The Annals of Statistics
1988, Vol. 16, No. 4, 1450–1474

ESTIMATING A REAL PARAMETER IN A CLASS OF SEMIPARAMETRIC MODELS

BY A. W. VAN DER VAART

University of Leiden

We study semiparametric models where for a fixed value of the finite-dimensional parameter there exists a sufficient statistic for the nuisance parameter. An asymptotically normal sequence of estimators for the parametric component is constructed, which is efficient under the assumption that projecting on the set of nuisance scores is equivalent to taking conditional expectations given the sufficient statistic. The latter property is checked for a number of examples, in particular for mixture models. We discuss the relation of our approach to conditional maximum likelihood estimation.

1. Introduction. Let $X_1, X_2, \ldots, X_n$ be independent random elements, X_j having a density $p_j(\cdot, \theta, \eta)$ with respect to a σ-finite measure μ on a measurable space $(\mathscr{X}, \mathscr{B})$. Here the parameter of interest θ belongs to an open subset Θ of $\mathbb{R}$, and $\eta \in H$ is arbitrary. We assume that the $p_j(\cdot, \theta, \eta)$ have the following structure. For every $(\theta, \eta) \in \Theta \times H$, there exist measurable functions $h_j(\cdot, \theta)$ and $\psi_j(\cdot, \theta)$: $(\mathscr{X}, \mathscr{B}) \to \mathbb{R}$, $j = 1, 2, \ldots, n$, and $g(\cdot, \theta, \eta)$: $\mathbb{R} \to \mathbb{R}$ and a measure ν_θ on $\mathbb{R}$ with

$$p_j(\cdot, \theta, \eta) = h_j(\cdot, \theta) g\big(\psi_j(\cdot, \theta), \theta, \eta\big) \quad \text{a.e. } [\mu], \tag{1.1}$$

$$\psi\big(X_j, \theta\big) \text{ has density } g(\cdot, \theta, \eta) \text{ w.r.t. } \nu_\theta. \tag{1.2}$$

Of course, this means that for every fixed θ and j, $\psi_j(X_j, \theta)$ is sufficient for $\eta \in H$ (with respect to X_j), and the n sufficient statistics are i.i.d. real-valued random variables by (1.2). These two ways of characterizing the model will be used interchangeably.

A number of interesting examples have this structure and are described in Sections 5–7. In Sections 2 and 3 of this paper we aim at the construction of estimators for θ, based on $X_1, \ldots, X_n$, for the general model given by (1.1)–(1.2), where we are particularly interested in obtaining asymptotically efficient sequences of estimators ($n \to \infty$). A sufficient condition for asymptotic efficiency is given in Section 4 and later checked for the examples.

We now give an informal discussion of the paper. As a starting point we take the *score* for θ, which in this Introduction is defined as $\sum_{j=1}^n \dot{l}_j(x_j, \theta, \eta)$, where

$$\dot{l}_j(x, \theta, \eta) = \partial/\partial\theta \log p_j(x, \theta, \eta).$$

As is well known, Fisher's information, $E_{\theta\eta}(\sum_{j=1}^n \dot{l}_j(X_j, \theta, \eta))^2$, measures how well θ can be estimated when η is known, i.e., when H consists of a single element. In

Received October 1986; revised February 1988.

AMS 1980 *subject classifications.* 62F12, 62F35, 62F10, 62G05, 62G20.

Key words and phrases. Semiparametric model, asymptotic efficient estimation, adaptation, mixture model, conditional maximum likelihood.

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

the situation that η is unknown, we expect the information for θ to be smaller. As in Begun, Hall, Huang and Wellner (1983) the loss in information for θ results from a loss in the score function. First, *score functions in the η-direction* are defined as $\sum_{j=1}^n b_j(x_j)$, where $b_j(x)$ is the derivative of the log density along a suitable sequence $\{\eta_t\} \subset H$ (independent of $j = 1, 2, \ldots, n$) in the sense of

$$b_j(x) = \partial/\partial t_{|t=0} \log p(x, \theta, \eta_t),$$

$\eta_0 = \eta$. We remark here that for mathematical rigor, scores can better be defined as derivatives in quadratic mean of the root density. Scores in that sense are introduced in Section 2 and underlie the results of the paper.

Having defined the score for θ and a set of scores for η, the efficient (or "effective") information for θ is defined as $nI_{ne}(\theta, \eta)$, where

$$I_{ne}(\theta, \eta) = n^{-1} \inf E_{\theta\eta} \left(\sum_{j=1}^n \dot{l}_j(X_j, \theta, \eta) - \sum_{j=1}^n b_j(X_j) \right)^2. \tag{1.3}$$

Here the infimum is taken over all η-scores $\sum_{j=1}^n b_j(x_j)$. Informally, asymptotically efficient estimator sequences $\{T_n\}$, where $T_n = t_n(X_1, \ldots, X_n)$, are characterized by the property that $\mathcal{L}_{\theta\eta}(\sqrt{n}(T_n - \theta))$ is approximately $N(0, I_{ne}^{-1}(\theta, \eta))$ for large n. This statement can be made precise in the sense of a convolution and local asymptotic minimax (LAM) theorem, as is explained for i.i.d. models in Begun, Hall, Huang and Wellner (1983).

Because of (1.1) it is clear that in the present model scores for η have the form

$$\sum_{j=1}^n \mathbf{b}(\psi_j(x_j, \theta)),$$

for some function $\mathbf{b}$ not depending on j. Inserting this in (1.3), we obtain

$$I_{ne}(\theta, \eta) = n^{-1} \inf E_{\theta\eta} \left(\sum_{j=1}^n \dot{l}_j(X_j, \theta, \eta) - \sum_{j=1}^n \mathbf{b}(\psi_j(X_j, \theta)) \right)^2, \tag{1.4}$$

where the infimum is now taken over a set $\mathbf{B}(\theta, \eta)$ of functions $\mathbf{b}$. In many interesting examples having the structure (1.1)–(1.2), this set turns out to be the set of all functions $\mathbf{b}$ for which the expression makes sense. It can be checked that in the latter case the infimum in (1.4) is taken for $\mathbf{b}(s) = \bar{l}_n(s, \theta, \eta)$ given by

$$\bar{l}_n(s, \theta, \eta) = n^{-1} \sum_{j=1}^n E_\theta[\dot{l}_j(X_j, \theta, \eta) | \psi_j(X_j, \theta) = s]. \tag{1.5}$$

Of course, under the much weaker condition

$$\bar{l}_n(s, \theta, \eta) \in \mathbf{B}(\theta, \eta), \tag{1.6}$$

this still goes through.

Let

$$\tilde{l}_{nj}(\cdot, \theta, \eta) = \dot{l}_j(\cdot, \theta, \eta) - \bar{l}_n(\psi_j(\cdot, \theta), \theta, \eta). \tag{1.7}$$

Then under (1.6), $\sum_{j=1}^n \tilde{l}_{nj}(x_j, \theta, \eta)$ can be considered the *efficient score function*

for θ. An estimator sequence $\{T_n\}$ is *asymptotically efficient* for θ if it satisfies

$$\sqrt{n}(T_n - \theta) = n^{-1/2} \sum_{j=1}^{n} \tilde{I}_n^{-1}(\theta, \eta)\tilde{l}_{nj}(X_j, \theta, \eta) + o_{P_{\theta\eta}}(1), \tag{1.8}$$

where $\tilde{I}_n(\theta, \eta) = n^{-1}\sum_{j=1}^{n} E_{\theta\eta}\tilde{l}_{nj}^2(X_j, \theta, \eta)$.

One idea to obtain T_n would be to define it by the estimating equation

$$\sum_{j=1}^{n} \tilde{l}_{nj}(X_j, T_n, \eta) = 0. \tag{1.9}$$

Indeed, the usual arguments invoking a Taylor expansion would imply (1.8). However, as η is unknown, (1.9) cannot serve as an estimating equation defining T_n. A way around this problem is to replace $\tilde{l}_{nj}(\cdot, \theta, \eta)$ in (1.9) by an estimated version $\hat{l}_{nj}(\cdot, \theta)$ and to solve for T_n from

$$\sum_{j=1}^{n} \hat{l}_{nj}(X_j, T_n) = 0. \tag{1.10}$$

This route will be followed, though with some modifications. First, handling (1.10) by way of a Taylor expansion requires quite a number of regularity conditions. Now it is usually possible to obtain an accurate initial estimate $\hat{\theta}_n$ for θ. Using $\hat{\theta}_n$ as the starting point for solving (1.10) by the Newton–Raphson scheme, we obtain as a second estimate

$$T_n = \hat{\theta}_n + n^{-1} \sum_{j=1}^{n} \hat{I}_n^{-1}(\hat{\theta}_n)\hat{l}_{nj}(X_j, \hat{\theta}_n). \tag{1.11}$$

Here $\hat{I}_n(\hat{\theta}_n)$ should estimate $\tilde{I}_n(\theta, \eta)$. Next we forget about the foregoing motivation and *define* T_n by (1.11), choosing a convenient estimator $\hat{I}_n(\hat{\theta}_n)$ for $\tilde{I}_n(\theta, \eta)$. It turns out that this *one-step method* works well if $\{\mathcal{L}_{\theta\eta}(\sqrt{n}(\hat{\theta}_n - \theta))\}$ is tight, a property which is usually called $\sqrt{n}$*-consistency*. Furthermore, it works particularly well when combined with another trick, *discretization*. This consists of using an initial estimator $\hat{\theta}_n$ for which $\sqrt{n}(\hat{\theta}_n - \theta)$ has a discrete support, the number of support points within each interval $[-M, M]$ being bounded uniformly in n. Any $\sqrt{n}$-consistent estimator can be discretized without destroying $\sqrt{n}$-consistency, by projecting it on a grid with mesh width $n^{-1/2}$. There is little motivation for discretization, except that it is very convenient in the proofs. Indeed, it is that much convenient that there is ample motivation not to try and do without it.

The one-step method and discretization are clever devices introduced by Le Cam to handle maximum likelihood estimators in parametric models. For semiparametric models they have to be complemented with a method for estimating $\tilde{l}_{nj}(\cdot, \theta, \eta)$, for given θ. In the special model determined by (1.1)–(1.2) this is usually possible. Suppose that ν_θ in (1.2) is Lebesgue measure and that $g(\cdot, \theta, \eta)$ is smooth. We have that

$$\begin{aligned} \dot{l}_j(x, \theta, \eta) = {} & \dot{h}_j/h_j(x, \theta) + \dot{\psi}_j(x, \theta)g'/g\big(\psi_j(x, \theta), \theta, \eta\big) \\ & + \dot{g}/g\big(\psi_j(x, \theta), \theta, \eta\big). \end{aligned} \tag{1.12}$$

Here g' is the derivative with respect to s of $g(s, \theta, \eta)$ and $\dot{h}_j$, $\dot{g}$ and $\dot{\psi}_j$ are partial derivatives with respect to θ. Hence [cf. (1.7) and (1.5)]

$$\tilde{l}_{nj}(x, \theta, \eta) = \tilde{H}_{nj}(x, \theta) + \tilde{\psi}_{nj}(x, \theta) g'/g\big(\psi_j(x, \theta), \theta, \eta\big), \tag{1.13}$$

where

$$\tilde{H}_{nj}(x, \theta) = \dot{h}_j/h_j(x, \theta) - n^{-1} \sum_{i=1}^{n} E_\theta\big(\dot{h}_i/h_i(X_i, \theta) | \psi_i(X_i, \theta) = \psi_j(x, \theta)\big), \tag{1.14}$$

$$\tilde{\psi}_{nj}(x, \theta) = \dot{\psi}_j(x, \theta) - n^{-1} \sum_{i=1}^{n} E_\theta\big(\dot{\psi}_i(X_i, \theta) | \psi_i(X_i, \theta) = \psi_j(x, \theta)\big). \tag{1.15}$$

The key to the construction of an estimate $\hat{l}_{nj}(\cdot, \theta)$ for $\tilde{l}_{nj}(\cdot, \theta, \eta)$ is that (1.13) depends on η only through g'/g. Now, for given θ, $\psi(X_1, \theta), \ldots, \psi(X_n, \theta)$ is an i.i.d. sample from the distribution with density $g(\cdot, \theta, \eta)$. The *kernel method* gives an estimate

$$\hat{g}(s, \theta) = n^{-1} \sum_{j=1}^{n} \sigma_n^{-1} \omega\big(\sigma_n^{-1}\big(s - \psi_j(X_j, \theta)\big)\big), \tag{1.16}$$

for $g(s, \theta, \eta)$, where the kernel ω is a probability density on $\mathbb{R}$. Then $\hat{g}'/\hat{g}(s, \theta)$ should estimate $g'/g(s, \theta, \eta)$, and substituting this in (1.13), we get a candidate for $\hat{l}_{nj}(\cdot, \theta)$. In the present paper we restrict ourselves to kernel estimators, but of course other estimators, perhaps better tuned to the special structure of $g(\cdot, \theta, \eta)$, could perform the same role.

Estimating a *location score* g'/g is a problem with a long history and appears in many constructions of adaptive estimators for the centre of symmetry of a distribution on $\mathbb{R}$ [cf. Stone (1975) and Bickel (1982) and references cited therein]. In the present model finding a suitable candidate for g'/g is complicated by the factor $\tilde{\psi}_{nj}(\cdot, \theta)$ appearing in (1.13). A construction of such a candidate is given in Section 4 under weak conditions, which show up as natural in the examples, but necessitated a long and tedious proof (for which we refer to a technical report).

We note that the fact that g is the density of the sufficient statistic is of crucial importance. Thus far we have used the presence of the sufficient statistic $\psi_j(X_j, \theta)$ both as a means to ensure a special form of the scores for the nuisance parameter and to suggest the possibility of estimating $\tilde{l}_{nj}(\cdot, \theta, \eta)$. Sufficiency also plays a most important role to ensure unbiasedness of the scores $\tilde{l}_{nj}(\cdot, \theta, \eta)$ with respect to the nuisance parameter. Indeed, for all $(\theta, \eta, \eta') \in \Theta \times H \times H$,

$$E_{\theta\eta'} \sum_{j=1}^{n} \tilde{l}_{nj}(X_j, \theta, \eta) = 0. \tag{1.17}$$

The importance of (1.17) is clear by reference to the general set-up for constructing one-step estimators with estimated score functions of Klaassen (1987) and Schick (1986). These authors require an estimator for the score function to be

both consistent and asymptotically unbiased. Relation (1.17) implies that in our special model we need only worry about consistency. In fact, exploiting the sufficiency structure still further, we shall be able to improve upon the general construction methods and use all (except one) of the observations to obtain $\hat{l}_{nj}(\cdot, \theta)$.

An important class of examples of the model (1.1)–(1.2) is given by mixtures over exponential families. In connection to these examples the importance of the score function $\tilde{l}_{nj}(\cdot, \theta, \eta)$ has been noted by Lindsay (1983). The main novelty of the present paper is the introduction of the estimators $\hat{l}_{nj}(\cdot, \theta)$, which allow adaptation to the underlying distribution, with as a result asymptotically improved, indeed efficient, estimators. In addition, we essentially show *compact differentiability* of $p_j(\cdot, \theta, \eta)$ in (θ, η), by means of which we establish the existence of a least favourable submodel in the direction of $\tilde{l}_{nj}(\cdot, \theta, \eta)$ in the sense of Begun, Hall, Huang and Wellner (1983).

In the case that the sufficient statistics are independent of θ we have that $\tilde{\psi}_{nj}(\cdot, \theta) = 0$, so that $\tilde{l}_{nj}(\cdot, \theta, \eta)$ is independent of η. Clearly, estimation of $g'/g(\cdot, \theta, \eta)$ is unnecessary. The estimator constructed in Section 2 is now a one-step version of a *conditional maximum likelihood estimator*, discussed by Andersen (1970). Efficiency of conditional maximum likelihood estimators in mixture models has been shown by Pfanzagl (1982), Chapter 14.

The paper is organized as follows. In Section 2 we introduce score functions in a more formal manner and give the first part of the construction of an estimator of θ, assuming a suitable estimator for the score function g'/g given. A construction of such an estimator for g'/g is given in Section 3. Next we formulate a convolution and local asymptotic minimax (LAM) theorem in Section 4, together with a sufficient condition for efficiency of the estimator of Sections 2 and 3. Examples of model (1.1)–(1.2) can be found in Sections 5–7. Here Section 5 is concerned with mixture models and contains concrete examples as well as some general results.

The results obtained here can be extended in many directions. For instance the constructions go through for the parameter θ ranging through $\mathbb{R}^k$ and sufficient statistics with values in a general Euclidean space. This is shown in van der Vaart (1988), where also a companion model is discussed where the marginal distributions of the sufficient statistics are allowed to depend on j, whereas the conditional distributions are fixed. In the case of mixture models this leads to adaptively constructed estimators of the structural parameter in models with infinitely many nuisance parameters (so-called functional models).

For efficiency of our estimator in the case of mixtures over an exponential family, it makes a difference whether the support of the mixing distribution contains a limit point or not. In the present paper we limit ourselves mainly to the first case. However, even for a finite discrete support our estimator is usually LAM and a best regular estimator. This result follows from an extension of the LAM and convolution theorem, established in van der Vaart (1986b), involving the relaxation of condition (S) in Begun, Hall, Huang and Wellner (1983).

2. Construction of an estimator. Let $L_2(p_j(\cdot, \theta, \eta))$ be the Hilbert space of measurable functions g: $(\mathcal{X}, \mathcal{B}) \to \mathbb{R}$ with $\int g^2(x) p_j(x, \theta, \eta)\, d\mu(x) < \infty$ and

define $L_2(g(\cdot, \theta, \eta))$ and so on analogously. Furthermore, let the addition of an asterisk (*) mean zero expectation; thus any $b \in L_{2*}(g(\cdot, \theta, \eta))$ has $\int b(s)g(s, \theta, \eta)\, d\nu_\theta(s) = 0$.

Instead of defining scores as pointwise derivatives as in the Introduction, we assume from now on that scores for θ exist as elements $\dot{l}_j(\cdot, \theta, \eta)$ of $L_{2*}(p_j(\cdot, \theta, \eta))$ satisfying for $\theta_n - \theta = O(n^{-1/2})$ and every $\varepsilon > 0$,

$$\sum_{j=1}^{n} \int \Big[p_j^{1/2}(x, \theta_n, \eta) - p_j^{1/2}(x, \theta, \eta) - \tfrac{1}{2}(\theta_n - \theta)'\dot{l}_j(x, \theta, \eta) p_j^{1/2}(x, \theta, \eta) \Big]^2 d\mu(x) \to 0, \tag{2.1}$$

$$I_n(\theta, \eta) = n^{-1} \sum_{j=1}^{n} \int \dot{l}_j^2(x, \theta, \eta) p_j(x, \theta, \eta)\, d\mu(x) = O(1), \tag{2.2}$$

$$n^{-1} \sum_{j=1}^{n} \int \dot{l}_j^2(x, \theta, \eta) 1_{\{|\dot{l}_j(x, \theta, \eta)| \geq \varepsilon\sqrt{n}\}} p_j(x, \theta, \eta)\, d\mu(x) \to 0. \tag{2.3}$$

Next we define $\tilde{l}_n$ and $\tilde{l}_{nj}$ by (1.5) and (1.7), respectively, and assume that

$$n^{-1} \sum_{j=1}^{n} \int \tilde{l}_{nj}^2(x, \theta_n, \eta) 1_{\{|\tilde{l}_{nj}(x, \theta_n, \eta)| \geq \varepsilon\sqrt{n}\}} p_j(x, \theta_n, \eta)\, d\mu(x) \to 0, \tag{2.4}$$

$$n^{-1} \sum_{j=1}^{n} \int \Big[\tilde{l}_{nj}(x, \theta_n, \eta) p_j^{1/2}(x, \theta_n, \eta) - \tilde{l}_{nj}(x, \theta, \eta) p_j^{1/2}(x, \theta, \eta) \Big]^2 d\mu(x) \to 0, \tag{2.5}$$

$$\liminf_{n \to \infty} \tilde{I}_n(\theta, \eta) > 0, \tag{2.6}$$

where $\tilde{I}_n(\theta, \eta) = n^{-1}\sum_{j=1}^{n} \int \tilde{l}_{nj}^2(x, \theta_n, \eta) p_j(x, \theta_n, \eta)\, d\mu(x)$. We note that if $X_1, \ldots, X_n$ are i.i.d., then (2.2)–(2.4) are implied by the other assumptions and (2.1)–(2.6) simplify to (5.3)–(5.5) (where unnecessary indices have been deleted).

Now $\dot{l}_j(\cdot, \theta, \eta)$ is defined by (2.1), we do no longer have the decomposition (1.12). However, motivated by (1.12)–(1.15), we assume the existence of measurable functions $\tilde{H}_{nj}(\cdot, \theta)$ and $\tilde{\psi}_{nj}(\cdot, \theta)$: $(\mathscr{X}, \mathscr{B}) \to \mathbb{R}$ and $Q(\cdot, \theta, \eta)$: $\mathbb{R} \to \mathbb{R}$ such that

$$\tilde{l}_{nj}(x, \theta, \eta) = \tilde{H}_{nj}(x, \theta) + \tilde{\psi}_{nj}(x, \theta) Q\big(\psi_j(x, \theta), \theta, \eta\big), \tag{2.7}$$

$$\sum_{j=1}^{n} E_\theta\big(\tilde{\psi}_{nj}(X_j, \theta) | \psi_j(X_j, \theta) = s\big) = 0. \tag{2.8}$$

The one-step method requires an initial estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \ldots, X_n)$ which is $\sqrt{n}$*-consistent*, i.e.,

$$\big\{ \mathscr{L}_{\theta\eta}\big(\sqrt{n}\,(\hat{\theta}_n - \theta)\big) \big\} \text{ is tight on } \mathbb{R}. \tag{2.9}$$

Finally, we need a suitable estimator for $Q(\cdot, \theta, \eta)$. For every fixed θ we assume the existence of measurable functions $Q_n(s, \theta, v_1, \ldots, v_{n-1})$: $\mathbb{R} \times \mathbb{R}^{n-1} \to \mathbb{R}$ such that for an i.i.d. sample $V_1, V_2, \ldots, V_n$ from the distribution with density $g(\cdot, \theta, \eta)$

with respect to ν_θ, and $\theta_n - \theta = O(n^{-1/2})$,

$$E_{\theta_n\eta}\int [Q_n(s,\theta_n,V_1,\ldots,V_{n-1}) - Q(s,\theta_n,\eta)]^2 \times \beta_n^2(s,\theta_n)g(s,\theta_n,\eta)\,d\nu_{\theta_n}(s) \to 0. \tag{2.10}$$

Here

$$\beta_n(s,\theta) = \left\{n^{-1}\sum_{j=1}^n E_\theta\big(\tilde{\psi}_{nj}^2(X_j,\theta)|\psi_j(X_j,\theta) = s\big)\right\}^{1/2}. \tag{2.11}$$

Condition (2.10) will be treated in detail in the next section. As for $\sqrt{n}$-consistent estimators, it is usually not too difficult to find candidates in specific models. General methods that may work are the following. For a fixed, conveniently chosen $\eta' \in H$, one may try defining an estimator $\hat{\theta}_n$ as the solution to

$$\sum_{j=1}^n \tilde{l}_{nj}(X_j,\theta,\eta') = 0,$$

which is an unbiased estimating equation by (1.17). In the same spirit it may work to solve for $\hat{\theta}_n$ from

$$\sum_{j=1}^n \tilde{H}_{nj}\big(X_j,\hat{\theta}_n\big) = 0.$$

The main result of this section is

THEOREM 2.1. *Let* (1.1)–(1.2) *and* (2.1)–(2.10) *hold. Then there exists an estimator sequence* $\{T_n\}$ *satisfying* (1.8). *Under the assumptions this implies*

$$\mathcal{L}_{\theta_n\eta_n}\big(\sqrt{n}\,\tilde{I}_n^{1/2}(\theta,\eta)(T_n - \theta_n)\big) \to N(0,1),$$

for every sequence $\{(\theta_n,\eta_n)\}$ *in* $\Theta \times H$ *such that*

$$n^{-1}\sum_{j=1}^n \int\Big[\sqrt{n}\big(p_j^{1/2}(x,\theta_n,\eta_n) - p_j^{1/2}(x,\theta,\eta)\big) - \tfrac{1}{2}g_{nj}(x)p_j^{1/2}(x,\theta,\eta)\Big]^2 d\mu(x) \to 0,$$

for a triangular array $\{g_{nj}\}$, *where* $g_{nj} \in L_2(p_j(\cdot,\theta,\eta))$, *satisfying*

$$g_{nj}(x) = \sqrt{n}\,(\theta_n - \theta)\dot{l}_j(x,\theta,\eta) - b_n\big(\psi_j(x,\theta)\big),$$

$$n^{-1}\sum_{j=1}^n \int g_{nj}^2(x)p_j(x,\theta,\eta)\,d\mu(x) = O(1),$$

$$n^{-1}\sum_{j=1}^n \int g_{nj}^2(x)1_{\{|g_{nj}(x)| \ge \varepsilon\sqrt{n}\}}p_j(x,\theta,\eta)\,d\mu(x) \to 0.$$

The last assertion of the theorem implies that $\{T_n\}$ is *regular*. We come back to this in Section 4, when discussing efficiency.

A candidate for T_n can be constructed as follows. Let $V_j(\theta) = \psi_j(X_j, \theta)$ and assume without loss of generality that $Q_n(s, \theta, v_1, \ldots, v_{n-1})$ is symmetric in $v_1, \ldots, v_{n-1}$. Let

$$\hat{Q}_n^j(s, \theta) = Q_n\big(s, \theta, V_1(\theta), \ldots, V_{j-1}(\theta), V_{j+1}(\theta), \ldots, V_n(\theta)\big), \tag{2.12}$$

$$\hat{l}_{nj}^j(x, \theta) = \tilde{H}_{nj}(x, \theta) + \tilde{\psi}_{nj}(x, \theta)\hat{Q}_n^j\big(\psi_j(x, \theta), \theta\big), \tag{2.13}$$

$$\hat{I}_n(\theta) = n^{-1/2} \sum_{j=1}^{n} \Big(\hat{l}_{nj}^j\big(X_j, \theta - n^{-1/2}\big) - \hat{l}_{nj}^j\big(X_j, \theta\big)\Big). \tag{2.14}$$

Now, let $\hat{\theta}_n$ be a discretized, $\sqrt{n}$-consistent estimator for θ and set

$$T_n = \hat{\theta}_n + n^{-1} \sum_{j=1}^{n} \hat{I}_n^{-1}\big(\hat{\theta}_n\big)\hat{l}_{nj}^j\big(X_j, \hat{\theta}_n\big), \tag{2.15}$$

whenever $\hat{I}_n(\hat{\theta}_n)$ is positive, and 0 otherwise.

The proof of Theorem 2.1 is accomplished through a series of lemmas. Here we shall use that (2.1)–(2.3) imply contiguity of the laws of $(X_1, X_2, \ldots, X_n)$ under (θ_n, η) and (θ, η) if $\theta_n - \theta = O(n^{-1/2})$, so that convergence to 0 of a function of $(X_1, X_2, \ldots, X_n)$ in $P_{\theta\eta}$-probability is equivalent to convergence to 0 in $P_{\theta_n\eta}$-probability.

The first lemma contains the main part of the technical work. Its proof can be found in Appendix A.2 of van der Vaart (1988).

LEMMA 2.1. *Let* (2.1)–(2.6) *hold. Then for* $\theta_n - \theta = O(n^{-1/2})$,

$$n^{-1/2} \sum_{j=1}^{n} \Big(\tilde{l}_{nj}\big(X_j, \theta_n, \eta\big) - \tilde{l}_{nj}\big(X_j, \theta, \eta\big)\Big) + \tilde{I}_n(\theta, \eta)\sqrt{n}\,(\theta_n - \theta) \to_{P_{\theta\eta}} 0.$$

The second lemma uses the sufficiency structure of the model in an essential way.

LEMMA 2.2. *Under the conditions of Theorem* 2.1, *for* $\theta_n - \theta = O(n^{-1/2})$,

$$E_{\theta_n\eta}\left[n^{-1/2} \sum_{j=1}^{n} \Big(\hat{l}_{nj}^j\big(X_j, \theta_n\big) - \tilde{l}_{nj}\big(X_j, \theta_n, \eta\big)\Big)\right]^2 \to 0.$$

PROOF. Write $V_j(\theta)$ and $\tilde{V}_{nj}(\theta)$ for $\psi(X_j, \theta)$ and $\tilde{\psi}_{nj}(X_j, \theta)$, respectively. By (2.12)–(2.13) and (2.7) we must show convergence to 0 of

$$E_{\theta_n\eta}\left[n^{-1/2} \sum_{j=1}^{n} \tilde{V}_{nj}(\theta_n)\big(\hat{Q}_n^j\big(V_j(\theta_n), \theta_n\big) - Q\big(V_j(\theta_n), \theta_n, \eta\big)\big)\right]^2$$

$$= n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} E_{\theta_n\eta}\tilde{V}_{ni}(\theta_n)\tilde{V}_{nj}(\theta_n)\big[\hat{Q}_n^i\big(V_i(\theta_n), \theta_n\big) - Q\big(V_i(\theta_n), \theta_n, \eta\big)\big]$$

$$\times\big[\hat{Q}_n^j\big(V_j(\theta_n), \theta_n\big) - Q\big(V_j(\theta_n), \theta_n, \eta\big)\big].$$

Taking first the conditional expectation with respect to $V_1(\theta_n), \ldots, V_n(\theta_n)$ and

remembering that $[\hat{Q}_n^j(V_j(\theta_n), \theta_n) - Q(V_j(\theta_n), \theta_n, \eta)]$ depends on $V_1(\theta_n), \ldots, V_n(\theta_n)$ only, we see that this equals

$$n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} E_{\theta_n \eta} E_{\theta_n}\big(\tilde{V}_{ni}(\theta_n)\tilde{V}_{nj}(\theta_n) | V_i(\theta_n), V_j(\theta_n)\big)\alpha_n^{ij}\big(V_i(\theta_n), V_j(\theta_n)\big), \tag{2.16}$$

where

$$\alpha_n^{ij}(s,t) = E_{\theta_n \eta}\Big(\big[\hat{Q}_n^i(V_i(\theta_n), \theta_n) - Q(V_i(\theta_n), \theta_n, \eta)\big] \times \big[\hat{Q}_n^j(V_j(\theta_n), \theta_n) - Q(V_j(\theta_n), \theta_n, \eta)\big] | V_i(\theta_n) = s, V_j(\theta_n) = t\Big).$$

The sums of the diagonal terms in (2.16) equals

$$n^{-1} \sum_{j=1}^{n} \int E_{\theta_n}\big(\tilde{V}_{nj}^2(\theta_n) | V_j(\theta_n) = s\big)\alpha_n^{jj}(s,s) g(s, \theta_n, \eta)\, d\nu_{\theta_n}(s)$$

$$= \int \beta_n^2(s, \theta_n) E_{\theta_n \eta}\big[\hat{Q}_n^1(s, \theta_n) - Q(s, \theta_n, \eta)\big]^2 g(s, \theta_n, \eta)\, d\nu_{\theta_n}(s) \to 0$$

by (2.10).

The sum of the off-diagonal terms in (2.16) equals

$$n^{-1} \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} \int\int E_{\theta_n}\big(\tilde{V}_{ni}(\theta_n) | V_i(\theta_n) = s\big) E_{\theta_n}\big(\tilde{V}_{nj}(\theta_n) | V_j(\theta_n) = t\big) \times \alpha_n^{12}(s,t) g(s, \theta_n, \eta) g(t, \theta_n, \eta)\, d\nu_{\theta_n}(s)\, d\nu_{\theta_n}(t).$$

By (2.8) this is equal to

$$-n^{-1} \sum_{j=1}^{n} \int\int E_{\theta_n}\big(\tilde{V}_{nj}(\theta_n) | V_j(\theta_n) = s\big) E_{\theta_n}\big(\tilde{V}_{nj}(\theta_n) | V_j(\theta_n) = t\big) \times \alpha_n^{12}(s,t) g(s, \theta_n, \eta) g(t, \theta_n, \eta)\, d\nu_{\theta_n}(s)\, d\nu_{\theta_n}(t),$$

which, by using the Cauchy–Schwarz inequality on the double integral, can be dominated in absolute value by

$$n^{-1} \sum_{j=1}^{n} E_{\theta_n \eta} \int\int \big[Q_n(s, \theta_n, t, V_3(\theta_n), \ldots, V_n(\theta_n)) - Q(s, \theta_n, \eta)\big]^2 \times E_{\theta_n}\big(\tilde{V}_{nj}(\theta_n) | V_j(\theta_n) = s\big)^2 g(s, \theta_n, \eta) g(t, \theta_n, \eta)\, d\nu_{\theta_n}(s)\, d\nu_{\theta_n}(t) \tag{2.17}$$

$$\leq E_{\theta_n \eta} \int \beta_n^2(s, \theta_n)\big[\hat{Q}_n^1(s, \theta_n) - Q(s, \theta_n, \eta)\big]^2 g(s, \theta_n, \eta)\, d\nu_{\theta_n}(s),$$

which converges to 0, again by (2.10). □

PROOF OF THEOREM 2.1. We show first that for $\theta_n - \theta = O(n^{-1/2})$,

$$\hat{I}_n(\theta_n)\tilde{I}_n^{-1}(\theta, \eta) \to_{P_{\theta\eta}} 1. \tag{2.18}$$

Indeed, by (2.14) and Lemma 2.2 the left-hand side of (2.18) equals

$$\left[n^{-1/2}\sum_{j=1}^{n}\left(\tilde{l}_{nj}\left(X_j,\theta_n-n^{-1/2},\eta\right)-\tilde{l}_{nj}(X_j,\theta_n,\eta)\right)+o_{P_{\theta\eta}}(1)\right]\tilde{I}_n^{-1}(\theta,\eta)$$
$$=1+o_{P_{\theta\eta}}(1)$$

by Lemma 2.1.

For T_n given by (2.15) we have

$$P_{\theta\eta}\left(\left|\sqrt{n}\,(T_n-\theta)-n^{-1/2}\sum_{j=1}^{n}\tilde{I}_n^{-1}(\theta,\eta)\tilde{l}_{nj}(X_j,\theta,\eta)\right|\geq\varepsilon\right)$$
$$\leq P_{\theta\eta}\left(|\sqrt{n}\,(\hat{\theta}_n-\theta)|\geq M\right)$$
$$+\sum P_{\theta\eta}\left(\left|\sqrt{n}\,(\theta_n-\theta)+n^{-1/2}\sum_{j=1}^{n}\left(\hat{I}_n^{-1}(\theta_n)\hat{l}_{nj}^{j}(X_j,\theta_n)\right.\right.\right.$$
$$\left.\left.\left.-\tilde{I}_n^{-1}(\theta,\eta)\tilde{l}_{nj}(X_j,\theta,\eta)\right)\right|\geq\varepsilon\right)+o(1), \tag{2.19}$$

where the sum is over the set of $\theta_n\in\mathbb{R}$ in the support of $\hat{\theta}_n$ with $\sqrt{n}\,|\theta_n-\theta|\leq M$. By $\sqrt{n}$-consistency of $\hat{\theta}_n$, M can be chosen such that the first term in (2.19) is arbitrarily small. Then, as $\hat{\theta}_n$ is discretized, the number of terms in the sum is finite and bounded uniformly in n, and it suffices to prove that the maximum over the terms converges to 0. This would follow if for any sequence of numbers $\{\theta_n\}$ in $\mathbb{R}$ with $\theta_n-\theta=O(n^{-1/2})$,

$$\sqrt{n}\,(\theta_n-\theta)+n^{-1/2}\sum_{j=1}^{n}\left(\hat{I}_n^{-1}(\theta_n)\hat{l}_{nj}^{j}(X_j,\theta_n)-\tilde{I}_n^{-1}(\theta,\eta)\tilde{l}_{nj}(X_j,\theta,\eta)\right)\to_{P_{\theta\eta}}0.$$

This is a consequence of Lemmas 2.1 and 2.2, (2.18) and the tightness of $\{\mathscr{L}_{\theta\eta}(n^{-1/2}\Sigma_{j=1}^{n}\tilde{l}_{nj}(X_j,\theta_n,\eta))\}$.

The second assertion is a consequence of local asymptotic normality and the third lemma of Le Cam. □

3. Estimation of $g'/g(\cdot,\theta,\eta)$. In this section it is shown that estimators for g'/g needed for the construction in Section 2, typically exist. More precisely, we present a set of sufficient conditions for (2.10), where, motivated by (1.13), $Q(\cdot,\theta,\eta)$ is replaced by $g'/g(\cdot,\theta,\eta)$ and where it is assumed that ν_θ is Lebesgue measure λ on $\mathbb{R}$. The general result presented in the following discussion is purely asymptotical and should foremost be considered an existence result. For application of adaptive methods in practice much work remains to be done.

We assume that $g(\cdot,\theta,\eta)$ is a density with respect to Lebesgue measure, which vanishes outside an interval $(a,b)\subset\mathbb{R}$ [independent of (θ,η)]. Moreover we assume that it is *absolutely continuous* on (a,b) in the sense that there exists a measurable function $g'(\cdot,\theta,\eta)$, also vanishing outside (a,b), such that for

$a < c < d < b$,

$$g(d, \theta, \eta) - g(c, \theta, \eta) = \int_c^d g'(s, \theta, \eta)\, ds. \tag{3.1}$$

We suppose that for every $\theta \in \Theta$ there exists a measurable real function $\beta_\infty(\cdot, \theta)$ on $\mathbb{R}$ such that for $\theta_n - \theta = O(n^{-1/2})$ and $n \to \infty$,

$$\beta_n(s, \theta_n) - \beta_\infty(s, \theta) \to 0 \quad \text{a.e. } [\lambda], \tag{3.2}$$

$$\int \left[g'/g^{1/2}(s, \theta, \eta)\beta_\infty(s, \theta)\right]^2 ds < \infty, \tag{3.3}$$

$$\int \left[g'/g^{1/2}(s, \theta_n, \theta)\beta_n(s, \theta_n) - g'/g^{1/2}(s, \theta, \eta)\beta_\infty(s, \theta)\right]^2 ds \to 0, \tag{3.4}$$

$$g(s, \theta_n, \eta) - g(s, \theta, \eta) \to 0 \quad \text{a.e. } [\lambda]. \tag{3.5}$$

In i.i.d. models one expects $\beta_\infty(s, \theta) = \{E_\theta(\tilde{\psi}^2(X_j, \theta)|\psi(X_j, \theta) = s)\}^{1/2}$, of course. In non-i.i.d. models (3.2) may be restrictive. However, usually it is satisfied along subsequences, which is sufficient for applicability of Theorem 3.1.

Finally, we need that the functions $\beta_n(\cdot, \theta_n)$ satisfy a Lipschitz condition in their first argument. For some $\kappa > 0$ and constants M_s,

$$|\beta_n(s + h, \theta_n) - \beta_n(s, \theta_n)| \le M_s|h|^\kappa \quad \text{a.e. } [\lambda],\ n = 1, 2, \ldots. \tag{3.6}$$

While the first conditions are all natural, condition (3.6) is less transparent. It is relatively weak, though. Also, the condition can be relaxed in the sense that we only need that (3.6) holds for all h in a neighbourhood of a.a. s, which is implied for instance by equidifferentiability of $\{\beta_n(s, \theta_n)\colon n = 1, 2, \ldots\}$ in s a.e.

Let $\omega\colon \mathbb{R} \to \mathbb{R}$ be any twice continuously differentiable probability density with respect to Lebesgue measure, with support contained in $[-1, 1]$. Given $\sigma \in (0, \infty) \subset \mathbb{R}$ and an i.i.d. sample $V_1, V_2, \ldots, V_n$ from $g(\cdot, \theta, \eta)$, define

$$\hat{g}_{n\sigma}(s) = n^{-1} \sum_{i=1}^{n-1} \sigma^{-1}\omega\left(\sigma^{-1}(s - V_i)\right).$$

THEOREM 3.1. *Let* (3.1)–(3.6) *hold. Define*

$$\hat{Q}_n(s, \theta) = \hat{g}'_{n\sigma_n}(s)/\left(\hat{g}_{n\sigma_n}(s) + \delta_n\right)1_{C_n(\theta)}(s),$$

where for sequences $\alpha_n, \gamma_n, \delta_n, \varepsilon_n, \sigma_n \downarrow 0$ *and* $b_n, c_n \to \infty$ *in* $\mathbb{R}$,

$$C_n(\theta) = \Big\{a + \varepsilon_n < s < b - \varepsilon_n,\ \left|\hat{g}'_{n\sigma_n}(s)\beta_n(s, \theta)\right| < c_n\left(\hat{g}_{n\sigma_n}(s) + \delta_n\right),$$
$$\alpha_n < \beta_n(s, \theta) < b_n,\ \sup\{|\beta_n(s + \sigma_n y, \theta) - \beta_n(s, \theta)|\colon |y| < 1\} \le \gamma_n\Big\}.$$

If $\sigma_n\varepsilon_n^{-1} \to 0$, $\gamma_n\alpha_n^{-1} \to 0$, $c_n^2\sigma_n\alpha_n^{-1} \to 0$, $\delta_n^{-2}\sigma_n^{-4}b_n^2n^{-1} \to 0$ *and* $\gamma_n^{-1}\sigma_n^\kappa \to 0$, *then*

$$E_{\theta_n\eta}\int \left[\hat{Q}_n(s, \theta_n) - g'/g(s, \theta, \eta)\right]^2\beta_n^2(s, \theta_n)g(s, \theta_n, \eta)\, d\lambda(s) \to 0.$$

The proof of Theorem 3.1 is long and tedious. We refer to van der Vaart (1988), Section 5.3 or (1986a).

4. Efficiency. In Sections 2 and 3 it is shown how to construct an estimator sequence which is *asymptotically linear* with the influence function of the jth observation equal to $\tilde{l}_{nj}(\cdot, \theta, \eta)$ [cf. (1.8)]. Under the appropriate conditions, this estimator sequence is asymptotically optimal in the sense of the convolution and local asymptotic minimax (LAM) theorem. Statements of slight extensions of existing results [cf. Begun, Hall, Huang, and Wellner (1983)] are included in this section for reference.

The following theorems are based on the assumption that a given *least favourable* one-dimensional submodel exists. Begun, Hall, Huang and Wellner (1983) establish the existence of a least favourable submodel under the condition of joint differentiability of the underlying densities $p_j(\cdot, \theta, \eta)$ with respect to θ and η, and a condition (S) on the form of the set of η-scores. Here we rather choose to assume the existence of a least favourable submodel directly.

It is argued in Section 1 that the one-dimensional submodel in the direction $\tilde{l}_{nj}(\cdot, \theta, \eta)$ is often least favourable. The existence of this submodel is of course sufficient to make the estimator sequence of Sections 2 and 3 efficient. Therefore, we assume

$$\text{for all } h \in \mathbb{R} \text{ there exists } \{\eta_n(h)\}_{n=1}^{\infty} \subset H \text{ such that}$$

$$n^{-1} \sum_{j=1}^{n} \int \Big[p_j^{1/2}(x, \theta_n(h), \eta_n(h)) - p_j^{1/2}(x, \theta, \eta) - \tfrac{1}{2} h \tilde{l}_{nj}(x, \theta, \eta) p_j^{1/2}(x, \theta, \eta) \Big]^2 d\mu(x) \to 0, \tag{4.1}$$

$n \to \infty$. Here $\theta_n(h) = \theta + n^{-1/2}h$. In Sections 5–7 it is shown by examples that condition (4.1) is often satisfied.

The LAM and convolution theorem stated in the following discussion are in their strongest form in the sense that we take the maximum risk and require regularity of the estimator sequence over the least favourable submodel only. Using "full" Hellinger neighbourhoods as in Begun, Hall, Huang and Wellner (1983) is more natural on the one hand, but on the other requires the introduction of more technical detail and, besides, gives weaker results.

PROPOSITION 4.1 (LAM theorem). *Let* (1.1)–(1.2), (2.1)–(2.6) *and* (4.1) *hold. Then for any loss function* $l: \mathbb{R} \to [0, \infty)$ *with*

$$l(x) = l(|x|), \qquad l(|x|) \le l(|y|) \quad \text{if } |x| \le |y|,$$

and any estimator sequence $\{T_n\}$, *we have*

$$\lim_{c \to \infty} \liminf_{n \to \infty} \sup_{|h| < c} E_{\theta_n(h)\eta_n(h)} l\big(\sqrt{n}\, \tilde{I}_n^{1/2}(\theta, \eta)(T_n - \theta_n(h))\big) \ge \int l(x)\, dN(0,1)(x).$$

Moreover, we can have equality for a nonzero loss function l *satisfying* $\int l(x) e^{\eta x}\, dN(0,1)(x) < \infty$ *for some* $\eta > 0$, *only if* $\{T_n\}$ *satisfies* (1.8).

Suppose that (4.1) holds. Then call an estimator sequence $\{T_n\}$ in the model (1.1)–(1.2) regular at (θ, η) if $\{\mathcal{L}_{\theta\eta}(\sqrt{n}(T_n - \theta))\}$ is tight and for any $h \in \mathbb{R}$,

$\{\mathcal{L}_{\theta_n(h),\eta_n(h)}(\sqrt{n}(T_n - \theta_n(h)))\}$ and $\{\mathcal{L}_{\theta\eta}(\sqrt{n}(T_n - \theta))\}$ have the same limit points for the same subsequences.

PROPOSITION 4.2 (Convolution theorem). *Let* (1.1)–(1.2), (2.1)–(2.6) *and* (4.1) *hold. If* $\{T_n\}$ *is regular at* (θ, η) *and* L *a weak limit point of*

$$\left\{\mathcal{L}_{\theta\eta}\left(\sqrt{n}\,\tilde{I}_n^{1/2}(\theta,\eta)(T_n - \theta)\right)\right\},$$

then

$$L = N(0,1) * M,$$

for a probability measure M *on* $\mathbb{R}$.

Propositions 4.1 and 4.2 can be proved by adapting Hájek's (1970, 1972) theorems to the present situation. For this the following *local asymptotic normality* lemma is of crucial importance.

LEMMA 4.1 (LAN). *Let* (1.1)–(1.2), (2.1)–(2.6) *and* (4.1) *hold. Set*

$$\Lambda_n(h) = \log \prod_{j=1}^{n} p_j\left(X_j, \theta_n(h), \eta_n(h)\right)/p_j\left(X_j, \theta, \eta\right),$$

where $\log a/b$ *is* $-\infty$ *if* $a = 0 < b$, $+\infty$ *if* $b = 0 < a$ *and* 0 *if* $a = b = 0$. *Then*

$$\Lambda_n(h) - n^{-1/2} \sum_{j=1}^{n} h\tilde{l}_{nj}\left(X_j, \theta, \eta\right) + \tfrac{1}{2}h^2\tilde{I}_n(\theta,\eta) \to_{P_{\theta\eta}} 0.$$

A final question to be answered is whether the estimator sequence $\{T_n\}$ constructed in the foregoing sections is *regular* in the sense defined previously. In Theorem 2.1 we already noted that $\mathcal{L}_{\theta_n\eta_n}(\sqrt{n}\,\tilde{I}_n^{1/2}(\theta,\eta)(T_n - \theta_n))$ converges to a standard normal distribution along certain sequences $\{(\theta_n, \eta_n)\}$. All we need to check is whether the sequences $\{\theta + n^{-1/2}h, \eta_n(h)\}$ are among these sequences. We state this without proof.

LEMMA 4.2. *Let* (1.1)–(1.2), (2.1)–(2.6) *and* (4.1) *hold. Then the triangular array* $\{\tilde{l}_{nj}(\cdot, \theta, \eta)\}$ *satisfies the conditions imposed on* $\{g_{nj}\}$ *in Theorem* 2.1. *Hence* $\{T_n\}$ *satisfying* (1.8) *is regular.*

5. Mixture models. Important examples of models with the structure (1.1)–(1.2) belong to the class of mixture models. In this section we first derive scores for mixture models in a rigorous manner, which leads to the establishment of a property called local completeness, which implies the sufficient condition for efficiency (4.1) for a large set of mixture models. Next follow several concrete examples.

The models we consider are all i.i.d. models. For convenience of notation we drop all unnecessary indices j and n. In particular (1.1)–(1.2) become

$$p(\cdot, \theta, \eta) = h(\cdot, \theta)g(\psi(\cdot, \theta), \theta, \eta) \quad \text{a.e. } [\mu], \tag{5.1}$$

$$\psi(X_j, \theta) \text{ has density } g(\cdot, \theta, \eta) \quad \text{w.r.t. } \nu_\theta. \tag{5.2}$$

Furthermore, we set

$$\tilde{l}(\cdot, \theta, \eta) = \dot{l}(\cdot, \theta, \eta) - E_\theta\big[\dot{l}(X_1, \theta, \eta) | \psi(X_1, \theta) = \psi(\cdot, \theta)\big].$$

Relations (2.1)–(2.6) reduce to

$$\int \big[t^{-1}\big(p^{1/2}(x, \theta + t, \eta) - p^{1/2}(x, \theta, \eta)\big) - \tfrac{1}{2}\dot{l}(x, \theta, \eta)p^{1/2}(x, \theta, \eta)\big]^2 d\mu(x) \to 0, \tag{5.3}$$

$$\int \big[\tilde{l}(x, \theta + t, \eta)p^{1/2}(x, \theta + t, \eta) - \tilde{l}(x, \theta, \eta)p^{1/2}(x, \theta, \eta)\big]^2 d\mu(x) \to 0, \tag{5.4}$$

$$\tilde{I}(\theta, \eta) = \int \tilde{l}^2(x, \theta, \eta)p(x, \theta, \eta)\, d\mu(x) > 0. \tag{5.5}$$

Finally, the condition (4.1) on the existence of a least favourable submodel simplifies to the existence of $\{\eta_t\} \subset H$, $|t| < 1$, such that

$$\int \big[t^{-1}\big(p^{1/2}(x, \theta + t, \eta_t) - p^{1/2}(x, \theta, \eta)\big) - \tfrac{1}{2}\tilde{l}(x, \theta, \eta)p^{1/2}(x, \theta, \eta)\big]^2 d\mu(x) \to 0 \tag{5.6}$$

as $t \to 0$.

One way to establish (5.6) is to show that any function of $\psi(\cdot, \theta)$ occurs as an η-score (cf. Section 1). This can and will be done for mixture models. The situation is important enough to give it a name.

DEFINITION 5.1. Let (5.1)–(5.3) hold. Then $\psi(X_1, \theta)$ is called (strongly) locally complete at (θ, η) if for any $\mathbf{b} \in L_{2*}(g(\cdot, \theta, \eta))$ there exists $\{\eta_t\} \subset H$ such that for any $h \in \mathbb{R}$,

$$\int \big[t^{-1}\big(p^{1/2}(x, \theta + th, \eta_t) - p^{1/2}(x, \theta, \eta)\big) - \tfrac{1}{2}\big(h\dot{l}(x, \theta, \eta) + \mathbf{b}(\psi(x, \theta))\big)p^{1/2}(x, \theta, \eta)\big]^2 d\mu(x) \to 0 \tag{5.7}$$

as $t \to 0$.

There is no simple relation between ordinary completeness of $\psi(X_1, \theta)$ and local completeness, though under regularity conditions the latter implies the first.

5.1. *Local completeness in mixture models.* We now introduce *mixture models*. Let Θ be an open subset of $\mathbb{R}$ and let H be a collection of probability measures on a measurable space $(\mathscr{Z}, \mathscr{A})$. For each $(\theta, z) \in \Theta \times \mathscr{Z}$ we have a probability density $\mathbf{p}(\cdot, \theta, z)$ with respect to a σ-finite measure μ on a measurable space $(\mathscr{X}, \mathscr{B})$; we assume that $\mathbf{p}(x, \theta, z)$ is measurable as a function of (x, z)

and set

$$p(x, \theta, \eta) = \int \mathbf{p}(x, \theta, z)\, d\eta(z). \tag{5.8}$$

In a *mixture model* $X_1, X_2, \ldots, X_n$ are i.i.d. random elements with density given by (5.8). We shall be concerned with the subclass where there exist measurable functions $h(\cdot, \theta)$ and $\psi(\cdot, \theta)$: $(\mathscr{X}, \mathscr{B}) \to \mathbb{R}$ and $g(\cdot, \theta, z)$: $\mathbb{R} \to \mathbb{R}$ with

$$\mathbf{p}(\cdot, \theta, z) = h(\cdot, \theta)\mathbf{g}(\psi(\cdot, \theta), \theta, z) \quad \text{a.e. } [\mu]. \tag{5.9}$$

Clearly, (5.1)–(5.2) hold with

$$g(s, \theta, \eta) = \int \mathbf{g}(s, \theta, z)\, d\eta(z),$$

and ν_θ defined by

$$\nu_\theta(B) = \int 1_{\{\psi(x,\theta) \in B\}} h(x, \theta)\, d\mu(x).$$

The form taken by scores in mixture models is well known. We now set out to establish this in a rigorous manner. Begun, Hall, Huang, and Wellner (1983) require $p^{1/2}(\cdot, \theta, \eta)$ to be *Fréchet differentiable* in θ and η. Theorem 5.1 essentially asserts under a weak regularity condition that it is *compactly differentiable*, a weaker form of differentiability, which, however, is sufficient to obtain the results of Begun, Hall, Huang and Wellner (1983). For later use it is helpful to introduce a set $T(\eta, H)$ of directions in which η can be approximated within H.

DEFINITION 5.2. Given $\eta \in H$ (a class of probability distributions) $T(\eta, H)$ is the subset of $L_2(\eta)$ consisting of all c for which there exists a sequence $\{\eta_t\} \subset H$ and a σ-finite measure τ with $\eta_t \ll \tau$, $\eta \ll \tau$ and

$$\int \left[t^{-1}\left((d\eta_t/d\tau)^{1/2} - (d\eta/d\tau)^{1/2} \right) - \tfrac{1}{2} c (d\eta/d\tau)^{1/2} \right]^2 d\tau \to 0, \qquad t \downarrow 0. \tag{5.10}$$

It is easily shown that $T(\eta, H) = L_{2*}(\eta)$ in the case that H is the set of all probability measures on $(\mathscr{X}, \mathscr{A})$. Restrictions on the set of η, which may be needed in examples, such as finite moments or absolute continuity, typically do not affect this conclusion.

THEOREM 5.1. *Let* (5.8)–(5.10) *hold and suppose that for measurable functions* $\dot{\mathbf{l}}(x, \theta, z)$: $(\mathscr{X} \times \mathscr{Z}, \mathscr{B} \times \mathscr{A}) \to \mathbb{R}$ *and* $t \to 0$,

$$\int\int \left[t^{-1}\left(\mathbf{p}^{1/2}(x, \theta + t, z) - \mathbf{p}^{1/2}(x, \theta, z) \right) - \tfrac{1}{2} \dot{\mathbf{l}}(x, \theta, z) \mathbf{p}^{1/2}(x, \theta, z) \right]^2 d\mu(x)\, d\eta(z) \to 0. \tag{5.11}$$

Define

$$\mathscr{g}(x) = p^{-1}(x, \theta, \eta) \int \left(\dot{\mathbf{l}}(x, \theta, z) + c(z) \right) \mathbf{p}(x, \theta, z)\, d\eta(z).$$

Then

$$\int \left[t^{-1}\left(p^{1/2}(x, \theta + t, \eta_t) - p^{1/2}(x, \theta, \eta) \right) - \tfrac{1}{2} \dot{g}(x) p^{1/2}(x, \theta, \eta) \right]^2 d\mu(x) \to 0. \tag{5.12}$$

PROOF. See van der Vaart (1988), Theorem 5.13, or van der Vaart (1986a). □

It is important to note that Theorem 5.1 does not always give an exhaustive set of η-scores. This is because we allow only sequences $\{\eta_t\}$ which satisfy (5.10). Since (5.10) implies that the part of η_t which is singular with respect to η, disappears at rate $o(t^2)$, this is especially restrictive when η has discrete support. Indeed, in two examples which follow it turns out that a least favourable one-dimensional submodel as in (5.6) cannot be obtained for a sequence $\{\eta_t\}$ satisfying (5.10), but still does exist.

Using Theorem 5.1, we shall derive a result relating in the case of mixture models, ordinary completeness of $\psi(X_1, \theta)$ in the model given by the $\mathbf{g}(\cdot, \theta, z)$, to local completeness of $\psi(X_1, \theta)$. In view of the applications to follow, we need to work with a slightly weaker form of completeness than ordinary completeness, which we call L_2-completeness.

DEFINITION 5.3. A set of probability distributions $\mathscr{P}$ on a measurable space $(\mathscr{X}, \mathscr{B})$ is called L_2-complete if $b \in L_2(P)$ and $\int b\, dP = 0$ for all $P \in \mathscr{P}$, implies that $b = 0$ P-a.e. for all $P \in \mathscr{P}$.

THEOREM 5.2. *For the mixture model* (5.8)–(5.9), *let* (5.11) *hold. Assume that* $T(\eta, H) = L_{2*}(\eta)$ *and*

$$\{\mathbf{g}(\cdot, \theta, z)\, dv_\theta : z \in A\} \text{ is } L_2\text{-complete for every } A \in \mathscr{A} \text{ with } \int_A d\eta = 1. \tag{5.13}$$

Then $\psi(X_1, \theta)$ *is locally complete at* (θ, η).

PROOF. Given $c \in L_{2*}(\eta)$, we choose $\{\eta_t\} \subset H$, $t > 0$, such that (5.10) holds. By Theorem 5.1 we conclude that

$$\int \left[t^{-1}\left(p^{1/2}(x, \theta + th, \eta_t) - p^{1/2}(x, \theta, \eta) \right) - \tfrac{1}{2}\left(h\dot{l}(x, \theta, \eta) + Ac(\psi(x, \theta)) \right) p^{1/2}(x, \theta, \eta) \right]^2 d\mu(x)$$

converges to 0, where

$$\dot{l}(x, \theta, \eta) = p^{-1}(x, \theta, \eta) \int \dot{\mathbf{l}}(x, \theta, z) \mathbf{p}(x, \theta, z)\, d\eta(z),$$

$$Ac(s) = g^{-1}(s, \theta, \eta) \int c(z) \mathbf{g}(s, \theta, z)\, d\eta(z).$$

Next we prove that the linear space $\{Ac(\cdot)\colon c \in L_{2*}(\eta)\}$ is dense in $L_{2*}(g(\cdot,\theta,\eta))$. Indeed, suppose that $\mathbf{b} \in L_{2*}(g(\cdot,\theta,\eta))$ and $\mathbf{b} \perp Ac$ for all $c \in L_{2*}(\eta)$. Then

$$\int\int \mathbf{b}(s)1_{\{g(s,\theta,\eta)>0\}}\mathbf{g}(s,\theta,z)\,dv_\theta(s)\,c(z)\,d\eta(z) = 0.$$

Hence

$$\int \mathbf{b}(s)1_{\{g(s,\theta,\eta)>0\}}\mathbf{g}(s,\theta,z)\,dv_\theta(s) = 0, \quad \eta - \text{a.a. } z.$$

By L_2-completeness

$$\mathbf{b}(s)1_{\{g(s,\theta,\eta)>0\}} = 0, \quad \mathbf{g}(\cdot,\theta,z) - \text{a.e., } \eta\text{-a.a. } z.$$

Hence $\mathbf{b} = 0$, $g(\cdot,\theta,\eta)$-a.e.

Finally, let $\mathbf{b} \in L_{2*}(g(\cdot,\theta,\eta))$, arbitrary. Then by the preceding argument there exists $\{\mathbf{b}_n\} \subset L_{2*}(g(\cdot,\theta,\eta))$ (of the form Ac_n) with $\mathbf{b}_n \to \mathbf{b}$ and such for every $n = 1,2,\ldots$ there is $\{\eta_{nt}\} \subset H$ such that

$$r_{nt} = \int\Big[t^{-1}\big(p^{1/2}(x,\theta+th,\eta_{nt}) - p^{1/2}(x,\theta,\eta)\big) - \tfrac{1}{2}\big(h\dot{l}(x,\theta,\eta) + \mathbf{b}_n(\psi(x,\theta))\big)p^{1/2}(x,\theta,\eta)\Big]^2 d\mu(x) \to 0$$

as $t \downarrow 0$. Choose a sequence $\{t_n\}$ with $t_n \downarrow 0$ such that $r_{nt} < n^{-1}$ if $t \le t_n$. Next let η_t be η_{nt} if $t_{n+1} < t \le t_n$. Then $\{\eta_t\}$ satisfies (5.7). □

5.2. *Mixtures over an exponential family.* Before giving concrete examples of mixture models we specialize to mixtures over an exponential family. Suppose that $\mathbf{g}(\cdot,\theta,z)$ in (5.9) takes the form

$$\mathbf{g}(s,\theta,z) = c(z,\theta)\,d(s,\theta)e^{sze(\theta)}, \qquad s,z \in \mathbb{R}. \tag{5.14}$$

Let $\mathscr{Z}(\theta)$ be the set of z-values for which the family is defined, i.e.,

$$\mathscr{Z}(\theta) = \left\{z \in \mathbb{R}\colon \int d(s,\theta)\exp(sze(\theta))\,dv_\theta(s) < \infty\right\}.$$

Next let H be a set of probability distributions on $\mathscr{Z} = \bigcap\{\mathscr{Z}(\theta)\colon \theta \in \Theta\}$.

It is well known that an exponential family is complete if the parameter set contains an open interval. In the following lemma this result is adapted to our purposes. Recall that the support of a probability distribution on $\mathscr{Z} \subset \mathbb{R}$ is the smallest closed set with probability 1.

LEMMA 5.1. *Let* (5.14) *hold with* $e(\theta) \neq 0$ *and suppose that the support of* η *contains a limit point within the interior of* $\mathscr{Z}$. *Then* $\{\mathbf{g}(\cdot,\theta,z)\,dv_\theta\colon z \in A\}$ *is* L_2*-complete for every* $A \in \mathscr{A}$ *with* $\int_A d\eta = 1$.

PROOF. Let $\int_A d\eta(z) = 1$ and let $b\colon \mathbb{R} \to \mathbb{R}$ be measurable and satisfy

$$\begin{aligned} &\int b^2(s)\mathbf{g}(s,\theta,z)\,dv_\theta(s) < \infty, \quad \text{all } z \in A,\\ &\int b(s)\mathbf{g}(s,\theta,z)\,dv_\theta(s) = 0, \quad \text{all } z \in A. \end{aligned} \tag{5.15}$$

Set

$$\mathscr{Z}_b(\theta) = \left\{ z \in \mathscr{Z} \colon \int |b(s)| \mathbf{g}(s, \theta, z)\, d\nu_\theta(s) < \infty \right\}.$$

Since support$(\eta) \subset \bar{A}$, we are guaranteed an infinite sequence $\{\bar{z}_j\} \subset \bar{A}$ converging to a point $z_0 \in \operatorname{Int} \mathscr{Z}$ and hence a sequence $\{z_j\} \subset A$ with $z_j \to z_0$. We first show that $z_0 \in \operatorname{Int} \mathscr{Z}_b(\theta)$. For any z and z_j,

$$\begin{aligned}
&\int |b(s)| e^{sze(\theta)} c(z, \theta)\, d(s, \theta)\, d\nu_\theta(s) \\
&\quad \le \left[\int b^2(s) \mathbf{g}(s, \theta, z_j)\, d\nu_\theta(s) \right]^{1/2} \left[\int e^{4s(z - z_0)e(\theta)} \mathbf{g}(s, \theta, z_j)\, d\nu_\theta(s) \right]^{1/4} \\
&\qquad \times \left[\int e^{4s(z_0 - z_j)e(\theta)} \mathbf{g}(s, \theta, z_j)\, d\nu_\theta(s) \right]^{1/4} c(z, \theta)/c(z_j, \theta).
\end{aligned}$$

But, as $z_0 \in \operatorname{Int} \mathscr{Z}$, this is finite for sufficiently large j and small $|z - z_0|$.

It is well known that the function $\zeta \to \int b(s) \mathbf{g}(s, \theta, \zeta)\, d\nu_\theta(s)$, is analytic on $\mathscr{G} = \{\zeta \in \mathbb{C} \colon \operatorname{Re} \zeta \in \operatorname{Int} \mathscr{Z}_b(\theta)\}$. Since it is identically 0 at the sequence $\{z_j\}$ which has a limit point in $\operatorname{Int} \mathscr{Z}_b(\theta)$, we see by analytic continuation that

$$\int b(s) \mathbf{g}(s, \theta, \zeta)\, d\nu_\theta(s) = 0, \quad \text{all } \zeta \in G.$$

Hence there exists $u \in \operatorname{Int} \mathscr{Z}_b(\theta) e(\theta)$ such that for all $v \in \mathbb{R}$,

$$\int e^{isv} b^+(s) e^{su}\, d(s, \theta)\, d\nu_\theta(s) = \int e^{isv} b^-(s) e^{su}\, d(s, \theta)\, d\nu_\theta(s).$$

By uniqueness of Fourier transforms the finite measures given by $\tau^+(B) = \int b^+(s) e^{su}\, d(s, \theta)\, d\nu_\theta(s)$ and $\tau^-(B) = \int b^-(s) e^{su}\, d(s, \theta)\, d\nu_\theta(s)$, respectively, are equal. Hence

$$b^+ = b^- \quad \text{a.e. } [\mathbf{g}(\cdot, \theta, z)\, d\nu_\theta]. \qquad \square$$

Thus in the case of a mixture over an exponential family, we have been able to show the existence of a least favourable submodel as in (4.1) and (5.6) provided that the support of the mixing distribution contains a limit point in $\operatorname{Int} \mathscr{Z}$. In fact, we have established the much stronger property of local completeness. In general, local completeness of $\psi(X_1, \theta)$ will fail if the support of the mixing distribution is finitely discrete or countable without limit point. However, as is shown in the following two examples, a least favourable submodel as in (4.1) and (5.6) may still exist.

We also note that it can be proved, that irrespective of the support of the mixing distribution, the estimator constructed in Sections 2 and 3 is typically LAM and best regular (in a more general sense than in Section 4 [cf. van der Vaart (1986b)]. However, in such situations there may exist estimators which behave strictly better than the estimator constructed in Sections 2 and 3. These improved estimators will not be regular, but, of course, will be LAM. In the case that a least favourable submodel as in (4.1) and (5.6) exists, a LAM estimator is

necessarily asymptotically linear, according to the second statement of Proposition 4.1. Thus in this case the estimator constructed in Sections 2 and 3 is the essentially unique LAM estimator. In this sense it is interesting to know whether (4.1) and (5.6) is satisfied or not; i.e., in the case of mixture models, whether the support of the mixing distribution contains a limit point or not.

EXAMPLE 5.1 (Errors in variables). Let H be a class of probability distributions on $\mathbb{R}$ and let $\mathbf{p}(\cdot, \theta, z)$ be the density of the bivariate normal distribution with mean $(z, \theta z)'$ and covariance matrix equal to the identity matrix I. Let the observations be pairs (X_j, Y_j) having the distribution with density $p(\cdot, \theta, \eta)$ given by (5.8). This model may be structurally written

$$\begin{pmatrix} X_j \\ Y_j \end{pmatrix} = \begin{pmatrix} Z_j \\ \theta Z_j \end{pmatrix} + \begin{pmatrix} e_j \\ f_j \end{pmatrix},$$

where Z_j, e_j and f_j are independent, Z_j has distribution η and $(e_j, f_j)'$ is bivariate standard normal. Thus an independent variable Z_j is observed with error and a dependent variable Y_j is a regression on Z_j with slope θ. This model can be enlarged with more parameters, but we restrict ourselves in this paper to one-dimensional θ.

As a sufficient statistic we choose $\psi(X_j, Y_j, \theta) = X_j + \theta Y_j$ which conditionally on $Z_j = z$ has a $N(z(1 + \theta^2), 1 + \theta^2)$ distribution. Under the condition that $\eta \in H$ has $0 < \int z^2 \, d\eta(z) < \infty$, the conditions of both Theorems 2.1 and 3.1 are satisfied, and Sections 2 and 3 thus yield a construction of an estimator satisfying (1.8). Here we have $\dot{l}(x, y, \theta, z) = -z\phi'/\phi(y - \theta z)$, so that $\iiint \dot{l}^2(x, \theta, z)\mathbf{p}(x, y, \theta, z) \, dx \, dy \, d\eta(z) = \int z^2 \, d\eta(z)$. Theorem 5.1 may be used to obtain the score for θ and we find that the partition (2.7) holds with

$$\tilde{H}_{nj}(x, y, \theta) = (1 + \theta^2)^{-2}(x + \theta y)(y - \theta x)$$

and

$$\tilde{\psi}_{nj}(x, y, \theta) = (1 + \theta^2)^{-1}(y - \theta x).$$

Furthermore,

$$\beta_n^2(s, \theta) = (1 + \theta^2)^{-1}.$$

A $\sqrt{n}$-consistent estimator can be obtained from the estimating equation

$$\sum_{j=1}^{n} \tilde{H}_{nj}(X_j, Y_j, \theta) = 0.$$

Indeed, it can be checked that one of the solutions,

$$\hat{\theta}_n = \left(2 \sum_{j=1}^{n} X_j Y_j\right)^{-1} \Bigg[\sum_{j=1}^{n} \left(Y_j^2 - X_j^2\right) + \left\{ \left(\sum_{j=1}^{n} \left(Y_j^2 - X_j^2\right) \right)^2 + \left(2 \sum_{j=1}^{n} X_j Y_j\right)^2 \right\}^{1/2} \Bigg] 1\{\Sigma_{j=1}^{n} X_j Y_j \neq 0\},$$

is such that $\mathcal{L}_{\theta\eta}(\sqrt{n}\,(\hat{\theta}_n - \theta))$ converges to a normal distribution as $n \to \infty$.

Finally, consider efficiency of the estimator sequence $\{T_n\}$ satisfying (1.8). In the case that $T(\eta, H) = L_{2^*}(\eta)$ for all $\eta \in H$, we obtain immediately from Theorem 5.2 and Lemma 5.1 that $\{T_n\}$ is efficient at all (θ, η) for which the support of η contains a limit point. We now show independently that in this special case $\{T_n\}$ is in fact efficient at all $(\theta, \eta) \in \Theta \times H$ as long as for every $\eta \in H$ the scale family $\{\sigma^{-1}\eta(\sigma^{-1}\cdot)\colon \sigma > 0\}$ belongs to H too.

Let $\dot{\mathbf{l}}_z(x, y, \theta, z) = \partial/\partial z \log \mathbf{p}(x, y, \theta, z)$. Then

$$E_\theta\big(\dot{\mathbf{l}}(X_1, Y_1, \theta, z) | X_1 + \theta Y_1 = x + \theta y\big) = z\theta(1 + \theta^2)^{-1}\dot{\mathbf{l}}_z(x, y, \theta, z).$$

By Theorem 5.1

$$\dot{l}(x, y, \theta, \eta) = p^{-1}(x, y, \theta, \eta)\int \dot{\mathbf{l}}(x, y, \theta, z)\mathbf{p}(x, y, \theta, z)\, d\eta(z).$$

Thus we see that

$$\begin{aligned} &E_\theta\big(\dot{l}(X_1, Y_1, \theta, \eta) | X_1 + \theta Y_1 = x + \theta y\big) \\ &\quad = p^{-1}(x, y, \theta, \eta)\int z\theta(1 + \theta^2)^{-1}\dot{\mathbf{l}}_z(x, y, \theta, z)\mathbf{p}(x, y, \theta, z)\, d\eta(z). \end{aligned} \tag{5.16}$$

Informally this is

$$\begin{aligned} &-\partial/\partial t_{|t=0}\log \int \mathbf{p}\Big(x, y, \theta, z\big(1 - t\theta(1 - \theta^2)^{-1}\big)\Big)\, d\eta(z) \\ &\quad = -\partial/\partial t_{|t=0}\log \int \mathbf{p}(x, y, \theta, z)\, d\eta_t(z), \end{aligned}$$

where η_t is the measure defined by $\eta_t(B) = \eta((1 - t\theta(1 + \theta^2)^{-1})^{-1}B)$.

This argument shows that the left-hand side of (5.16) is an η-score, the condition needed for efficiency of $\{T_n\}$ [cf. (1.6)]. It is straightforward to make this argument precise by checking (5.6) for the sequence $\{\eta_t\}$ given previously. The same argument implies that estimating θ in the model where (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$ are i.i.d. and distributed according to a bivariate normal distribution with mean $(z, \theta z)'$ and covariance I, z unknown and fixed, is asymptotically not easier than estimating θ in our present model (of which the former is a submodel if H contains the point masses).

A more detailed study of this model can be found in Bickel and Ritov (1987).

EXAMPLE 5.2 [Paired exponentials, cf. Lindsay (1985)]. Let H be a class of probability distributions on $(0, \infty) \subset \mathbb{R}$ and let $p(\cdot, \theta, \eta)$ satisfy (5.8) with the following density with respect to Lebesgue measure on $\mathbb{R}^2$,

$$\mathbf{p}(x, y, \theta, z) = ze^{-zx}\theta ze^{-\theta zy}1_{\{x > 0,\, y > 0\}}.$$

Since the parameter in an exponential distribution equals the (constant) hazard rate, this means that we estimate θ, the ratio of the hazard rates within pairs, common to a sample of paired observations.

Write the pairs of observations as (X_j, Y_j). As a sufficient statistic we can choose $\psi(X_1, Y_1, \theta) = X_1 + \theta Y_1$. It has density $g(s, \eta) = \int z^2 s e^{-zs}1_{\{s > 0\}}\, d\eta(z)$

with respect to Lebesgue measure. Theorems 2.1 and 3.1 apply without further conditions. We have

$$\dot{l}(x, y, \theta, z) = \theta^{-1} - zy, \qquad \int\int \dot{l}^2(x, y, \theta, z)\mathbf{p}(x, y, \theta, z)\, dx\, dy = \theta^{-2}$$

and

$$\tilde{l}(x, y, \theta, \eta) = (2\theta(x + \theta y))^{-1}(x - \theta y) + Q(x + \theta y, \eta)(2\theta)^{-1}(\theta y - x),$$

where $Q(s, \eta) = g'(s, \eta)/g(s, \eta)$. Furthermore, $\mathscr{L}(X - \theta Y | X + \theta Y = s)$ is the uniform distribution on $[-s, s] \subset \mathbb{R}$, so that

$$\beta^2(s, \theta) = E_\theta\big(\tilde{\psi}^2(X_1, Y_1, \theta) | X_1 + \theta Y_1 = s\big) = (12\theta^2)^{-1}s^2.$$

A $\sqrt{n}$-consistent estimator $\hat{\theta}_n$ can be defined as the unique solution to

$$\sum_{j=1}^{n} \tilde{H}\big(X_j, Y_j, \theta\big) = 0,$$

where $\tilde{H}(x, y, \theta) = (2\theta)^{-1}(x + \theta y)^{-1}(x - \theta y)$.

Suppose that H is the class of all probability measures on $\mathbb{R}$, or more generally that $T(\eta, H) = L_{2^*}(\eta)$ for all $\eta \in H$. From Theorem 5.2 and Lemma 5.1 we immediately conclude that the resulting estimator sequence $\{T_n\}$ is efficient at all (θ, η) for which the support of η contains a limit point in $(0, \infty)$. An argument analogous to that in Example 5.1 shows that $\{T_n\}$ is in fact efficient at all $(\theta, \eta) \in (0, \infty)$ provided that for every $\eta \in H$ we have that $\{\sigma^{-1}\eta(\sigma^{-1}\cdot): \sigma > 0\} \subset H$.

van der Vaart (1988) contains a more extensive discussion of this model, where also the non-i.i.d. case in which η may differ from observation to observation is considered. It turns out that the estimator sequence remains asymptotically normal as long as the averages $n^{-1}\sum_{j=1}^{n}\eta_{nj}$ do not let mass escape to either 0 or ∞.

6. Conditional likelihood. In this section we give an informal discussion of the situation that $\psi(\cdot, \theta)$ is independent of θ. We consider the i.i.d. version of our model given by (5.1)–(5.2) and write $\psi(x)$ for $\psi(x, \theta)$.

We suppose that (5.3) holds and identify the derivative $\dot{l}(x, \theta, \eta)$ with the pointwise derivative $\partial/\partial\theta \log p(x, \theta, \eta)$. Clearly,

$$\dot{l}(x, \theta, \eta) = \dot{h}/h(x, \theta) + \dot{g}/g(\psi(x), \theta, \eta) \tag{6.1}$$

and

$$\tilde{l}(x, \theta, \eta) = \dot{h}/h(x, \theta) - E_\theta\big(\dot{h}/h(X_1, \theta) | \psi(X_1) = \psi(x)\big).$$

Here one usually finds that

$$E_\theta\big(\dot{h}/h(X_1, \theta) | \psi(X_1)\big) = 0, \tag{6.2}$$

a fact which can be explained by reference to the conditional distribution of X_1 given $\psi(X_1)$. Indeed, one expects the function $h(x, \theta)$ to be the density of X_1 in x, given that $\psi(X_1) = \psi(x)$. Then by the usual change in order of differentiation

and integration

$$E_\theta\big(\dot{h}/h(X_1,\theta)|\psi(X_1)=t\big)=\int \partial/\partial\theta\, p_{X_1|\psi(X_1)=t}(x)\,d\tau_t(x)=0.$$

We do not try to formalize the preceding argument in general, but from now on assume that (6.2) holds true.

In the present situation we have that $\beta_n(s,\theta)=0$, so that (2.10) is trivially satisfied: Estimation of $g'/g(\cdot,\theta,\eta)$ is unnecessary. The construction of Section 2 is valid and the resulting estimator may be considered a one-step version of a *conditional maximum likelihood estimator*, discussed (in a different setup) by Andersen (1970). Indeed, under (6.1)–(6.2) the estimator of Section 2 takes the form

$$T_n=\hat{\theta}_n+n^{-1}\sum_{j=1}^{n}\hat{I}_n^{-1}\big(\hat{\theta}_n\big)\dot{h}/h\big(X_j,\hat{\theta}_n\big),$$

while a conditional maximum likelihood estimator is defined as the value of θ that maximizes $\pi_{j=1}^n h(X_j,\theta)$.

Condition (4.1) is of course still sufficient to render this estimator efficient, and in particular we may apply the results of Section 5 to obtain efficiency of the conditional maximum likelihood estimator in mixtures over an exponential family, when the support of the mixing distribution contains a limit point. For the present case, where the sufficient statistic is independent of θ, this result was essentially obtained for mixture models by Pfanzagl (1982), Section 14.

Alternatively, one would expect that an estimator for θ based on the conditional distribution of X_1 given $\psi(X_1)$ is efficient if and only if the marginal distribution of $\psi(X_1)$ does not contain information about θ. Here, for information one should of course read efficient information, which in analogy with (1.3)–(1.4), is defined as

$$I_g(\theta,\eta)=\inf E_{\theta\eta}\big(\dot{g}/g(\psi(X_1),\theta,\eta)-\mathbf{b}(\psi(X_1))\big)^2.$$

Here the infimum is taken over the same set $\mathbf{B}(\theta,\eta)$ of functions $\mathbf{b}$ as in (1.4), i.e., the set of $\mathbf{b}$ such that $\mathbf{b}(\psi(\cdot))$ is an η-score. As explained in Section 1 one expects the estimator of Section 2 to be efficient at (θ,η) if (1.6) is satisfied. Under the assumptions made so far the latter is true if and only if $I_g(\theta,\eta)=0$.

LEMMA 6.1. *Let* (5.1)–(5.3) *and* (6.1)–(6.2) *hold with* $\psi(\cdot,\theta)=\psi(\cdot)$, $\nu_\theta=\nu$. *Then* (1.6) *holds if and only if* $I_g(\theta,\eta)=0$.

PROOF. We have

$$\tilde{l}_n(\cdot,\theta,\eta)=E_\theta\big[\dot{l}(X_j,\theta,\eta)|\psi(X_j)=\cdot\big]=\dot{g}/g(\cdot,\theta,\eta).\qquad\square$$

Following Andersen (1970), call $\psi(X_1)$ *weakly ancillary* if $\{g(\cdot,\theta,\eta)\colon \eta\in H\}$ is the same for each $\theta\in\Theta$. Weak ancillarity implies that for any t there exists $\eta_t\in H$ such that $g(\cdot,\theta+t,\eta)=g(\cdot,\theta,\eta_t)$. Thus it implies that $I_g(\theta,\eta)=0$, and we therefore expect it to be sufficient to render the conditional maximum likelihood estimator efficient.

This statement can be made precise under regularity conditions, or alternatively in concrete examples by checking condition (4.1). In the latter case the preceding discussion may reveal the least favourable one-dimensional submodel.

EXAMPLE 6.1 (Neyman and Scott). Let H be a set of probability distributions on $\mathbb{R}$, $\theta \in \Theta = (0, \infty)$ and let $p(\cdot, \theta, \eta)$ satisfy (5.8) with $\mathbf{p}(\cdot, \theta, z)$ equal to the density of a $N_2((z, z)', \theta I)$ distribution. The model satisfies (5.1)–(5.2) with sufficient statistic $\psi(X_1, Y_1, \theta) = X_1 + Y_1$. The conditional distribution of (X_1, Y_1) given $X_1 + Y_1 = s$ is restricted to straight lines in $\mathbb{R}^2$ and can be related to a $N(\frac{1}{2}s, \frac{1}{2}\theta)$ distribution on $\mathbb{R}$. Indeed, we can factorize $p(\cdot, \theta, \eta)$ as in (5.1) with

$$g(s, \theta, \eta) = \int (2\theta)^{-1/2} \phi\big((2\theta)^{-1/2}(s - 2z)\big)\, d\eta(z),$$

$$h(x, y, \theta) = \big(\tfrac{1}{2}\theta\big)^{-1} \phi\Big(\big(\tfrac{1}{2}\theta\big)^{-1/2}\big(x - \tfrac{1}{2}(x + y)\big)\Big),$$

where ϕ is the standard normal density. The conditional likelihood estimator is the solution to $\sum_{j=1}^n \dot{h}/h(X_j, Y_j, \theta) = 0$ and is given by $T_n = (2n)^{-1} \sum_{j=1}^n (X_j - Y_j)^2$.

This example became famous because the *direct* maximum likelihood estimator $\hat{\theta}_{mn}$ for the functional version of the model, obtained by maximizing $\prod_{j=1}^n \mathbf{p}(X_j, Y_j, \theta, z_j)$ over all $(\theta, z_1, z_2, \ldots, z_n) \in \mathbb{R}^{n+1}$, is inconsistent. (Indeed, $\hat{\theta}_{mn} = \frac{1}{2}\hat{\theta}_n \to \frac{1}{2}\theta$.) In contrast to this, T_n is not only consistent, but also efficient in the i.i.d. mixture model, for instance if H is the set of all distributions on $\mathbb{R}$ for which the support contains a limit point.

EXAMPLE 6.2 (Paired Poisson variables). Let μ be counting measure on the points $\{(x, y) \in \mathbb{R}^2\colon x, y = 0, 1, 2, \ldots\}$, let H be a set of probability distributions on $\mathbb{R}$ and let the observations (X_j, Y_j) be i.i.d. distributed according to a density given by (5.8), where

$$\mathbf{p}(x, y, \theta, z) = (\theta e^z)^x e^{-\theta e^z} (x!)^{-1} (e^z)^y e^{-e^z} (y!)^{-1}.$$

We can choose $\psi(X_1, Y_1, \theta) = X_1 + Y_1$ which has a density $g(\cdot, \theta, \eta)$ with respect to counting measure on $\{0, 1, 2, \ldots\}$ given by

$$g(s, \theta, \eta) = \int [(\theta + 1)e^z]^s e^{-(\theta+1)e^z} (s!)^{-1}\, d\eta(z).$$

As is easily checked $\psi(X_1, Y_1, \theta)$ is weakly ancillary if for any $\eta \in H$ the location family $\{\eta(\cdot - \mu)\colon \mu \in \mathbb{R}\}$ belongs to H. The conditional maximum likelihood estimator is given by $\sum_{j=1}^n X_j / \sum_{j=1}^n Y_j$. If H is the set of all probability distributions on $\mathbb{R}$ it can be checked that it is efficient at all $(\theta, \eta) \in \Theta \times H$. Note that this example differs from the foregoing one in that the least favourable submodel as in (4.1) always exists, irrespective of the support of η.

7. Other examples. Mixture models are not the only source of examples for the structure given by (1.1)–(1.2). In this section we discuss the problem of estimating a centre of symmetry, regression models and two-sample problems, where the two samples differ by a transformation depending on θ.

EXAMPLE 7.1 (Symmetric location). Let H be a set of probability densities η with respect to Lebesgue measure λ on $\mathbb{R}$, absolutely continuous, symmetric about 0 and with finite and positive Fisher information for location $I_1(\eta) = \int(\eta'/\eta)^2\eta\, d\lambda$. Let $\sigma_1, \sigma_2, \ldots$ be known positive numbers such that $\max\{\sigma_j^{-1}: j = 1, 2, \ldots, n\} = o(\sqrt{n})$ and $0 < \liminf n^{-1}\Sigma_{j=1}^n \sigma_j^{-2} \leq \limsup n^{-1}\Sigma_{j=1}^n \sigma_j^{-2} < \infty$. Set $p_j(x, \theta, \eta) = \sigma_j^{-1}\eta(\sigma_j^{-1}(x - \theta))$. Then (1.1)–(1.3) are satisfied with the sufficient statistics $\psi_j(X_j, \theta) = \sigma_j^{-1}|X_j - \theta|$, which have density $g(\cdot, \theta, \eta) = 2\eta(\cdot)1_{(0,\infty)}(\cdot)$ with respect to λ. We have $\dot{l}_j(x, \theta, \eta) = -\sigma_j^{-1}\eta'/\eta(\sigma_j^{-1}(x - \theta))$ and $\bar{l}_n(s, \theta, \eta) = 0$. The decomposition (2.7) holds with $\tilde{\psi}_{nj}(x, \theta) = -\sigma_j^{-1}\operatorname{sgn}(x - \theta)$, so that $\beta_n^2(s, \theta) = n^{-1}\Sigma_{j=1}^n \sigma_j^{-2}$. A $\sqrt{n}$-consistent estimator can be found by the M-method with η' equal to the logistic density. It can be checked that Theorems 2.1 and 3.1 apply and, as is well known, the estimator is efficient even if H equals $\{\eta\}$: Of course, the i.i.d. version of this model has been treated by many authors and our only contribution is to analyse the model in terms of sufficient statistics. Note here that the usual symmetrization of the kernel estimator for the density η, has now been taken care of automatically, as we estimate the density of the sufficient statistic.

EXAMPLE 7.2 (Regression). Let $\Theta = \mathbb{R}$ and μ Lebesgue measure on $\mathbb{R}$. Let $\xi_1, \xi_2, \ldots, \xi_n$ be known constants, satisfying $\max\{|\xi_j|: j = 1, 2, \ldots, n\} = o(\sqrt{n})$ and $\liminf n^{-1}\Sigma_{j=1}^n(\xi_j - \bar{\xi}_n)^2 < \limsup n^{-1}\Sigma_{j=1}^n(\xi_j - \bar{\xi}_n)^2 < \infty$, where $\bar{\xi}_n = n^{-1}\Sigma_{j=1}^n \xi_j$. Given unobservable i.i.d. error terms $e_1, e_2, \ldots, e_n$, distributed according to an absolutely continuous density η with $I_1(\eta) = \int(\eta'/\eta)^2\eta\, d\lambda < \infty$, we set $X_j = \theta\xi_j + e_j$. In other words $p_j(x, \theta, \eta) = \eta(x - \theta\xi_j)$, $j = 1, 2, \ldots, n$. Consider two cases:

7.2.1. η is symmetric about 0. A sufficient statistic is $|X_j - \theta\xi_j|$. The further analysis resembles Example 7.1.

7.2.2. H is an arbitrary set of densities with finite Fisher information. A sufficient statistic is $X_j - \theta\xi_j$. We have $\bar{l}_n(s, \theta, \eta) = -\bar{\xi}_n\eta'/\eta(s)$, $\tilde{l}_{nj}(x, \theta, \eta) = -(\xi_j - \bar{\xi}_n)\eta'/\eta(x - \theta\xi_j)$ and $\beta_n^2(s, \theta) = n^{-1}\Sigma_{j=1}^n(\xi_j - \bar{\xi}_n)^2$. The conditions of Theorems 2.1 and 3.1 are satisfied and thus Sections 2 and 3 yield an asymptotically linear estimator. Because $\bar{l}_n(\cdot, \theta, \eta)$ is proportional to $\eta'/\eta(\cdot)$, this estimator is efficient if for each $\eta \in H$ the location family $\{\eta(\cdot - \mu): \mu \in \mathbb{R}\}$ is also contained in H. More formally, one can choose $\eta_n(h)(\cdot) = \eta(\cdot + \bar{\xi}_n n^{-1/2}h)$ to obtain the least favourable one-dimensional submodel as in (4.1).

EXAMPLE 7.3 (Two-sample problems). Let H consist of probability densities with respect to Lebesgue measure on $\mathbb{R}$ and let μ be Lebesgue measure. Let a first sample $X_1, \ldots, X_m$ be i.i.d. with density $p_j(\cdot, \theta, \eta) = \eta(\cdot)$, $j = 1, 2, \ldots, m$, and let a second sample $X_{m+1}, \ldots, X_n$ be i.i.d. with density $p_j(\cdot, \theta, \eta) = \eta(A_\theta \cdot)\Delta(\cdot, \theta)$, $j = m + 1, m + 2, \ldots, n$, where A_θ: $\mathbb{R} \to \mathbb{R}$ is a sufficiently regular transformation and $m = m_n$. Clearly, as sufficient statistics can be chosen $\psi_j(X_j, \theta) = X_j$, $j = 1, 2, \ldots, m$, and $\psi_j(X_j, \theta) = A_\theta X_j$, $j = m + 1, \ldots, n$.

In particular, consider the transformation $A_\theta x = \theta^{-1}x$, $\theta > 0$, which leads to the two-sample scale model. Let η be absolutely continuous with finite and

positive Fisher information for scale, $I_s(\eta) = \int(1 + x\eta'/\eta(x))^2\eta(x)\,d\lambda(x)$ and assume that $0 < \liminf n^{-1}m \leq \limsup n^{-1}m < 1$. We have $\dot{l}_n(s, \theta, \eta) = -\theta^{-1}(n - m)n^{-1}(1 + s\eta'/\eta(s))$, $\beta_n(s, \theta) = (m(n - m))^{1/2}n^{-1}\theta^{-1}|s|$ and $g(s, \theta, \eta) = \eta(s)$. It is easily checked that the conditions of Theorems 2.1 and 3.1 are satisfied. The resulting estimator is efficient if for each $\eta \in H$ we have that $\{\sigma^{-1}\eta(\sigma^{-1}\cdot)\colon \sigma > 0\} \subset H$.

Acknowledgments. I thank W. R. van Zwet and C. A. J. Klaassen for introducing me to the model of type (1.1)–(1.2) [cf. Klaassen and van Zwet (1985)]. I also thank Chris Klaassen for helpful remarks during the preparation of this paper.

REFERENCES

ANDERSEN, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *J. Roy. Statist. Soc. Ser. B* **32** 283–301.

BEGUN, J. M., HALL, W. J., HUANG, W.-M. and WELLNER, J. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist.* **11** 432–452.

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.

HÁJEK, J. (1970). A characterization of limiting distributions of regular estimators. *Z. Wahrsch. verw. Gebiete* **14** 323–330.

HÁJEK, J. (1972). Local asymptotic minimax and admissibility in estimation. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **1** 175–194. Univ. California Press.

KLAASSEN, C. A. J. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.* **15** 1548–1562.

KLAASSEN, C. A. J. and VAN ZWET, W. R. (1985). On estimating a parameter and its score function. In *Proc. Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. M. Le Cam and R. A. Olshen, eds.) **2** 827–839. Wadsworth, Monterey, Calif.

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

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

PFANZAGL, J. (1982) (with W. Wefelmeyer). *Contributions to a General Asymptotic Statistical Theory. Lecture Notes in Statist.* **13**. Springer, New York.

SCHICK, A. S. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14** 1139–1151.

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

VAN DER VAART, A. W. (1986a). Estimating a real parameter in a class of semi-parametric models. Report 86-9, Dept. Mathematics, Univ. Leiden.

VAN DER VAART, A. W. (1986b). On the asymptotic information bound. Report 86-13, Dept. Mathematics, Univ. Leiden.

VAN DER VAART, A. W. (1988). Statistical estimation in large parameter spaces. Thesis. CWI tract 44, Centrum voor Wiskunde en Informatica, Amsterdam.

DEPARTMENT OF MATHEMATICS
FREE UNIVERSITY
DE BOELELAAN 1081
1081 HV AMSTERDAM
THE NETHERLANDS