| > k_n M) = 0.$$

Similarily, $Z_n = o_{\mathbf{P}}(k_n)$ means that for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}(|Z_n| > k_n\epsilon) = 0.$$

Condition 1. *(consistency and rates). The estimators $\hat{F}_n$ and $\hat{\theta}_n$ are consistent, i.e.*

$$d(\hat{F}_n, F_0) = o_{\mathbf{P}}(1), \tag{2.2}$$

and

$$|\hat{\theta}_n - \theta_0| = o_{\mathbf{P}}(1). \tag{2.3}$$

Moreover, for some $\delta_n^2 = o(n^{-1/2})$,

$$\int \log(\frac{2p_{\hat{F}_n}}{p_{\hat{F}_n} + p_{F_0}})dP_n = O_{\mathbf{P}}(\delta_n^2). \tag{2.4}$$

Condition 2 *(differentiability in a neighbourhood of F_0 and existence of efficient influence curves). For some $\epsilon > 0$, and for all $0 \leq \alpha < 1$ and $F \in \Lambda$ with $d(F, F_0) \leq \epsilon$, we have*

$$A^* b_{F_\alpha} = \psi_{F_\alpha},\ \ F_\alpha\text{–a.s.}, \tag{2.5}$$

for b_{F_α} satisfying

$$b_{F_\alpha}(x) = A_{F_\alpha} h_{F_\alpha}(x),\ p_{F_\alpha}(x) > 0, \tag{2.6}$$

for some $h_{F_\alpha} \in L_\infty(F_\alpha)$, with $\int h_{F_\alpha} dF_\alpha = 0$.

CONDITION 3. *(control on the direction h_{F_α}). For some $\epsilon > 0$ and $M < \infty$,*

$$\sup_{d(F,F_0)\leq\epsilon} \sup_{0\leq\alpha<1} \sup_{y\in\text{support}(F_\alpha)} |(1-\alpha)h_{F_\alpha}(y)| \leq M. \tag{2.7}$$

CONDITION 4. *(control on the efficient influence curves b_{F_α}). The information for estimating θ_0 is positive:*

$$\int b_{F_0}^2 dP > 0. \tag{2.8}$$

Moreover, the influence curves are uniformly bounded: for some $\epsilon > 0$,

$$\sup_{d(F,F_0)\leq\epsilon} \sup_{0\leq\alpha<1} \sup_{p_{F_\alpha}(x)>0} |b_{F_\alpha}(x)| < \infty. \tag{2.9}$$

Finally, for $0 < \epsilon_n \to 0$,

$$\sup_{d(F,F_0)\leq\epsilon_n} \sup_{0\leq\alpha<1} \int b_{F_\alpha}^2 dP_n = \int b_{F_0}^2 dP + o_{\mathbf{P}}(1), \tag{2.10}$$

and

$$\sup_{d(F,F_0)\leq\epsilon_n} \sup_{0\leq\alpha<1} \int (b_{F_\alpha} - b_{F_0}) d(P_n - P) = o_{\mathbf{P}}(n^{-1/2}). \tag{2.11}$$

THEOREM 2.1. *Assume that conditions 1-4 are met. Then*

$$\hat{\theta}_n - \theta_0 = \int b_{F_0} d(P_n - P) + o_{\mathbf{P}}(n^{-1/2}). \tag{2.12}$$

Proof. By (2.4), we can find a non-random increasing function $\gamma : (0,\infty) \to (0,\infty)$, satisfying

$$\gamma(x) \to 0 \text{ for } x \to 0, \tag{2.13}$$

$$\frac{x}{\gamma(x)} \leq \frac{1}{2} \text{ for all } x > 0, \tag{2.14}$$

and

$$\frac{x}{\gamma(x)} \to 0 \text{ for } x \to 0, \tag{2.15}$$

such that $\delta_n^2 = o(n^{-1/2})\gamma(n^{-1/2})$. This gives

$$\int \log\left(\frac{2p_{\hat{F}_n}}{p_{\hat{F}_n} + p_{F_0}}\right) dP_n = o_{\mathbf{P}}(n^{-1/2})\gamma(n^{-1/2}). \tag{2.16}$$

Choose

$$1-\hat{\alpha}_n=\frac{|\hat{\theta}_n-\theta_0|+n^{-1/2}}{\gamma(|\hat{\theta}_n-\theta_0|+n^{-1/2})}. \tag{2.17}$$

Then $\hat{\alpha}_n<1$ and by (2.14), $\hat{\alpha}_n\geq 1/2$. Since $|\hat{\theta}_n-\theta_0|=o_{\mathbf{P}}(1)$, (2.15) implies that $(1-\hat{\alpha}_n)=o_{\mathbf{P}}(1)$. Furthermore, (2.13) yields

$$\frac{|\hat{\theta}_n-\theta_0|}{1-\hat{\alpha}_n}=o_{\mathbf{P}}(1), \tag{2.18}$$

as well as

$$\frac{n^{-1/2}}{1-\hat{\alpha}_n}=o_{\mathbf{P}}(1). \tag{2.19}$$

Define

$$\tilde{F}_n=\hat{\alpha}_n\hat{F}_n+(1-\hat{\alpha}_n)F_0. \tag{2.20}$$

Let $\epsilon>0$ be small enough, so that (2.5), (2.6), (2.7) and (2.9) in conditions 2-4 are fulfilled for this value of ϵ, and let D_n be the set

$$D_n=\{d(\hat{F}_n,F_0)\leq\epsilon\}.$$

Because $\hat{\alpha}_n<1$, we can find on D_n an influence curve $b_{\tilde{F}_n}$ and a direction $h_{\tilde{F}_n}$, such that

$$A^*b_{\tilde{F}_n}=\psi_{\tilde{F}_n},$$

and

$$b_{\tilde{F}_n}(x)=A_{\tilde{F}_n}h_{\tilde{F}_n}(x),\ x\in\{p_{\tilde{F}_n}>0\}.$$

We can write

$$\int b_{\tilde{F}_n}\,dP_n\mathrm{l}\{D_n\}=\int b_{\tilde{F}_n}\,d(P_n-P)\mathrm{l}\{D_n\}+\int b_{\tilde{F}_n}\,dP\mathrm{l}\{D_n\}. \tag{2.21}$$

Because $d(\hat{F}_n,F_0)=o_{\mathbf{P}}(1)$, we have that $\mathrm{l}\{D_n\}=1+o_{\mathbf{P}}(1)$. So, using (2.11),

$$\int b_{\tilde{F}_n}d(P_n-P)\mathrm{l}\{D_n\}=\int b_{F_0}d(P_n-P)+o_{\mathbf{P}}(n^{-1/2})=O_{\mathbf{P}}(n^{-1/2}). \tag{2.22}$$

Since $\tilde{F}_n$ dominates F_0, and $\hat{\alpha}_n=1+o_{\mathbf{P}}(1)$,

$$\int b_{\tilde{F}_n}\,dP\mathrm{l}\{D_n\}=-\hat{\alpha}_n(\hat{\theta}_n-\theta_0)\mathrm{l}\{D_n\}=-(1+o_{\mathbf{P}}(1))(\hat{\theta}_n-\theta_0). \tag{2.23}$$

Insert (2.22) and (2.23) into (2.21) to get that

$$\int b_{\tilde{F}_n}\,dP_n\mathrm{l}\{D_n\}=\int b_{F_0}d(P_n-P)+o_{\mathbf{P}}(n^{-1/2})-(1+o_{\mathbf{P}}(1))(\hat{\theta}_n-\theta_0). \tag{2.24}$$

Let

$$\hat{t}_n=\frac{\int b_{\tilde{F}_n}\,dP_n}{(1-\hat{\alpha}_n)\int b_{F_0}^2\,dP}\mathrm{l}\{D_n\}. \tag{2.25}$$

Equality (2.24), together with (2.8), (2.18) and (2.19) imply that

$$\hat{t}_n = o_{\mathbf{P}}(1). \tag{2.26}$$

Take M as in (2.7) and $E_n = D_n \cap \{|\hat{t}_n| < 1/M\}$. Define on $\{E_n\}$,

$$d\tilde{F}_n(\hat{t}_n) = (1 + \hat{t}_n(1-\hat{\alpha}_n)h_{\tilde{F}_n})d\tilde{F}_n.$$

Then on $\{E_n\}$,

$$\tilde{F}_n(\hat{t}_n) \in \Lambda. \tag{2.27}$$

Moreover, from (2.26) we know that $\mathrm{l}\{E_n\} = 1 + o_{\mathbf{P}}(1)$, so that by (2.9) and (2.10),

$$\int \log(\frac{p_{\tilde{F}_n(\hat{t}_n)}}{p_{\tilde{F}_n}})dP_n\mathrm{l}\{E_n\} = \int \log(1 + \hat{t}_n(1-\hat{\alpha}_n)b_{\tilde{F}_n})dP_n\mathrm{l}\{E_n\} \tag{2.28}$$

$$= \hat{t}_n(1-\hat{\alpha}_n)\int b_{\tilde{F}_n}dP_n\mathrm{l}\{E_n\} - \frac{1}{2}\hat{t}_n^2(1-\hat{\alpha}_n)^2\int b_{\tilde{F}_n}^2 dP_n\mathrm{l}\{E_n\}(1+o_{\mathbf{P}}(1))$$

$$= \frac{1}{2}\frac{(\int b_{\tilde{F}_n}dP_n)^2}{\int b_{F_0}^2 dP}\mathrm{l}\{E_n\}(1+o_{\mathbf{P}}(1)).$$

Since $\hat{F}_n$ maximizes the likelihood, and (2.27) holds on E_n, we have

$$\int \log p_{\hat{F}_n}dP_n\mathrm{l}\{E_n\} \geq \int \log p_{\tilde{F}_n(\hat{t}_n)}dP_n\mathrm{l}\{E_n\}. \tag{2.29}$$

Furthermore, the concavity of the log-function and the fact that $\hat{\alpha}_n \geq 1/2$ yield

$$\int \log p_{\tilde{F}_n}dP_n \geq (2\hat{\alpha}_n - 1)\int \log p_{\hat{F}_n}dP_n + 2(1-\hat{\alpha}_n)\int \log(\frac{p_{\hat{F}_n}+p_{F_0}}{2})dP_n. \tag{2.30}$$

Combine (2.28), (2.29) and (2.30) to find

$$2(1-\hat{\alpha}_n)\int \log(\frac{2\hat{p}_{\hat{F}_n}}{p_{\hat{F}_n}+p_{F_0}})dP_n \geq \frac{1}{2}\frac{(\int b_{\tilde{F}_n}dP_n)^2}{\int b_{F_0}^2 dP}\mathrm{l}\{E_n\}(1+o_{\mathbf{P}}(1)). \tag{2.31}$$

From (2.16) and (2.17), we know that the left-hand side of this equality is

$$(1-\hat{\alpha}_n)o_{\mathbf{P}}(n^{-1/2})\gamma(n^{-1/2}) = (|\hat{\theta}-\theta_0|+n^{-1/2})\frac{\gamma(n^{-1/2})}{\gamma(|\hat{\theta}_n-\theta_0|+n^{-1/2})}o_{\mathbf{P}}(n^{-1/2})$$

$$= (|\hat{\theta}_n-\theta_0|+n^{-1/2})o_{\mathbf{P}}(n^{-1/2}),$$

where in the last step, we used that γ is increasing. In view of (2.24), the right-hand side of (2.31) is of the form

$$(O_{\mathbf{P}}(n^{-1/2}) - (1+o_{\mathbf{P}}(1))(\hat{\theta}_n-\theta_0))^2(\frac{1+o_{\mathbf{P}}(1)}{\int b_{F_0}^2 dP}).$$

So we find from (2.31) that

$$|\hat{\theta}_n - \theta_0|^2 \leq \max\{O_{\mathbf{P}}(n^{-1}), (|\hat{\theta}_n - \theta_0| + n^{-1/2}) o_{\mathbf{P}}(n^{-1/2})\},$$

which implies $|\hat{\theta}_n - \theta_0| = O_{\mathbf{P}}(n^{-1/2})$. But then, the left-hand side of (2.31) is $o_{\mathbf{P}}(n^{-1})$, so that it reads

$$(\int b_{F_0} d(P_n - P) + o_{\mathbf{P}}(n^{-1/2}) - (1 + o_{\mathbf{P}}(1))(\hat{\theta}_n - \theta_0))^2 (1 + o_{\mathbf{P}}(1)) = o_{\mathbf{P}}(n^{-1}).$$

In other words,

$$\hat{\theta}_n - \theta_0 = \int b_{F_0} d(P_n - P) + o_{\mathbf{P}}(n^{-1/2}).$$

□

3. Comments on conditions 1-4

Conditions 1 and 4 can be verified using the concept of *entropy*. Therefore, we introduce the following definitions. Let Q be a probability measure on $(\mathcal{X}, \mathcal{A})$ and $\mathcal{G} \subset \mathcal{L}_q(Q)$, $q \geq 1$.

Definition 1 The δ-covering number $N_q(\delta, \mathcal{G}, Q)$ of $\mathcal{G}$ is defined as the number of balls with radius δ necessary to cover $\mathcal{G}$. More precisely, let $\{g_j\}_{j=1}^m$ be such that for each $g \in \mathcal{G}$ there is a $j \in \{1, \ldots, m\}$ such that

$$\int (g - g_j)^q dQ \leq \delta^q. \tag{3.1}$$

Then $N_q(\delta, \mathcal{G}, Q)$ is the smallest m for which such a collection $\{g_j\}_{j=1}^m$ exists. The δ-entropy of $\mathcal{G}$ is $H_q(\delta, \mathcal{G}, Q) = \log N_q(\delta, \mathcal{G}, Q) \vee 1$.

Definition 2. Let $\{[g_j^L, g_j^U]\}_{j=1}^m$ be such that for all $g \in \mathcal{G}$ there is a $j \in \{1 \ldots, m\}$ such that

$$g_j^L \leq g \leq g_j^U, \tag{3.2}$$

and

$$\int (g_j^U - g_j^L)^q dQ \leq \delta^q. \tag{3.3}$$

Write $N_q^B(\delta, \mathcal{G}, Q)$ for the smallest m for which such a collection $\{[g_j^L, g_j^U]\}_{j=1}^m$ exists. Then $H_q^B(\delta, \mathcal{G}, Q) = \log N_q^B(\delta, \mathcal{G}, Q) \vee 1$ is called the δ-entropy with bracketing.

3.1. On Condition 1

Suppose that $\mathcal{Y}$ is a locally compact Hausdorff space with countable base, and that $\mathcal{B}$ is the Borel σ-algebra. Let $\mathcal{C}_0$ be the class of all functions $c : \mathcal{Y} \to \mathbf{R}$ that vanish at infinity. If $k(x, \cdot) \in \mathcal{C}_0$ for μ-almost all $x \in \mathcal{X}$, then $d(\hat{F}_n, F_0) \to 0$ almost surely (see Pfanzagl (1988)).

Denote the class of all measures F on $(\mathcal{Y}, \mathcal{B})$ with $F(\mathcal{Y}) \leq 1$ by Λ^*. The *vague* topology on Λ^* is the smallest topology such that $F \mapsto \int c dF$ is continuous for every $c \in \mathcal{C}_0$. Let τ be the metric corresponding to the vague topology. We say that F_0 is *identifiable* (for the metric τ) if for all $F \in \Lambda^*$, $d(F, F_0) = 0$ implies $\tau(F, F_0) = 0$. If F_0 is identifiable, then $d(\hat{F}_n, F_0) \to 0$ almost surely implies $\tau(\hat{F}_n, F_0) \to 0$ almost surely. So in particular, then $|\hat{\theta}_n - \theta_0| \to 0$ almost surely, whenever $a \in \mathcal{C}_0$. More details can be found in e.g. Pfanzagl (1988) or Van de Geer (1993a).

Let us now investigate the rate of convergence for the log-likelihood ratio. Note first of all that

$$0 \leq \int \log(\frac{2p_{\hat{F}_n}}{p_{\hat{F}_n} + p_{F_0}}) dP_n \leq \frac{1}{2} \int \log(\frac{p_{\hat{F}_n}}{p_{F_0}}) dP_n,$$

so that nothing is lost (and indeed something could be gained) by comparing $p_{\hat{F}_n}$ with the convex combination $(p_{\hat{F}_n} + p_{F_0})/2$ instead of with p_{F_0}.

LEMMA 3.1. *Suppose that*

$$\int_0^1 \sqrt{H_2^B(\delta, \mathcal{G}, P)} d\delta < \infty, \tag{3.4}$$

where

$$\mathcal{G} = \{\sqrt{\frac{p_F + p_{F_0}}{p_{F_0}}} : \ F \in \Lambda\}.$$

Then

$$\int \log(\frac{p_{\hat{F}_n}}{p_{F_0}}) dP_n = O_{\mathbf{P}}(\delta_n^2), \text{ with } \delta_n^2 = o(n^{-1/2}).$$

Proof. See Van de Geer (1995). A slight modification can be found in Wong and Shen (1992). □

We call the left-hand side of (3.4) the entropy integral. If the entropy integral diverges, suboptimal rates can emerge (see Birgé and Massart (1993)).

Lemma 3.2 below makes use of the special structure of the mixing model. We need the following notation: for $\sigma \geq 0$,

$$\tau_1^2(\sigma) = \int_{p_{F_0} \leq \sigma} p_{F_0} d\mu,$$

and

$$\tau_2^2(\sigma) = \int_{p_{F_0} > \sigma} \frac{1}{p_{F_0}} d\mu.$$

LEMMA 3.2. *Let* $\mathcal{K} = \{k(\cdot, y) : \ y \in \mathcal{Y}\}$. *Suppose that the functions in* $\mathcal{K}$ *are uniformly bounded, and that for all probability measures* Q *with finite support,*

$$N_2(\delta, \mathcal{K}, Q) \leq A\delta^{-w}, \ \delta > 0, \tag{3.5}$$

where the constants A and w do not depend on Q. Let $0 \leq \sigma_n \to 0$, $\delta_n \geq \tau_1(\sigma_n) \vee n^{-(2+w)/(4+4w)}(\tau_2(\sigma_n))^{w/(2+2w)}$. Then

$$\int \log(\frac{2p_{\hat{F}_n}}{p_{\hat{F}_n} + p_{F_0}})dP_n = O_{\mathbf{P}}(\delta_n^2). \tag{3.6}$$

Proof See Van de Geer (1993b). □

Lemma 3.2 may not yield the optimal rate, but all we need here is a rate faster than $n^{-1/2}$. This is the case if $\tau_2(\sigma_n) = o(n^{\frac{1}{2w}})$. To put it differently, if we define

$$\tau_1^{-1}(\delta) = \sup\{\sigma : \ \tau_1(\sigma) \leq \delta^2\},$$

then $\delta_n^2 = o(n^{-1/2})$ if

$$\tau_2(\tau_1^{-1}(\delta)) = o(\delta^{-\frac{2}{w}}) \text{ for } \delta \downarrow 0.$$

3.2. On Condition 2

It is natural to require (2.5) and (2.6) for $\alpha = 0$. The fact that we need these equalities for all $0 \leq \alpha < 1$ is closely related to being able to estimate the efficient influence curve. However, it should be noted that we only assume b_F to exist for certain F that dominate F_0.

In some applications, $b_{\hat{F}_n}$ does exist, but this usually will not help to simplify the proofs (see Section 4 for an example).

3.3. On Condition 3

Here, we assume that h_{F_α} behaves like $1/(1-\alpha)$. Clearly, this allows misbehaviour for $\alpha \to 1$, but it also requires h_{F_α} to be bounded. In some applications, this reduces to assuming that h_{F_0} is bounded. The proof of Theorem 2.1 reveals that we need that for all t sufficiently small, $dF_\alpha(t)/dF_\alpha = 1+t(1-\alpha)h_{F_\alpha}$ exists and is non-negative, i.e., that $F_\alpha(t) \in \Lambda$. If for some function g, dg/dF_0 exists and is bounded, say by C, then also dg/dF_α exists and $(1-\alpha)dg/dF_\alpha$ is bounded by C. This is the reason why in Example 4.1 and 4.3 our approach works. But it fails in Example 4.2!

3.4. On Condition 4

Let us present a brief overview of some results from empirical process theory, that can be applied in this context. We cite them from Pollard (1984) and Ossiander (1987). Throughout, we assume that the necessary measurability conditions are satisfied.

Consider a class $\mathcal{G}$ of functions on $(\mathcal{X}, \mathcal{A})$, with envelope

$$G = \sup_{g \in \mathcal{G}} |g|.$$

The class $\mathcal{G}$ is called a *Glivenko-Cantelli* class if

$$\sup_{g \in \mathcal{G}} |\int g d(P_n - P)| \to 0, \text{ almost surely.}$$

It is called a *Donsker* class if $\sqrt{n}\int gd(P_n - P)$ converges in distribution to a mean zero Gaussian process on $\mathcal{G}$. This limiting process is assumed to have continuous sample paths with respect to $\rho(\cdot,\cdot)$, with

$$\rho(g_1, g_2)^2 = \int (g_1 - g_2)^2 dP - (\int (g_1 - g_2) dP)^2.$$

Necessary and sufficient conditions for $\mathcal{G}$ to be a Glivenko-Cantelli class are:

$$G \in L_1(P),$$

and

$$H_1(\delta, \mathcal{G}, P_n) = o_{\mathbf{P}}(n),\ \ \delta > 0.$$

If $\mathcal{G}$ is a Donsker class, then the *asymptotic equicontinuity* condition holds, i.e. for all $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\limsup_{n\to\infty} \mathbf{P}(\sup_{g\in\mathcal{G}_\delta} |\sqrt{n}\int g_\delta d(P_n - P)| > \epsilon) < \epsilon, \tag{3.7}$$

with $\mathcal{G}_\delta = \{(g_1 - g_2) :\ g_1, g_2 \in \mathcal{G},\ \rho(g_1, g_2) \leq \delta\}$. Sufficient conditions for $\mathcal{G}$ to be a Donsker class are:

$$G \in L_2(P),$$

and

$$\int_0^1 \sqrt{H_2^B(\delta, \mathcal{G}, P)}\, d\delta < \infty. \tag{3.8}$$

Condition (3.8) may be replaced by

$$\int_0^1 \sqrt{\lambda(\delta, \mathcal{G})}\, d\delta < \infty, \tag{3.9}$$

where

$$\lambda(\delta, \mathcal{G}) \geq \sup_Q H_2(\delta(\int G^2 dQ)^{1/2}, \mathcal{G}, Q), \tag{3.10}$$

and where the supremum is taken over all probability measures Q with finite support.

Suppose now that (2.9) is met, and that for $0 < \epsilon_n \to 0$,

$$\sup_{d(F,F_0)\leq\epsilon_n} \sup_{0\leq\alpha<1} \int (b_{F_\alpha} - b_{F_0})^2 dP \to 0. \tag{3.11}$$

Then it is clear from the above that (2.10) and (2.11) are fulfilled if one of the entropy conditions (3.8) or (3.9) holds, with

$$\mathcal{G} = \{b_{F_\alpha} :\ d(F, F_0) \leq \epsilon,\ 0 \leq \alpha < 1\}.$$

In many applications however, no explicit expressions are available for the efficient influence curves, so that it may still be difficult to check the entropy conditions.

4. Examples

Example 4.1: interval censored obervations, case I

Suppose one observes $X_i = (T_i, \beta_i)$, where $\beta_i = \mathrm{l}\{Y_i \le T_i\}$, Y_i and T_i are independent random variables, both with values in a bounded interval, say $(0,1]$, and where T_i has (unknown) distribution G, $i = 1, \ldots, n$. This is one of the models studied in Groeneboom and Wellner (1992). The density of X_i with respect to $G \times \nu$, ν being the counting measure on $\{0,1\}$, is

$$p_{F_0}(t,\beta) = \beta F_0(t) + (1-\beta)(1-F_0(t)) = \int k(t,\beta,y)dF_0(y),$$

with $k(t,\beta,y) = \beta \mathrm{l}\{y \le t\} + (1-\beta)\mathrm{l}\{y > t\}$.

Lemma 4.1. *Suppose G has density g with respect to Lebesgue measure. Assume that $\dot{a}(y) = da(y)/dy$ exists and*

$$|\frac{\dot{a}(t)}{g(t)}| \le C_1 < \infty, \ t \in (0,1], \tag{4.1}$$

and

$$|\frac{d(\dot{a}(t)/g(t))}{dF_0(t)}| \le C_2 < \infty, \ t \in (0,1]. \tag{4.2}$$

Then

$$\hat{\theta}_n - \theta_0 = \int b_{F_0} d(P_n - P) + o_{\mathbf{P}}(n^{-1/2}),$$

where

$$b_{F_0}(t,\beta) = -\beta \frac{(1-F_0(t))\dot{a}(t)}{g(t)} + (1-\beta)\frac{F_0(t)\dot{a}(t)}{g(t)}, \ t \in (0,1], \ \beta \in \{0,1\}.$$

Proof Condition 1 is met: $d(\hat{F}_n, F_0) \to 0$ almost surely, $|\hat{\theta}_n - \theta_0| \to 0$ almost surely and

$$\int \log(\frac{p_{\hat{F}_n}}{p_{F_0}})dP_n = O_{\mathbf{P}}(n^{-2/3}).$$

This follows from the theory in Subsection 3.1 (see also Groeneboom and Wellner (1992) and Van de Geer (1993a)). Define

$$b_{F_\alpha}(t,\beta) = -\beta \frac{(1-F_\alpha(t))\dot{a}(t)}{g(t)} + (1-\beta)\frac{F_\alpha(t)\dot{a}(t)}{g(t)}, \ t \in (0,1], \ \beta \in \{0,1\}.$$

Then

$$A^* b_{F_\alpha}(y) = -\int_{y-}^1 (1-F_\alpha(t))\dot{a}(t)dt + \int_0^{y-} F_\alpha(t)\dot{a}(t)dt$$

$$= a(y) - a(0) - \int_0^1 (1-F_\alpha(t-))\dot{a}(t)dt = \psi_{F_\alpha}(y), \ y \in (0,1].$$

Because F_α dominates F_0, dF_0/dF_α exists. So

$$h_{F_\alpha}(y) = -\frac{d(F_\alpha(y)(1-F_\alpha(y))\dot{a}(y)/g(y))}{dF_\alpha(y)}$$

$$= \frac{(2F_\alpha(y)-1)\dot{a}(y)}{g(y)} + \frac{F_\alpha(y-)(1-F_\alpha(y-))}{g^2(y)} \frac{d(\dot{a}(y)/g(y))}{dF_0(y)} \frac{dF_0(y)}{dF_\alpha(y)}$$

exists too. Moreover, for $p_{F_\alpha}(t,\beta) > 0$,

$$A_{F_\alpha} h_{F_\alpha}(t,\beta) = \beta \frac{\int_0^t h_{F_\alpha} dF_\alpha}{F_\alpha(t)} + (1-\beta) \frac{\int_t^1 h_{F_\alpha} dF_\alpha}{1-F_\alpha(t)} = b_{F_\alpha}(t,\beta).$$

Thus, Condition 2 is fulfilled.

Clearly,

$$|h_{F_\alpha}(y)| \leq C_1 + \frac{C_2}{(1-\alpha)}.$$

Therefore, Condition 3 holds too.

Finally, we shall verify Condition 4. Note first that

$$\int b_{F_0}^2 dP = \int \frac{F_0(t)(1-F_0(t))(\dot{a}(t))^2}{g(t)} dt > 0,$$

and

$$|b_{F_\alpha}(t,\beta)| \leq C_1.$$

To check (2.10) and (2.11), we use the theory of Subsection 3.4. We have

$$d^2(F,F_0) = \frac{1}{2}\Big(\int (\sqrt{F} - \sqrt{F_0})^2 dG + \int (\sqrt{1-F} - \sqrt{1-F_0})^2 dG\Big),$$

and

$$\int (b_{F_\alpha} - b_{F_0})^2 dP = \int \frac{(F_\alpha(t) - F_0(t))^2 (\dot{a}(t))^2}{g(t)} dt \leq C_1^2 d^2(F_\alpha, F_0).$$

So indeed, (3.11) is satisfied: for $0 < \epsilon_n \to 0$,

$$\sup_{d(F,F_0)\leq\epsilon_n} \sup_{0\leq\alpha<1} \int (b_{F_\alpha} - b_{F_0})^2 dP \to 0.$$

Also (3.8) holds, namely for $\mathcal{G} = \{b_{F_\alpha} : \ d(F,F_0) \leq \epsilon, \ 0 \leq \alpha < 1\}$, we have for some constant A,

$$H_2^B(\delta, \mathcal{G}, P) \leq A\frac{1}{\delta}, \ \delta > 0.$$

This follows from entropy calculations for monotone functions (see Birman and Solomjak (1967) and, for the extension to entropy with bracketing, Van de Geer (1991)). Therefore, (2.10) and (2.11) are fulfilled. In view of Theorem 2.1, this completes the proof. □

The result of Lemma 4.1 was established earlier in Groeneboom and Wellner (1992). They use specific properties of the maximum likelihood estimator $\hat{F}_n$, such as the distance between successive jumps of $\hat{F}_n$ being smaller than $n^{-1/3}\log n$ with large probability. In this sense, local properties of $\hat{F}_n$ had to be obtained first, before one could arrive at the asymptotic behaviour of such global quantities as the mean of the maximum likelihood estimator. It inspired us to develop an alternative proof, which is hopefully applicable in more general situations.

The model for interval censored observations gives a good insight into the difficulties that arise due to the fact that $\hat{F}_n$ does not dominate F_0. Let us have a closer look for the case $a(y) = y$. It is known that $\hat{F}_n$ has finite support, say $\hat{z}_1 < \ldots < \hat{z}_m$. Define

$$\hat{g}(t) = \frac{G(\hat{z}_j) - G(\hat{z}_{j-1})}{\hat{z}_j - \hat{z}_{j-1}},\ t \in (\hat{z}_{j-1}, \hat{z}_j],$$

and

$$b_{\hat{F}_n}(t,\beta) = -\beta\frac{1-\hat{F}_n(t)}{\hat{g}(t)} + \beta\frac{\hat{F}_n(t)}{\hat{g}(t)}.$$

Then for $y \in \{\hat{z}_1, \ldots, \hat{z}_m\}$,

$$A^* b_{\hat{F}_n}(y) = y - \hat{\theta}_n = \psi_{\hat{F}_n}(y).$$

However,

$$\int \psi_{\hat{F}_n} dF_0 = -(\hat{\theta}_n - \tilde{\theta}_n),$$

with

$$\tilde{\theta}_n = \int\int^y \frac{1}{\hat{g}(t)} dG(t) dF_0(y).$$

In general, $\tilde{\theta}_n \neq \theta_0$, unless G happens to be the uniform distribution on $(0,1]$.

Write

$$h_{\hat{F}_n}(y) = -\frac{d(\hat{F}_n(y)(1-\hat{F}_n(y))/\hat{g}(y))}{d\hat{F}_n(y)}$$

$$= \frac{2\hat{F}_n(y) - 1}{\hat{g}(y)} + \frac{\hat{F}_n(y-)(1-\hat{F}_n(y-))}{\hat{g}(y)\hat{g}(y-)}\frac{d\hat{g}(y)}{d\hat{F}_n(y)}.$$

In order to have that at $\hat{F}_n$, the derivative of the likelihood equals zero in the direction $h_{\hat{F}_n}$, one must have that this direction is bounded. This leads to showing that the jumps of $\hat{F}_n$ are large enough.

Example 4.2: interval censored observations: case II

Let $X_i = (T_i, U_i, \beta_i, \gamma_i)$, with $\beta_i = 1\{Y_i \le T_i\}$, $\gamma_i = 1\{T_i < Y_i \le U_i\}$ and Y_i independent of (T_i, U_i), $i = 1, \ldots, n$. We assume bounded support, say T_i, U_i and $Y_i \in (0,2]$, and that $T_i < U_i$. This model is also studied in Groeneboom and Wellner (1992). The rate of convergence of the log-likelihood ratio can be found in Van de Geer (1993b). In general, no explicit expression for the influence curve can be given. However, it can be shown that under fairly mild conditions, $b_{\hat{F}_n}$ and $h_{\hat{F}_n}$ exist and are bounded. One of the conditions here is that T_i and U_i can be arbitrary close to each other. We shall now consider a situation where this is not true, but where explicit expressions are available. Namely, we suppose that $U_i = T_i + 1$. Let G (F_0) be the distribution of T_i (Y_i), with density g (f_0) w.r.t. Lebesgue measure.

Suppose that

$$0 < 1/C_0 \le f_0(y) \le C_0 < \infty,\ y \in (0,2], \tag{4.3}$$

$$\frac{|\dot{a}(t+1)| \vee |\dot{a}(t)|}{g(t)} \le C_1 < \infty,\ t \in (0,1], \tag{4.4}$$

and

$$|\frac{d(\dot{a}(t+1)/g(t))}{dF_0(t+1)}| \vee |\frac{d(\dot{a}(t)/g(t))}{dF_0(t)}| \le C_2 < \infty,\ t \in (0,1]. \tag{4.5}$$

Then

$$b_{F_0}(t,\beta,\gamma) = \beta\frac{H_{F_0}(t)}{F_0(t)} + \gamma\frac{H_{F_0}(t+1) - H_{F_0}(t)}{F_0(t+1) - F_0(t)} \tag{4.6}$$

$$-(1-\beta-\gamma)\frac{H_{F_0(t+1)}}{1 - F_0(t+1)},$$

and

$$H_{F_0}(y) = \begin{cases} F_0(y)(F_0(y) + F_0(y+1) - 2)\frac{\dot{a}(y)}{g(y)}, & 0 < y \le 1, \\ (F_0(y-1) + F_0(y))(1 - F_0(y))\frac{\dot{a}(y)}{g(y-1)}, & 1 < y \le 2, \end{cases} \tag{4.7}$$

The directions are $h_{F_\alpha} = dH_{F_\alpha}/dF_\alpha$, with H_{F_α} of the form (4.7), with F_0 replaced by F_α. However, condition 3 is not met, because $dF(y+1)/dF_0(y)$ and $dF(y)/dF_0(y+1)$, $0 < y \le 1$, can be arbitrary large. It is not clear to us whether the maximum likelihood estimator will be efficient. So this example shows that Theorem 2.1 certainly does not always provide an answer.

Example 4.3: convolution.

Let $X_i = Y_i + Z_i$, with Y_i and Z_i independent, Y_i has unknown distribution F_0 and Z_i has known distribution K, $i = 1, \ldots, n$. Suppose K (F_0) has density k (f_0) w.r.t. Lebesgue measure and that K and F_0 have support in $(0, 1]$. We consider the special case

$$k(z) = c_1^z/c_2,\ 0 < z \le 1,\ c_2 = \frac{\log c_1}{c_1 - 1},$$

where $c_1 \ge 1$ is fixed.

Lemma 4.3. *Suppose*

$$0 < 1/C_0 \le f_0(y) \le C_0 < \infty,\ 0 < y \le 1, \tag{4.8}$$

and

$$|d\dot{a}(y)/dy| \leq C_1 < \infty. \tag{4.9}$$

Then

$$\hat{\theta}_n - \theta_0 = \int b_{F_0} d(P_n - P) + o_{\mathbf{P}}(n^{-1/2}),$$

where

$$b_{F_0}(x) = \begin{cases} \frac{\Phi_{F_0}(x)}{L_{F_0}(x)}, & 0 < x \leq 1, \\ \frac{\Phi_{F_0}(1) - \Phi_{F_0}(x-1)}{L_{F_0}(1) - L_{F_0}(x-1)}, & 1 < x \leq 2 \;, \end{cases}$$

$$\Phi_{F_0}(y) = \frac{c_1 L_{F_0}(y)\Phi_{F_0}(1) - (\dot{a}(y) + \log c_1(a(y) - \theta_0))c_2 L_{F_0}(y)(L_{F_0}(1) - L_{F_0}(y))}{L_{F_0}(1) + (c_1 - 1)L_{F_0}(y)},$$

and

$$L_{F_0}(y) = \int_0^y c_1^{-u} dF_0(u).$$

Proof Let us only check (2.4) of Condition 1. The other conditions can be verified in the same way as in Lemma 4.1. If $c_1 = 1$, then (2.4) also follows from the same arguments as in Lemma 4.1, and we find

$$\int \log \frac{p_{\hat{F}_n}}{p_{F_0}} dP_n = O_{\mathbf{P}}(n^{-2/3}).$$

For $c_1 > 1$, the class $\mathcal{K} = \{k(\cdot, y) = k(\cdot - y), \; y \in (0, 1]\}$ satisfies for some A,

$$N(\delta, \mathcal{K}, Q) \leq A\delta^{-1},$$

for all $\delta > 0$ and all probability measures Q. Moreover, $\mathcal{K}$ is uniformly bounded. Now,

$$P_{F_0}(x) = \begin{cases} \frac{c_1^x}{c_2} \int_0^x c_1^{-y} dF_0(y), & 0 < x \leq 1, \\ \frac{c_1^x}{c_2} \int_{x-1}^1 c_1^{-y} dF_0(y), & 1 < x \leq 2. \end{cases}$$

Using (4.8), one sees that in Lemma 3.2, one can take $\tau_1^2(\sigma) \asymp \sigma$ and $\tau_2^2(\sigma) \asymp \log(1/\sigma)$ for $\sigma > 0$ small. So, inserting $w = 1$ in this lemma, one obtains

$$\int \log(\frac{2p_{\hat{F}_n}}{p_{\hat{F}_n} + p_{F_0}}) dP_n = O_{\mathbf{P}}(n^{-3/4}(\log n)^{1/4}).$$

□

References

Birgé, L. and Massart, P., (1993), Rates of convergence for minimum contrast estimators, *Probab. Th. Relat. Fields* **97** 113-150.

Birman, M. and Solomjak, M.J., (1967), Piece-wise polynomial approximations of functions of the classes W_p^α, *Mat. Sbornik* **73** 295-317.

Groeneboom, P. and Wellner, J.A., (1992), *Information Bounds and Nonparametric Maximum Likelihood Estimation*, BMW Seminar, Band 19, Birkhäuser.

OSSIANDER, M. (1987), A central limit theorem under metric entropy with L_2 bracketing, *Ann. Probab.* **15** 897-919.

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

POLLARD, D. (1984), *Convergence of Stochastic Processes*, Springer, New York.

VAN DE GEER, S. (1990), Estimating a regression function, *Ann. Statist.* **18** 907-924.

VAN DE GEER, S. (1991), *The entropy bound for monotone functions*, Techn. Report TW 91-10, Univ. of Leiden.

VAN DE GEER, S. (1993a), Hellinger consistency of certain nonparametric maximum likelihood estimators, *Ann. Statist.* **21** 14-44.

VAN DE GEER, S. (1993b), *Rates of convergence for the maximum likelihood estimator in mixture models*, Techn. Report TW 93-09, Univ. of Leiden.

VAN DE GEER, S. (1995), The method of sieves and minimum contrast estimators, *Mathematical Methods of Statistics* **4** 20-38.

VAN DER LAAN, M.J. (1993), *Efficient and Inefficient Estimation in Semiparametric Models* Thesis, Univ. of Utrecht. To appear as CWI tract **44**, Centre for Math. and Comp. Sci., Amsterdam.

VAN DER LAAN, M.J. (1994), *Proving efficiency of NPMLE and identities*, Techn. Report 44, Un. of California, Berkeley.

VAN DER VAART, A.W. (1991), On differentiable functionals, *Ann. Statist.* **19** 178-205.

WONG, W.H. and SHEN, X. (1992), *Probability inequalities for likelihood ratios and convergence rates for sieve MLE's*, Techn. Report 346, Univ. of Chicago.

MATHEMATICS DEPARTMENT, LEIDEN UNIVERSITY, NETHERLANDS