*The Annals of Statistics*
1996, Vol. 24, No. 2, 862–878

# EFFICIENT MAXIMUM LIKELIHOOD ESTIMATION IN SEMIPARAMETRIC MIXTURE MODELS

BY AAD VAN DER VAART

*Vrije Universiteit Amsterdam*

We consider maximum likelihood estimation in several examples of semiparametric mixture models, including the exponential frailty model and the errors-in-variables model. The observations consist of a sample of size $n$ from the mixture density $\int p_\theta(x|z)\, d\eta(z)$. The mixing distribution is completely unknown. We show that the first component $\hat{\theta}_n$ of the joint maximum likelihood estimator $(\hat{\theta}_n, \hat{\eta}_n)$ is asymptotically normal and asymptotically efficient in the semiparametric sense.

**1. Introduction.** In a semiparametric mixture model one observes a sample $X_1, \dots, X_n$ from a density of the type

$$p_{\theta,\eta}(x) = \int p_\theta(s|z)\, d\eta(z).$$

Here the *mixing distribution* $\eta$ is a completely unknown probability distribution on a measurable space $(\mathscr{Z}, \mathscr{C})$ and the *kernel* or mixture density $x \to p_\theta(x|z)$ is a family of probability densities with respect to a measure $\mu$ on a measurable space $(\mathscr{X}, \mathscr{A})$, which is known up to a parameter $\theta$ ranging over an open subset $\Theta$ of Euclidean space.

The maximum likelihood estimator $(\hat{\theta}_n, \hat{\eta}_n)$ maximizes the likelihood function

$$(\theta, \eta) \to \text{lik}(\theta, \nu) = \prod_{i=1}^{n} p_{\theta,\eta}(X_i).$$

Using Wald's approach, Kiefer and Wolfowitz (1956) show that in many cases the maximum likelihood estimator is consistent with respect to the product of the Euclidean and weak topology. In this paper we derive for several examples of kernels that the sequence $\hat{\theta}_n$ is asymptotically normal and asymptotically efficient in the semiparametric sense. The approach does apply to other examples as well, but is based on a property that must be established for particular examples. Since the examples are among the most frequently studied mixture models in the literature and nothing is known about the rate of convergence and asymptotic distribution of the maximum likelihood estimator, the present study appears worthwhile, even though it is not completely general.

Received August 1994; revised May 1995.

*AMS* 1991 *subject classifications*. 62G20, 62F12.

*Key words and phrases*. Maximum likelihood, semiparametric model, mixture model, Donsker class, asymptotic efficiency, efficient score equation.

The first example is a frailty model studied by, among others, Lindsay (1985), Kumon and Amari (1984), Heckman and Singer (1984), Van der Vaart (1988a) and Pfanzagl (1990). Alternatively to the frailty model considered by Murphy (1995), the survival times are modelled parametrically, while inhomogeneity of the hazards is modelled nonparametrically. The second example is a version of the errors-in-variables model in which the errors are modelled by a Gaussian distribution. Efficient estimators for this model, but not the maximum likelihood estimator, are studied by Bickel and Ritov (1987). See this paper and Anderson (1984) for an introduction to the large literature on the errors-in-variables problem. As a third example, we consider scale mixtures over symmetric densities.

Though the efficiency of the maximum likelihood estimator has been an open problem for many years, several other methods which yield efficient estimators of $\theta$ have been proposed during the last decade. In particular, Bickel and Ritov (1987), Van der Vaart (1988a, b) and Pfanzagl (1990) construct one-step estimators based on an estimated score function. Lindsay (1985) considers (inefficient) estimators based on specifying a parametric form of $\eta$.

We refer to Severini and Wong (1992) for an alternative method to prove asymptotic efficiency of maximum likelihood estimators, based on profile likelihood. In this paper we do not use profile likelihood, but follow a different route based on the efficient influence function.

For the computation of the maximum likelihood estimator the results of Lindsay (1983b) are of interest. These imply that for every fixed $\theta$ the likelihood is maximized with respect to $\eta$ by at least one discrete distribution $\hat{\eta}_n(\theta)$ with at most $n$ support points. The characterizations of the support of $\hat{\eta}_n(\theta)$ by Lindsay (1983b) in combination with a conjugate gradient algorithm or the concave majorant algorithm of Groeneboom (1991) to calculate the weights make feasible the efficient computation of $\hat{\eta}_n(\theta)$. The maximum likelihood estimator $\hat{\theta}_n$ can be calculated by maximizing the profile likelihood $\theta \to \text{lik}(\theta, \hat{\eta}_n(\theta))$ or, preferably, by building an updating procedure for initial estimators for $\theta$ into the iteration steps.

In Section 2 we give the main result of the paper. This is formulated in terms of empirical process theory reviewed in Section 3. Sections 4, 5 and 6 are concerned with the three examples mentioned previously.

## 2. A general result.

The examples treated in this paper admit a "statistic" $\psi_\theta(X)$ which is sufficient for $\eta$ given a fixed value of $\theta$. In every case the efficient score function for $\theta$ (the score for $\theta$ minus its projection on the set of nuisance scores) is given by

$$\tilde{l}_{\theta,\eta}(x) = \dot{l}_{\theta,\eta}(x) - E_\theta\Big(\dot{l}_{\theta,\eta}(X) | \psi_\theta(X) = \psi_\theta(x)\Big), \tag{2.1}$$

where $\dot{l}_{\theta,\eta}(x)$ is the score function for $\theta$: the vector of partial derivatives of the logarithm of the density $p_{\theta,\eta}(x)$ with respect to $\theta$. See Lindsay (1983a) or

Van der Vaart (1988a, c). As a consequence we have that

$$E_{\theta,\eta_0}\tilde{l}_{\theta,\eta}(X) = 0 \quad \text{for every } \theta, \eta, \eta_0. \tag{2.2}$$

This unbiasedness of the efficient score function plays an important role in the analysis of efficient one-step estimators in earlier papers. It will also be crucial for the study of the maximum likelihood estimator in the present paper. As a result of the convexity of the model in the parameter the unbiasedness is true for general mixture models. The explicit expression (2.1) for the efficient score function will be used in order to check technical conditions.

A second special property of the examples in this paper is that the efficient score function is an actual score function, in the sense that for every $(\theta, \eta)$ there exist finite-dimensional submodels $t \to \eta_t(\theta, \eta)$ indexed by a parameter $t$ of the same dimension as $\theta$ ranging over a neighbourhood of the origin such that $\eta_0(\theta, \eta) = \eta$ and

$$\tilde{l}_{\theta,\eta}(x) = \frac{\partial}{\partial t}\log p_{\theta+t,\,\eta_t(\theta,\eta)}(x)\bigg|_{t=0} \quad \text{for every } x. \tag{2.3}$$

This is not true for mixture models in general, not even the ones admitting a sufficient statistic studied here. Though $\dot{l}_{\theta,\eta}(x)$ is an actual score function by its definition, its conditional expectation $E_\theta(\dot{l}_{\theta,\eta}(X)|\psi_\theta(X) = \psi(x))$ may fail to have this property. In general the conditional expectation is in the closure of the linear span of the score functions for the nuisance parameter, but the score functions for the nuisance parameter may form a convex cone rather than a linear space. This is true in particular at $(\theta, \eta)$ for which $\eta$ is a discrete distribution. See Lindsay (1983a) or Van der Vaart (1988a, c). Since often discrete distributions are the only maximizers of the likelihood, this is relevant for the approach of this paper, since it is the maximum likelihood estimator that we wish to perturb in the given manner.

If $\tilde{l}_{\hat{\theta}_n,\hat{\eta}_n}(x)$ is a score function at $(\hat{\theta}_n, \hat{\eta}_n)$ in the sense of (2.3), then it follows that

$$\sum_{i=1}^{n} \tilde{l}_{\hat{\theta}_n,\hat{\eta}_n}(X_i) = 0. \tag{2.4}$$

Indeed, by definition of the maximum likelihood estimator the map $t \to \text{lik}(\hat{\theta}_n + t, \eta_t(\hat{\theta}_n, \hat{\eta}_n))$ is maximal at the point $t = 0$, whence its derivative at $t = 0$ vanishes. We refer to (2.4) as the *efficient score equation*.

The argument may now proceed by a classical linearization scheme. If the efficient score function is smooth in $\theta$ we obtain

$$0 = \sum \tilde{l}_{\hat{\theta}_n,\hat{\eta}_n}(X_i) = \sum \tilde{l}_{\theta_0,\hat{\eta}_n}(X_i) + \sum \dot{\tilde{l}}_{\tilde{\theta}_n,\hat{\eta}_n}(X_i)\big(\hat{\theta}_n - \theta_0\big),$$

for a point $\tilde{\theta}_n$ between $\theta_0$ and $\hat{\theta}_n$. Thus

$$\sqrt{n}\,\big(\hat{\theta}_n - \theta_0\big) = -\left(\frac{1}{n}\sum \dot{\tilde{l}}_{\tilde{\theta}_n,\hat{\eta}_n}(X_i)\right)^{-1} \frac{1}{\sqrt{n}}\sum \tilde{l}_{\theta_0,\hat{\eta}_n}(X_i).$$

This differs from the classical expansion of the maximum likelihood equations in that the (efficient) score function depends on the random nuisance parameter $\hat{\eta}_n$. This difficulty may be overcome by application of empirical process theory. Here the unbiasedness (2.2) ensures that the right-hand side is properly centered. The final conclusion is that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with covariance equal to the inverse of the *efficient information matrix* $E_{\theta,\eta}\tilde{l}_{\theta,\eta}(X)\tilde{l}_{\theta,\eta}(X)'$ evaluated at $(\theta_0, \eta_0)$. Throughout we assume that this is nonsingular.

To minimize regularity conditions the derivation of the preceding paragraph may be replaced by an approach that does not use the second derivative $\dot{\tilde{l}}_{\theta,\eta}$. We impose the following regularity conditions:

$$(2.5)\qquad \int\Big[p_{\theta,\eta_0}^{1/2} - p_{\theta_0,\eta_0}^{1/2} - \tfrac{1}{2}(\theta-\theta_0)'\dot{l}_{\theta_0,\eta_0}p_{\theta_0,\eta_0}^{1/2}\Big]^2\,d\mu = o\big(\|\theta-\theta_0\|^2\big),$$

$$(2.6)\qquad \tilde{l}_{\theta,\eta} \to \tilde{l}_{\theta_0,\eta_0}, \qquad P_{\theta_0,\eta_0}\text{-a.s.},$$

$$(2.7)\qquad \int\|\tilde{l}_{\theta,\eta}\|^2(p_{\theta,\eta_0} + p_{\theta_0,\eta_0})\,d\mu = O(1).$$

These conditions should hold for $\theta \to \theta_0$ and $\eta \to \eta_0$ for a metric $d$ for which the maximum likelihood estimator is known to be consistent.

The following theorem is true for general semiparametric models and arbitrary functions $\tilde{l}_{\theta,\eta}$, though its conditions are motivated by the application to mixture models described previously. In the application to the errors-in-variables model we shall use the theorem in its general form with $\eta$ equal to the mixing distribution and the error variance jointly. It is useful to note that the efficient score equation is not necessary in its full strength. It suffices that

$$(2.8)\qquad \sum_{i=1}^{n}\tilde{l}_{\hat{\theta}_n,\hat{\eta}_n}(X_i) = o_P(\sqrt{n}).$$

This may be satisfied even if the efficient score function is not an actual score function, in which case the present approach still holds. In fact, the proof of the following theorem shows that under regularity conditions (2.8) is necessary for the asymptotic normality and efficiency of the maximum likelihood estimator. A further note is that within the context of the following theorem the estimators $\hat{\eta}_n$ need not be the maximum likelihood estimators. Any consistent estimators for which (2.8) is valid could be used. Furthermore the functions $\tilde{l}_{\theta,\eta}$ may be arbitrary except that $\tilde{l}_{\theta_0,\eta_0}$ should be the efficient score function at $(\theta_0, \eta_0)$. Finally we note that (2.2) may be replaced by

$$(2.9)\qquad \int\tilde{l}_{\hat{\theta}_n,\hat{\eta}_n}p_{\hat{\theta}_n,\eta_0}\,d\mu = o_P(n^{-1/2}).$$

The usefulness of these relaxations is the subject of further study. We do not need them for the examples in this paper.

The notion of a Donsker class is reviewed in Section 3.

THEOREM 2.1. *Suppose that* (2.8) *and* (2.9) *hold and that the class of functions* $\{\tilde{l}_{\theta,\eta}\colon \|\theta-\theta_0\|<\delta,\ d(\eta,\eta_0)<\delta\}$ *is* $P_{\theta_0,\eta_0}$-*Donsker for some* $\delta>0$ *and satisfies* (2.5)–(2.7). *If the maximum likelihood estimator* $(\hat{\theta}_n,\hat{\eta}_n)$ *is consistent for* $(\theta_0,\eta_0)$, *then the sequence* $\sqrt{n}(\hat{\theta}_n-\theta_0)$ *is asymptotically normal. If* $\tilde{l}_{\theta_0,\eta_0}$ *is the efficient score function at* $(\theta_0,\eta_0)$, *then the asymptotic covariance matrix equals the inverse of the efficient information matrix.*

PROOF. A Donsker class which is bounded in $L_1$ is totally bounded (or precompact) in $L_2$. Therefore any sequence $\tilde{l}_{\theta_n,\eta_n}$ with $\theta_n\to\theta_0$ and $\eta_n\to\eta_0$ has a further subsequence that converges in $L_2(P_{\theta_0,\eta_0})$. In view of condition (2.6) the function $\tilde{l}_{\theta_0,\eta_0}$ is the only limit point. Conclude that (2.6) is also valid in an $L_2$-sense.

The assumption that the functions $\tilde{l}_{\theta,\eta}$ form a Donsker class entails that the sequence of processes

$$G_n(\theta,\eta)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\tilde{l}_{\theta,\eta}(X_i)-\int\tilde{l}_{\theta,\eta}p_{\theta_0,\eta_0}\,d\mu\right)$$

converges in distribution in the space $l^\infty((\theta,\eta)\colon \|\theta-\theta_0\|<\delta,\ d(\eta,\eta_0)<\delta)$ of bounded functions on a neighbourhood of the true parameter $(\theta_0,\eta_0)$ to a tight Brownian bridge process $G(\theta,\eta)$. Almost all sample paths of $G$ are uniformly continuous with respect to the semimetric with square

$$\rho^2((\theta_1,\eta_1),(\theta_2,\eta_2))=\int\|\tilde{l}_{\theta_1,\eta_1}-\tilde{l}_{\theta_2,\eta_2}\|^2 p_{\theta_0,\eta_0}\,d\mu.$$

By the $L_2$-version of (2.6) we have that $\rho((\hat{\theta}_n,\hat{\eta}_n),(\theta_0,\eta_0))$ converges to zero in probability. As a consequence of the uniform convergence and continuity of the limit,

$$G_n\big(\hat{\theta}_n,\hat{\eta}_n\big)-G_n(\theta_0,\eta_0)\to_P 0.$$

The second part of the proof consists of showing that

$$\begin{aligned}G_n\big(\hat{\theta}_n,\hat{\eta}_n\big)&=-\sqrt{n}\int\tilde{l}_{\hat{\theta}_n,\hat{\eta}_n}p_{\theta_0,\eta_0}\,d\mu+o_P(1)\\ &=\sqrt{n}\int\tilde{l}_{\hat{\theta}_n,\hat{\eta}_n}\big(p_{\hat{\theta}_n,\eta_0}-p_{\theta_0,\eta_0}\big)\,d\mu+o_P(1)\\ &=\left(\int\tilde{l}_{\theta_0,\eta_0}\dot{l}'_{\theta_0,\eta_0}p_{\theta_0,\eta_0}\,d\mu+o_P(1)\right)\sqrt{n}\big(\hat{\theta}_n-\theta_0\big)+o_P(1).\end{aligned}\tag{2.10}$$

Since the integral in the last line equals the efficient information matrix, this would conclude the proof.

The first equality in (2.10) is the efficient score equation (2.8) and the second equality follows from the unbiasedness (2.9) of the efficient score function. We must prove the third equality. The difference between the second and last line of (2.10) can be written as the sum of three terms:

$$\begin{aligned}\sqrt{n}\int\tilde{l}_{\hat{\theta}_n,\hat{\eta}_n}\Big(p^{1/2}_{\hat{\theta}_n,\eta_0}+p^{1/2}_{\theta_0,\eta_0}\Big)\Big[\Big(p^{1/2}_{\hat{\theta}_n,\eta_0}-p^{1/2}_{\theta_0,\eta_0}\Big)-\tfrac{1}{2}\big(\hat{\theta}_n-\theta_0\big)\dot{l}_{\theta_0,\eta_0}p^{1/2}_{\theta_0,\eta_0}\Big]\,d\mu\\ +\int\tilde{l}_{\hat{\theta}_n,\hat{\eta}_n}\Big(p^{1/2}_{\hat{\theta}_n,\eta_0}-p^{1/2}_{\theta_0,\eta_0}\Big)\tfrac{1}{2}\dot{l}'_{\theta_0,\eta_0}p^{1/2}_{\theta_0,\eta_0}\,d\mu\,\sqrt{n}\,\big(\hat{\theta}_n-\theta_0\big)\\ -\int\Big(\tilde{l}_{\hat{\theta}_n,\hat{\eta}_n}-\tilde{l}_{\theta_0,\eta_0}\Big)\dot{l}'_{\theta_0,\eta_0}p_{\theta_0,\eta_0}\,d\mu\,\sqrt{n}\,\big(\hat{\theta}_n-\theta_0\big).\end{aligned}$$

The first and third terms can easily be seen to be $o_P(\sqrt{n}\,\|\hat{\theta}_n-\theta_0\|)$ by applying the Cauchy–Schwarz inequality together with (2.5)–(2.7). The square of the norm of the integral in the middle term can for every sequence of constants $m_n\to\infty$ be bounded by a multiple of

$$\begin{aligned}m_n^2\int\|\dot{l}_{\hat{\theta}_n,\hat{\eta}_n}\|\,p^{1/2}_{\theta_0,\eta_0}|p^{1/2}_{\hat{\theta}_n,\eta_0}-p^{1/2}_{\theta_0,\eta_0}|\,d\mu^2\\ +\int\|\tilde{l}_{\hat{\theta}_n,\hat{\eta}_n}\|^2\big(p_{\hat{\theta}_n,\eta_0}+p_{\theta_0,\eta_0}\big)\,d\mu\int_{\|\dot{l}_{\theta_0,\eta_0}\|>m_n}\|\dot{l}_{\theta_0,\eta_0}\|^2p_{\theta_0,\eta_0}\,d\mu.\end{aligned}$$

In view of (2.5) and the Cauchy–Schwarz inequality, the first term converges to zero in probability provided $m_n\to\infty$ sufficiently slowly to ensure that $m_n\|\hat{\theta}_n-\theta_0\|\to_P 0$. [Such a sequence exists. If $Z_n\to_P 0$, then there exists a sequence $\varepsilon_n\downarrow 0$ such that $P(|Z_n|>\varepsilon_n)\to 0$. Then $\varepsilon_n^{-1/2}Z_n\to_P 0$.] In view of (2.7) the second term converges to zero in probability for every $m_n\to\infty$. This concludes the proof of (2.10). □

The assumption of consistency of the maximum likelihood estimator allows us to localize the conditions to a neighbourhood of the true parameter. One possibility to establish consistency is the method of Wald (1949). This method is applied by Kiefer and Wolfowitz (1956) to obtain consistency of the maximum likelihood estimator in mixture models. Under some regularity conditions they prove consistency for a metric that generates the weak topology. Another possibility is to prove that

$$\sup_{\substack{\theta\in\Theta\\ \eta\in H}}\left|\frac{1}{n}\sum_{i=1}^{n}\log\frac{p_{\theta,\eta}}{p_{\theta_0,\eta_0}}(X_i)-\int\log\frac{p_{\theta,\eta}}{p_{\theta_0,\eta_0}}p_{\theta_0,\eta_0}\,d\mu\right|$$

converges to zero in outer probability. This requires that the class of functions $\log(p_{\theta,\eta}/p_{\theta_0,\eta_0})$ is Glivenko–Cantelli and is of similar character to the other conditions in this paper. If the true value $(\theta_0,\eta_0)$ is a "well-separated" maximum of the Kullback–Leibler information function $(\theta,\eta)\to\int\log(p_{\theta,\eta}/p_{\theta_0,\eta_0})p_{\theta_0,\eta_0}\,d\mu$, then consistency follows. Consistency of the maximum likelihood estimator for a mixture distribution in the absence of the parameter $\theta$ can be proved under very weak regularity conditions making use of the concavity of the likelihood, as shown by Pfanzagl (1988). This approach is not possible for estimating $(\theta, \eta)$ jointly, but variations on Pfanzagl's method may help to relax regularity conditions a little. (The idea of the method is not to use the log density as criterion function, but another better behaved function.) We do not address the matter of consistency in great detail in this paper.

Condition (2.5) simply requires that the score function for $\theta$ exists in an $L_2$-sense. For mixture models it is implied by differentiability of the kernels in the following manner:

$$\int\int\Big[p_\theta^{1/2}(x|z) - p_{\theta_0}^{1/2}(x|z) - \tfrac{1}{2}(\theta-\theta_0)'\dot{l}_{\theta_0}(x|z)p_{\theta_0}^{1/2}(x|z)\Big]^2\,d\mu(x)\,d\eta_0(z)$$
$$= o\big(\|\theta-\theta_0\|^2\big).$$

In this case the score function for $\theta$ in the mixture model is related to the score functions for $\theta$ in the model of the kernel through

$$\dot{l}_{\theta,\eta}(x) = \frac{\int\dot{l}_\theta(x|z)p_\theta(x|z)\,d\eta(z)}{\int p_\theta(x|z)\,d\eta(z)}. \tag{2.11}$$

See, for instance, Van der Vaart [(1988a), Lemma 5.18]. The conditional expectation of $\dot{l}_{\theta,\eta}$ can be found as

$$E_\theta\big(\dot{l}_{\theta,\eta}(X)|\psi_\theta(X)\big) = \frac{\int E_\theta\big(\dot{l}_\theta(X|z)|\psi_\theta(X)\big)p_\theta(X|z)\,d\eta(z)}{\int p_\theta(X|z)\,d\eta(z)}.$$

**3. Donsker classes.** In this section we review some results on empirical processes that are used repeatedly in later sections of the paper. Let $\mathscr{F}$ be a class of measurable functions $f\colon \mathscr{X}\to\mathbb{R}$ on the probability space $(\mathscr{X},\mathscr{A},P)$. The empirical measure $\mathbb{P}_n = \Sigma_{i=1}^n\delta_{X_i}$ of an i.i.d. sample from $P$ is the discrete random measure that puts mass $1/n$ at every observation. The *empirical process* $\mathbb{G}_n = \sqrt{n}\,(\mathbb{P}_n - P)$ evaluated at the function $f$ is

$$\mathbb{G}_n f = \frac{1}{\sqrt{n}}\sum_{i=1}^n\Big(f(X_i) - \int f\,dP\Big).$$

The class $\mathscr{F}$ is called *Donsker* if the empirical process $\{\mathbb{G}_n(f)\colon f\in\mathscr{F}\}$ converges in distribution in the metric space $l^\infty(\mathscr{F})$ of all bounded functions $z\colon \mathscr{F}\to\mathbb{R}$, which is equipped with the supremum norm. To avoid problems with measurability, convergence in distribution is defined in the sense of outer expectations as in Dudley (1985).

Let $\|f\|_{P,2}$ denote the $L_2(P)$-norm of a function $f$. Given a pair of functions $l\le u$ the bracket $[l,u]$ consists of all functions $f$ with $l\le f\le u$. The bracketing number $N_{[\,]}(\varepsilon,\mathscr{F},L_2(P))$ is the minimal number of brackets $[l,u]$ of size

$P(u-l)^2$ smaller than $\varepsilon^2$ needed to cover $\mathscr{F}$. According to a theorem of Ossiander (1987) a sufficient condition for $\mathscr{F}$ to be Donsker is that

$$\int_0^\infty \sqrt{\log N_{[\,]}(\varepsilon, \mathscr{F}, L_2(P))}\, d\varepsilon < \infty. \tag{3.1}$$

This is referred to as $\mathscr{F}$ having a finite bracketing (entropy) integral.

Important examples of classes with a finite bracketing entropy integral are classes of smooth functions on Euclidean spaces. To define such classes let, for a given function $f\colon I \subset \mathbb{R}^d \to \mathbb{R}$ and $\alpha > 0$,

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

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

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

For $\alpha \le 1$ the norm $\|\cdot\|_\alpha$ is the Lipschitz norm of order $\alpha$, while for larger values of $\alpha$ the norm involves bounds on the partial derivatives of $f$ together with a Lipschitz norm on the partial derivatives of highest order. Let $C_M^\alpha(I)$ be the set of all continuous functions $f\colon I \to \mathbb{R}$ with $\|f\|_\alpha \le M$.

A classical result of Kolmogorov and Tikhomorov (1961) asserts that the entropy numbers of $C_M^\alpha(I)$ for the uniform norm (the logarithm of the number of balls of radius $\varepsilon$ needed to cover the class) are of the order $(1/\varepsilon)^{d/\alpha}$ for each given convex, bounded subset $I$ of $\mathbb{R}^d$. Thus the bracketing entropy integral of the class $C_M^\alpha(I)$ is finite for $\alpha > d/2$.

Van der Vaart (1993) extends this to classes of smooth functions on an unbounded support. Let $\mathbb{R}^d = \bigcup_{j=1}^\infty I_j$ be a partition into cubes of uniformly bounded size. Let $\mathscr{F}$ be a class of functions $f\colon \mathbb{R}^d \to \mathbb{R}$ such that the restrictions $f_{|I_j}$ belong to $C_{M_j}^\alpha(I_j)$ for every $j$ and some fixed $\alpha > d/2$. Then the class $\mathscr{F}$ is Donsker if and only if

$$\sum_j M_j P^{1/2}(I_j) < \infty. \tag{3.2}$$

An earlier result in this direction was obtained by Giné and Zinn (1986). In some situations it is also useful to have an explicit upper bound for the bracketing entropy of these classes $\mathscr{F}$. A simple bound obtained by Van der Vaart (1994) is as follows: there exists a constant $K$ depending only on $\alpha$, $V$, $r$, $d$ and the uniform bound on the diameter of the sets $I_j$ such that for $V \ge d/\alpha$,

$$\log N_{\square}(\varepsilon, \mathscr{F}, L_r(P)) \le K\left(\frac{1}{\varepsilon}\right)^V \left(\sum_{j=1}^\infty M_j^{V_r/(V+r)} P(I_j)^{V/(V+r)}\right)^{V+r/r}. \tag{3.3}$$

This implies that the class $\mathscr{F}$ satisfies (3.1) if the series on the right is convergent for $r = 2$ and some $V < 2$. This is slightly worse than the necessary and sufficient condition (3.2).

The class of functions $\mathscr{F}$ is said to be Glivenko–Cantelli if $\sup_f |\mathbb{P}_n f - \int f\,dP|$ converges almost surely to zero. A sufficient condition is that the bracketing numbers $N_{[\,]}(\varepsilon, \mathscr{F}, L_1(P))$ are finite for every $\varepsilon > 0$. For the class $\mathscr{F}$ as considered in the preceding paragraph this is the case if and only $\sum_j M_j P(I_j) < \infty$.

## 4. A frailty model.

Write the observations as pairs $(X_i, Y_i)$ and consider the mixture model with kernel

$$p_\theta(x, y|z) = ze^{-zx}\theta ze^{-\theta zy}.$$

Thus given unobservable variables $Z_i = z$ each observation consists of a pair of exponentially distributed variables with hazards $z$ and $\theta z$, respectively. The problem is to estimate the common ratio of the hazards $\theta$.

As sufficient statistic we use $\psi_\theta(X, Y) = X + \theta Y$. In the parametric model given by the kernel the score function for $\theta$ equals

$$\dot{l}_\theta(x, y|z) = \frac{1}{\theta} - zy.$$

The score function for $\theta$ in the mixture model is given by (2.11). Given $X + \theta Y = s$ the variables $X$ and $\theta Y$ are uniformly distributed on the interval $[0, s]$. This yields the efficient score function as in (2.1) given by

$$\tilde{l}_{\theta,\eta}(x, y) = \frac{\int \frac{1}{2}(x - \theta y) z^3 \exp(-z(x + \theta y))\, d\eta(z)}{\int \theta z^2 \exp(-z(x + \theta y))\, d\eta(z)}.$$

The circumstance that this is an actual score function is a consequence of the even more special fact that in the parametric model of the kernel the conditional score function for $\theta$ is proportional to the score for $z$:

$$E_\theta\big(\dot{l}_\theta(X, Y|z)|X + \theta Y\big) = \frac{1}{\theta} - z\frac{X + \theta Y}{2\theta} = \frac{z}{2\theta}\frac{\partial}{\partial z}\log p_\theta(X, Y|z).$$

Suppose that the "true" model belongs to the parametric model of the kernel $p_\theta(x, y|z)$. Then the left side is the projection in the mixture model of the score function for $\theta$ on the closed linear span of the score function for the (unknown) mixing distribution. The right side is the projection of this same score function on the score for $z$ in the parametric submodel $\{p_\theta(x, y, |z)$: $\theta \in \Theta$, $z \in \mathscr{Z}\}$. Thus an interpretation of the identity in terms of information numbers is that estimating $\theta$ does not become harder if the parametric model given by the kernel is enlarged to the mixture model. The technical implication is that

$$
\begin{aligned}
-E_\theta\bigl(\dot{l}_{\theta,\eta}(X,Y)|X+\theta Y\bigr)
&= -\frac{\int(z/2\theta)(\partial/\partial z)\log p_\theta(X,Y|z)p_\theta(X,Y|z)\,d\eta(z)}{\int p_\theta(X,Y|z)\,d\eta(z)}\\
&= \frac{\partial}{\partial t}\log\int p_\theta\left(X,Y|z\frac{1-t}{2\theta}\right)d\eta(z)\bigg|_{t=0}.
\end{aligned}
$$

It follows that the efficient score function $\tilde{l}_{\theta,\eta}$ is the score function at $t=0$ of the one-dimensional "least favorable" submodel $t \to p_{\theta+t,\eta_t}(x,y)$ for the probability measures $\eta_t$ given by

$$\eta_t(B) = \eta\bigl(B(1-t/2\theta)^{-1}\bigr).$$

The asymptotic normality and efficiency of the maximum likelihood estimator $\hat{\theta}_n$ follow provided the regularity conditions of Theorem 2.1 hold.

Conditions (2.5) and (2.6) are satisfied for any true parameter $(\theta_0,\eta_0)$. We shall check the other conditions under moment conditions on the true mixing distribution $\eta_0$. The present approach does not seem to yield asymptotic efficiency without imposing some restrictions. In comparison Van der Vaart (1988a, c) has shown that asymptotically efficient estimators exist in complete generality, using the one-step method and extensive truncation. Even a refined version of the present argument does not appear to yield an equally strong result. It may be noted that already for the one-step estimators considered in Pfanzagl (1990) some moment conditions appear necessary. Maximum likelihood estimators may require stronger conditions.

COROLLARY 4.1. *Suppose that the true mixing distribution $\eta_0$ satisfies $\int(z^2+z^{-5})\,d\eta_0(z)<\infty$. Then maximum likelihood estimator for $\theta$ is asymptotically efficient.*

PROOF. Consistency of the maximum likelihood estimator $(\hat{\theta}_n,\hat{\eta}_n)$ for the product of the Euclidean and weak topology follows from Kiefer and Wolfowitz (1956).

By applying Lemma L.23 of Pfanzagl (1990) repeatedly it follows that there exists a constant $C$ and a weak neighborhood $V$ of the true mixing distribution such that

$$\sup_{\eta\in V}\frac{\int_0^\infty z^{k+l}e^{-zs}\,d\eta(z)}{\int_0^\infty z^k e^{-zs}\,d\eta(z)} \le \begin{cases} C^l\left(\dfrac{|\log s|}{s}\right)^l, & s<1/2\\ C^l, & s\ge 1/2.\end{cases} \tag{4.1}$$

If we write $h_\eta(s)$ for the quotient in the supremum on the left side for $k=2$ and $l=1$, then $\tilde{l}_{\theta,\eta}(x,y)=(x-\theta y)/2\theta h_\eta(x+\theta y)$, and we immediately obtain, with $U = \{\theta : \|\theta - \theta_0\| < \delta\}$,

$$\sup_{\theta \in U} \sup_{\eta \in V} |\tilde{l}_{\theta,\eta}(x, y)| \leq \sup_{\theta \in U} \frac{C}{2\theta}(|\log(x + \theta y)| + |x + \theta y|)$$
$$\leq C'(|\log x| + |x| + |y|).$$

Condition (2.7) follows if $|\log X|$, $X$ and $Y$ have a finite second moment under $(\theta, \eta_0)$ uniformly over $\theta$ in a neighborhood of $\theta_0$. This is valid under our assumptions on $\eta_0$.

The class of functions $\{\tilde{l}_{\theta,\eta} : \theta \in U, \eta \in V\}$ will be shown to be a Donsker class by verifying that it satisfies Ossiander's condition (3.1).

Consider first the class of functions $s \to sh_\eta(s)$ as $\eta$ ranges over $V$ on the domain $(0, \infty) \subset \mathbb{R}$. We shall construct brackets by first constructing brackets on the subdomains $(0, 1/2]$ and $(1/2, \infty)$ separately. In view of (4.1) we have for every $1/2 < \alpha < 1$ and every $\eta \in V$, letting $\lesssim$ denote less than equal up to a constant,

$$|sh_\eta(s)| \lesssim |\log s|, \qquad s < 1/2,$$

$$|s_1 h_\eta(s_1) - s_2 h_\eta(s_2)| \lesssim |s_1 - s_2|^\alpha \sup_{s_1 < s < s_2} \big(sh_\eta(s)'\big)^\alpha \sup_{s_1 < s < s_2} \big(2sh_\eta(s)\big)^{1-\alpha}$$
$$\lesssim |s_1 - s_2|^\alpha \frac{|\log s_1|^{1+\alpha}}{s_1^\alpha}, \qquad 0 < s_1 < s_2 < 1/2.$$

Thus the restrictions of the functions $s \to sh_\eta(s)$ to an interval $[a, b] \subset (0, 1/2]$ belong to the space $C_M^\alpha[a, b]$ for $M = |\log a|^{1+\alpha}/a^\alpha$. Similarly we have

$$|sh_\eta(s)| \lesssim |s|, \qquad s \geq 1/2,$$

$$|s_1 h_\eta(s_1) - s_2 h_\eta(s_2)| \lesssim |s_1 - s_2| s_2, \qquad 1/2 < s_1 < s_2.$$

Thus the restrictions of the functions $s \to sh_\eta(s)$ to an interval $[a, b] \subset [1/2, \infty)$ belong to the space $C_M^1[a, b]$ for $M = b$. We now apply (3.3) with the partitions $(0, 1/2) = \cup_i (2^{-i}, 2^{-i+1})$ and $[1/2, \infty) = [1/2, 1) \cup \cup_i [i, i + 1)$ to see that for every $W \geq 1/\alpha$,

$$\log N_{[\,]}\big(\varepsilon, \{sh_\eta(s) : \eta \in V\}, L_2(Q)\big) \geq K\left(\frac{1}{\varepsilon}\right)^W, \tag{4.2}$$

for a constant $K$ depending only on $\alpha$ and $W$ and the numbers

$$\sum_i \left(\frac{|\log 2^{-i}|^{2+2\alpha}}{2^{-2i\alpha}} Q[2^{-i}, 2^{-i+1})\right)^{W/(W+2)}; \qquad \sum_i \big(i^2 Q[i, i + 1)\big)^{W/(W+2)},$$

provided these numbers are finite.

We apply this inequality for the measure $Q$ equal to the distribution of $X + \theta Y$ for a given fixed $\theta$ and $(X, Y)$ distributed according to $(\theta_0, \eta_0)$. By a straightforward calculation, the density at $s$ of $X + \theta Y$ given $Z = z$ is bounded above by $\theta_0/\theta z^2 s \exp(-z(1 \wedge \theta_0/\theta)s$. It follows that $Q[2^{-i}, 2^{-i+1})$

$\lesssim 2^{-i^2}\theta_0/\theta \int z^2 \, d\eta_0(z)$ and the first series converges for every $W$. Similarly $Q[i, i+1)$ is bounded above by $Q[i, \infty) \lesssim i^{-k} \int z^{-k} \, d\eta_0(z)$ and the second series converges for some $W < 2$ sufficiently close to 2 provided $\int z^{-k} \, d\eta_0(z) < \infty$ for some $k > 4$. Both series are bounded uniformly in $\theta \in U$.

This concludes the proof that (4.2) is valid for some $W < 2$ and $Q$ equal to the distribution of $X + \theta Y$, for a constant $K$ not depending on $\theta \in U$. Alternatively this inequality can be formulated in terms of the functions $(x, y) \to (x + \theta y)h_\eta(x + \theta y)$. Letting $\mathscr{G}_\theta$ be the set of all such functions as $\eta$ varies over $V$ and $P_0$, the distribution of $(X, Y)$ under $(\theta_0, \eta_0)$, we have

$$\log N_{[]}(\varepsilon, \mathscr{G}_\theta, L_2(P_0)) \leq K\left(\frac{1}{\varepsilon}\right)^W.$$

Still for a fixed $\theta$ the functions $\tilde{l}_{\theta,\eta}$ can be written

$$\tilde{l}_{\theta,\eta}(x, y) = \frac{x - \theta y}{x + \theta y}(x + \theta y)h_\eta(x + \theta y).$$

Thus the class $\mathscr{F}_\theta$ of functions $\tilde{l}_{\theta,\eta}$ when $\eta$ varies over $V$ is obtained from $\mathscr{G}_\theta$ by multiplication by a fixed, uniformly bounded function. It is not hard to see that

$$\log N_{[]}(\varepsilon, \mathscr{F}_\theta, L_2(P_0)) \leq \log N_{[]}(\varepsilon, \mathscr{G}_\theta, L_2(P_0)).$$

Finally the class of interest $\mathscr{F} = \cup_{\theta \in U} \mathscr{F}_\theta$ can be seen to be Donsker by the lemma below upon noting that

$$\begin{aligned} \left| \frac{\partial}{\partial\theta} \tilde{l}_{\theta,\eta}(x, y) \right| &= |-yh_\eta(x + \theta_y) + (x - \theta y)h'_\eta(x + \theta y)y| \\ &\lesssim |\log(x + \theta y)|^2 + (x + \theta y)^2. \end{aligned}$$

The proof is complete, because the right side is bounded by a multiple of $|\log x|^2 + x^2 + y^2$, which is square integrable. □

LEMMA 4.2. *Suppose that for every $\theta$ in a bounded interval in $\mathbb{R}$ a class $\mathscr{F}_\theta = \{f_{\theta,\eta} : \eta \in V\}$ of measurable functions is given such that for some $W < 2$ and a constant K not depending on $\theta$*

$$\log N_{[]}(\varepsilon, \mathscr{F}_\theta, L_2(P)) \leq K\left(\frac{1}{\varepsilon}\right)^W.$$

*Moreover assume that for every $\theta_1$, $\theta_2$ and $\eta$*

$$|f_{\theta_1,\eta} - f_{\theta_2,\eta}| \leq F|\theta_1 - \theta_2|,$$

*for some function F with $PF^2 < \infty$. Then $\mathscr{F} = \cup_\theta \mathscr{F}_\theta$ is P-Donsker.*

PROOF. Choose an $\varepsilon$-net of points $\theta_1, \dots, \theta_p$. The number of elements $p$ can be chosen bounded by $(2/\varepsilon)$ times the length of the interval. For every $\theta_i$ form a collection of $\varepsilon$-brackets $[l_{i,j}, u_{i,j}]$ covering $\mathscr{F}_{\theta_i}$. The number of brackets can be chosen bounded by $\exp(K(1/\varepsilon)^W)$. Now form the brackets

$$[l_{i,j} - \varepsilon F, u_{i,j} + \varepsilon F].$$

These have size bounded by $2\|F\|\varepsilon + \varepsilon$ and as $i$ and $j$ range over all possible values they cover $\mathscr{F}$. The total number of brackets is bounded by a multiple of $(2/\varepsilon)\exp(K(1/\varepsilon)^W)$. Hence (3.1) is satisfied. □

**5. Errors-in-variables.** Let the observations be a sample of pairs $(X_i, Y_i)$ with the same distribution as

$$X = Z + e,$$
$$Y = \alpha + \beta Z + f,$$

for a bivariate normal vector $(e, f)$ with mean zero and covariance matrix $\Sigma$ and a random variable $Z$ with distribution $\gamma$, independent of $(e, f)$. Thus $Y_i$ is a linear regression on a variable $Z_i$ which is observed with error. The parameter of interest is $\theta = (\alpha, \beta)$ and the nuisance parameter is $\eta = (\Sigma, \gamma)$. To make the parameters identifiable one can put restrictions on either $\Sigma$ or $\gamma$. It suffices that $\gamma$ is not normal (where a degenerate distribution is considered normal with variance zero). Alternatively it can be assumed that $\Sigma$ is known up to a scalar. We refer to the extensive literature on the model, reviewed in Anderson (1984) and Bickel and Ritov (1987). The second paper gives a construction of an asymptotically efficient estimator of $\theta$ by the one-step method with estimated efficient score function.

We consider the case that $\Sigma$ is a diagonal matrix with diagonal elements $\sigma^2$ and $\tau^2$ of which the ratio $\sigma/\tau$ is known. This is called the *restrictive model* in Bickel, Klaassen, Ritov and Wellner (1993). The case that $\Sigma$ is known up to a scalar can be treated by the same method, but the formulas will be longer. The density of the observations is

$$p_{\theta,\eta}(x, y) = \int \frac{1}{\sigma}\phi\left(\frac{x - z}{\sigma}\right)\frac{1}{\tau}\phi\left(\frac{y - \alpha - \beta z}{\tau}\right) d\gamma(z).$$

Given $(\theta, \Sigma)$ a sufficient statistic for $\gamma$ is $\psi_{\theta,\Sigma}(X, Y) = \sigma^{-2}X + \tau^{-2}(Y - \alpha)\beta$. The efficient score function for $(\theta, \Sigma)$ can be computed as in (2.1). We shall only be interested in the components corresponding to $\alpha$ and $\beta$, which are given by

$$\tilde{l}_{\theta,\eta|\alpha}(x, y) = \frac{-\beta X + Y - \alpha}{\tau^2 + \sigma^2\beta^2},$$

$$\tilde{l}_{\theta,\eta|\beta}(x, y) = \frac{-\beta X + Y - \alpha}{\tau^2 + \sigma^2\beta^2}\,\frac{\int z p_{\theta,\Sigma}(X, Y|z)\, d\gamma(z)}{\int p_{\theta,\Sigma}(X, Y|z)\, d\gamma(z)}.$$

The unbiasedness condition (2.2) can be verified directly and takes the form

$$\left.\begin{array}{l} E_{\theta_0,\Sigma_0,\gamma_0}\tilde{l}_{\theta_0,\Sigma,\gamma|\alpha}(X, Y) = 0, \\ E_{\theta_0,\Sigma_0,\gamma_0}\tilde{l}_{\theta_0,\Sigma,\gamma|\beta}(X, Y) = 0, \end{array}\right. \quad \text{for every } \theta_0, \Sigma_0, \Sigma, \gamma_0, \gamma.$$

It is essential that the efficient score is also unbiased in $\Sigma$, which within the context of Theorem 2.1 can therefore be treated in the same manner as the mixing distribution $\gamma$. The validity of the second equation depends on the assumption that the ratio of the diagonal elements is the same for both $\Sigma_0$ and $\Sigma$. This equation is not valid for the "unrestricted version" of the model.

The circumstance that the efficient score function for $\theta$ is an actual score function follows in a similar manner as in the frailty model. As in the example of the paired exponential model, this has the interpretation that the problem of estimating $\theta$ does not become harder if the model given by the kernel is enlarged to the mixture model. This was first noted by Bickel and Ritov (1987). Indeed

$$E_{\theta,\Sigma}\big(\dot{l}_{\theta|\alpha}(X,Y|z)|\psi_{\theta,\Sigma}(X,Y)\big) = \frac{\beta\sigma^2}{\tau^2+\beta^2\sigma^2}\frac{\partial}{\partial z}\log p_{\theta,\Sigma}(X,Y|z).$$

By a similar argument as for the frailty model, it follows that $\tilde{l}_{\theta,\eta|\alpha}(x,y)$ and $\tilde{l}_{\theta,\eta|\beta}(x,y)$ are score functions at $t=0$ for the one-dimensional submodels $t \to p_{\theta+t,\eta_t}$ defined by the probability measures

$$\eta_t(B) = \eta\left(B + \frac{t\beta\sigma^2}{\tau^2+\beta^2\sigma^2}\right)$$

and

$$\eta_t(B) = \eta\left(B\left(1 - \frac{t\beta\sigma^2}{\tau^2+\beta^2\sigma^2}\right)^{-1}\right),$$

respectively. To prove that the maximum likelihood estimator for $\hat{\theta}_n$ is asymptotically normal and efficient, it suffices to establish consistency and to check the regularity conditions of Theorem 2.1.

COROLLARY 5.1. *In the "restrictive" model (where the ratio $\sigma/\tau$ is known) suppose that $\gamma_0$ possesses a finite absolute 11th moment and that $\sigma$ (and hence $\tau$) are known to belong to a known compact interval bounded away from zero and infinity. Then the maximum likelihood estimator for $(\alpha,\beta)$ is asymptotically efficient.*

PROOF. Conditions (2.5) and (2.6) are satisfied provided $\gamma_0$ has a finite second moment. For the verification of the other conditions it is useful to rewrite

$$\tilde{l}_{\theta,\eta|\beta}(x,y) = \frac{-\beta X + Y - \alpha}{\tau^2+\sigma^2\beta^2}h_{\theta,\eta}\big(\sigma^{-2}x + \tau^{-2}(y-\alpha)\beta\big),$$

for the function $h_{\theta,\eta}$ given by

$$h_{\theta,\eta}(s) = \frac{\int z\exp(sz)\exp(-z^2/2\sigma^2 - \beta^2z^2/2\tau^2)\,d\gamma(z)}{\int\exp(sz)\exp(-z^2/2\sigma^2 - \beta^2z^2/2\tau^2)\,d\gamma(z)}.$$

Extending Lemma L.27 of Pfanzagl (1990) we can show that for every $\eta_0 = (\Sigma_0, \gamma_0)$ there exists a neighbourhood $U$ around $\gamma_0$ in the weak topology such that

$$\sup_{\|\theta-\theta_0\|<\delta}\ \sup_{\|\Sigma-\Sigma_0\|<\delta}\ \sup_{\gamma\in U} |h^{(i)}_{\theta,\eta}(s)| \le C(1+|s|)^i, \qquad i = 0,1,2,$$

for a constant $C$ depending only on $\eta_0$, $\delta$ and $U$. As a first application this yields the bound

$$\sup_{\|\theta-\theta_0\|<\delta}\ \sup_{\|\Sigma-\Sigma_0\|<\delta}\ \sup_{\gamma\in U} |\tilde{l}_{\theta,\eta|\beta}(x,y)| \le C(1+x^2+y^2).$$

Since $E_{\theta,\eta}(X^2+Y^2)$ is bounded by a multiple of $\int z^2\,d\gamma(z) + \alpha^2 + \beta^2 + \sigma^2 + \tau^2$, condition (2.7) follows if $\gamma_0$ has a finite second moment.

The class of functions $\{\tilde{l}_{\theta,\eta|\beta}\colon \|\theta-\theta_0\| < \delta,\ \eta \in U\}$ can be shown to be Donsker with the help of (3.2). Elementary calculations show that the partial derivatives with respect to $x$ and $y$ of the first and second order are bounded by a fourth degree polynomial in $|x|$ and $|y|$. Precisely,

$$\sup_{\|\theta-\theta_0\|<\delta}\ \sup_{\|\Sigma-\Sigma_0\|<\delta}\ \sup_{\gamma\in U} |D^i\tilde{l}_{\theta,\eta|\beta}(x,y)| \le C\big(|x|^4+|y|^4+1\big),$$

for a constant $C$ depending only on $\delta$ and $U$. Thus we can apply (3.2) with the partition $\mathbb{R}^2 = \bigcup_{k,l}(k-1,k]\times(l-1,l]$ and the constants $M_{k,l} = k^4 + l^4 + 1$. The class is certainly Donsker if

$$\sum_{k,l}(k^4+l^4+1)P_0^{1/2}(k-1<X\le k,\ l-1<Y\le l) < \infty.$$

This is certainly ensured by a finite absolute 11th moment of $\gamma_0$.

Suppose that this is true and that $\sigma$ and $\tau$ are known to belong to an interval that is bounded away from zero. Then the consistency of $(\hat{\theta}_n, \hat{\eta}_n)$ may be proved by the arguments of Kiefer and Wolfowitz (1956), provided the true parameter $(\theta_0, \eta_0)$ is identifiable. The proof of asymptotic normality and efficiency of $\hat{\theta}_n$ is complete. □

**6. Scale mixture.** Let $\phi$ denote a probability density that is symmetric about zero and consider the mixture model with kernel

$$p_\theta(x|z) = \frac{1}{z}\phi\left(\frac{x-\theta}{z}\right).$$

In this example the mixture density is symmetric about $\theta$ as well and one may estimate $\theta$ asymptotically efficiently with a fully adaptive estimator. See Stone (1975) and Bickel (1982). Alternatively, Van der Vaart (1988a) constructs an efficient one-step estimator that takes the mixture form of the underlying distribution into account. In this section it is shown that the full maximum likelihood estimator is asymptotically efficient.

For simplicity it is assumed that the mixing distribution $\eta$ is supported on a fixed interval $[m, M] \subset (0,\infty)$. This may be relaxed to moment conditions, but the precise argument would have to take into account special properties of the density $\phi$. Here we are interested to show that the conditions of Theorem 2.1 are valid in fair generality. Let $H$ be the set of all probability distributions on this interval. Assume that $\phi$ is twice continuously differentiable with finite Fisher information for location.

Consistency of the sequence of maximum likelihood estimators $(\hat\theta_n, \hat\eta_n)$ for the product of the Euclidean and the weak topology can be proved by the method of Kiefer and Wolfowitz (1956). Since the model is fully adaptive in the sense of Bickel (1981), the efficient score function for $\theta$ equals the ordinary score function for $\theta$. Thus (2.4) is trivially satisfied. The unbiasedness (2.2) follows from the argument in the Introduction using the sufficient statistic $\psi_\theta(X) = |X - \theta|$ or can be verified directly. It suffices to verify the regularity conditions of the theorem.

Conditions (2.5) and (2.6) are satisfied for any $\theta_0$ and $\eta_0$. For condition (2.7) we use the bound

$$|\dot l_{\theta,\eta}(x)| = \frac{|\int z^{-2}\phi'((x-\theta)/z)\,d\eta(z)|}{\int z^{-1}\phi((x-\theta)/z)\,d\eta(z)} \le \sup_z \frac{1}{m}\left|\frac{\phi'}{\phi}\left(\frac{x-\theta}{z}\right)\right|.$$

Condition (2.7) is satisfied if the function on the right is square integrable uniformly in $\theta$ in a neighbourhood of $\theta_0$. The class of functions $\{\dot l_{\theta,\eta}: \|\theta - \theta_0\| < \delta,\ \eta \in H\}$ can be shown to be Donsker with the help of (3.2). We have

$$\begin{aligned}\left|\frac{\partial}{\partial x}\dot l_{\theta,\eta}(x)\right| &\le \frac{|\int z^{-3}\phi''((x-\theta)/z)\,d\eta(z)|}{\int z^{-1}\phi((x-\theta)/z)\,d\eta(z)} + \dot l^2_{\theta,\eta}(x)\\ &\le \frac{1}{m^2}\sup_z\left|\frac{\phi''}{\phi}\left(\frac{x-\theta}{z}\right)\right| + \frac{1}{m^2}\sup_z\left|\frac{\phi'}{\phi}\left(\frac{x-\theta}{z}\right)\right|^2.\end{aligned}$$

Thus we may apply (3.2) with the partition $\mathbb{R} = \bigcup_{j=-\infty}^{\infty}(j-1, j]$ and constants $M_j$ equal to the maximum values of the function on the right-hand side on the intervals $(j-1, j]$.

The preceding estimates can be used for a variety of kernels. We close this section with formal statements for two examples.

COROLLARY 6.1. *If $\phi$ is the normal or logistic density and the mixing distribution is known to belong to a known interval* $[m, M] \subset (0, \infty)$, *then the maximum likelihood estimator for $\theta$ is asymptotically efficient.*

PROOF. For the normal kernel $\phi$ the function $\phi'/\phi$ is bounded in absolute value by $|x|$ and the function $|\phi''/\phi|$ can be bounded by a multiple of $(x^2 + 1)$. We can apply (3.2) with constants $M_j = j^2$. The series

$$\sum j^2 P_0^{1/2}(j - 1 < X \le j)$$

is certainly finite if $E_0 X^6$ is finite. This translates into a moment condition on $\eta_0$, which is certainly satisfied if $\eta_0$ has compact support.

For the logistic kernel both the score function $\phi'/\phi$ and the function $\phi''/\phi$ are uniformly bounded. We may apply (3.2) with the constants $M_j$ equal to 1. By the same argument the class of score functions is a Donsker class. □

## REFERENCES

ANDERSON, T. W. (1984). Estimating linear statistical relationships. *Ann. Statist.* **12** 1–45.

BICKEL, P. J. (1982). On adaptive estimation. *Ann Statist.* **10** 647–671.

BICKEL, P. J. and RITOV, Y. (1987). Efficient estimation in the errors in variables model. *Ann. Statist.* **15** 513–540.

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

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.

GINÉ, E. and ZINN, J. (1986). Empirical processes indexed by Lipschitz functions. *Ann. Probab.* **14** 1329–1338.

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

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.

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.

KOLMOGOROV, A. N. and TIKHOMIROV, V. M. (1961). Epsilon-entropy and epsilon-capacity of sets in functions spaces. *Amer. Math. Soc. Transl. Ser. 2* **17** 277–364.

KUMON, M. and AMARI, S. (1984). Estimation of a structural parameter in the presence of a large number of nuisance parameters. *Biometrika* **71** 445–459.

LINDSAY, B. G. (1983a). Efficiency of the conditional score in a mixture setting. *Ann. Statist.* **11** 486–497.

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

LINDSAY, B. G. (1985). Using empirical Bayes inference for increased efficiency. *Ann. Statist.* **13** 914–931.

MURPHY, S. A. (1995). Asymptotic theory for the frailty model. *Ann. Statist.* **23** 182–198.

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.

PFANZAGL, J. (1990). *Estimation in Semiparametric Models*. Springer, New York.

SEVERINI, T. A. and WONG, W. H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20** 1768–1862.

STONE, C. J. (1975). Adaptive maximum likelihood estimation of a location parameter. *Ann. Statist.* **3** 267–284.

VAN DER VAART, A. W. (1988a). *Statistical Estimation in Large Parameter Spaces. CWI Tract* **44**. Math. Centrum, Amsterdam.

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

VAN DER VAART, A. W. (1988c). Estimating a real parameter in a class of semiparametric models. *Ann. Statist.* **16** 1450–1474.

VAN DER VAART, A. W. (1993). New Donsker classes. *Ann. Probab.* To appear.

VAN DER VAART, A. W. (1994). Bracketing smooth functions. *Stochastic Process. Appl.* **52** 93–105.

WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20** 595–601.

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