The Annals of Statistics
1995, Vol. 23, No. 3, 968–989

ON OPTIMAL B-ROBUST INFLUENCE FUNCTIONS IN SEMIPARAMETRIC MODELS

BY LARRY Z. SHEN

The Procter & Gamble Company

Bounded influence functions are used for robust estimation in semiparametric models. In this paper, we generalize Hampel's variational problem to semiparametric models and define the optimal B-robust influence function as the one solving the variational problem. We identify the lowest bounds for influence functions and establish the existence and uniqueness of the optimal influence functions in general semiparametric models. Explicit optimal influence functions are given for a special case. Examples are provided to illustrate the procedures for calculating the optimal influence functions and for constructing the corresponding optimal estimators.

1. Introduction. Efficient and adaptive estimation in semiparametric models has been considered by many authors for the past 10 years. Most of the effort has been devoted to calculating efficient influence functions and to constructing efficient and adaptive estimates from them. The monograph by Bickel, Klaassen, Ritov and Wellner (1993) (BKRW hereafter) contains an extensive theoretical treatment of semiparametric models.

Robust estimation in semiparametric models has been explored recently by some authors, for example, Beran (1978), Wu (1990) and Chen (1990). The main reason for such interest comes from the concern that a few outliers in a data set may totally distort the efficient or adaptive estimate. In his Ph.D. dissertation, Chen (1990) extended the notions of influence functions to adapt them to the structures of semiparametric models. He also introduced the idea of "adaptive robustness." Following the approach of Beran (1978), Wu (1990) considered shrinking the Hellinger neighborhoods of certain semiparametric models and explored the robust aspect of efficient estimates for these models. Sasieni (1993) derived a large class of robust estimates in Cox's models. Cheng and Van Ness (1992) considered robust estimates in special errors-in-variables models. However, there has been little consideration of the optimality problems related to semiparametric models. In this paper we shall extend the approach of Hampel (1968) and Hampel, Ronchetti, Rousseeuw and Stahel (1986) to semiparametric models.

Let $\mathbf{X} = (\Omega, \mathscr{B}, \mu)$ be a measure space, where μ is a σ-finite measure. A typical semiparametric model has the form

$$\mathscr{P} = \left\{P_{\theta, g}: \theta \in \Theta \subset R^d, g \in \mathbf{G}\right\}, \tag{1}$$

Received December 1992; revised September 1994.
AMS 1991 *subject classifications*. 62F35, 62G35.
Key words and phrases. B-robust, optimal bounded influence function, Hampel's problem, most robust estimate, semiparametric models.

where $P_{\theta, g}$ is a distribution function on $\mathbf{X}$. The parameter θ is of main interest, and $\mathbf{G}$ is a collection of functions. The nonparametric component $g(\cdot)$ is usually of secondary interest. Sometimes, the model contains a finite dimensional nuisance parameter η which requires different treatment than $g(\cdot)$. The following examples are semiparametric models to be considered in this article.

EXAMPLE 1. *Symmetric location models.* Assume that the data $x_1, \ldots, x_n \in R^1$ are i.i.d. and come from the model

$$X = \theta + \varepsilon,$$

where the center θ is the parameter to be estimated; the error ε has density $g(\cdot)$ that is symmetric about the origin and otherwise unknown. Adaptive estimates in this model have been constructed by Stone (1975), Beran (1978) and Bickel (1982).

EXAMPLE 2. *Heteroscedastic regression models.* Here we observe random variables $(x_1, y_1), \ldots, (x_n, y_n)$, where y_is are real values, x_is are d-vectors and they are related by

$$y_i = \theta^T x_i + \exp\big(r(x_i, \eta)\big)\varepsilon, \qquad \theta \in R^d, \eta \in R^p. \tag{2}$$

The error ε is independent of x_i and has density $g(\cdot) \in \mathbf{G}$ (the nonparametric component) that is symmetric about the origin. The main parameter here is $\theta \in R^d$, and the nuisance parameter is $\eta \in R^p$. Some related references include Carroll and Ruppert (1982) and Bickel (1978), although they have focused on the situation when the scales are functions of $\theta^T x$. However, model (2) is included among those studied by Jobson and Fuller (1980).

EXAMPLE 3. *Semiparametric mixture models.* Consider a semiparametric model in which $P_{\theta, g} = \int Q_{\theta, \eta}\, dG(\eta)$. Each $Q_{\theta, \eta}$ represents a distribution with density function $f(\cdot, \theta, \eta)$. The set $\mathbf{G}$ contains all possible distributions of η.

This model is motivated by the following considerations: while an estimate of θ is required in the parametric model $\{Q_{\theta, \eta}\}$, it is necessary to introduce an *incidental* parameter η_j, which indexes the sampling distribution of X_j. The number of parameters becomes large as the sample size increases. This poses certain difficulties for the consistent estimation of θ. To reduce the number of parameters to be estimated, we treat the η_js as random variables coming from an unknown distribution G. Neyman and Scott (1948) were the first to adopt such an approach. For a more complete discussion of the theory associated with semiparametric mixture models, see Lindsay (1980) and BKRW.

Influence functions are important tools in robust estimation. Hampel (1968) used influence functions to measure robustness of estimators. Influence functions were originally defined by von Mises (1947) for functionals. More specifically, assume that T is a R^d-valued functional on the set of all

distribution functions and satisfies $T(P_{\theta,g}) \equiv \theta$ (Fisher consistency). The influence function corresponding to $T(\cdot)$ is defined through the von Mises derivative:

$$\lim_{t\to 0} \frac{T\big((1-t)P_{\theta,g} + tH\big) - \theta}{t} = \int \underline{\psi}(x, P_{\theta,g})\, dH.$$

In particular, $H = \delta_x$, the point mass distribution at x,

$$\underline{\psi}(x, P_{\theta,g}) = \lim_{t\to 0} \frac{T\big((1-t)P_{\theta,g} + t\delta_x\big) - \theta}{t}.$$

The function $\underline{\psi}\colon X \times \mathscr{P} \to R^d$ is called the influence function of $T(\cdot)$. It is assumed that for any distribution $P \in \mathscr{P}$, $\|\underline{\psi}(\cdot, P)\| \in L_2(P)$ and $\int \underline{\psi}(X, P)\, dP = 0$.

Let $x_1, \ldots, x_n$ be i.i.d. samples from $P_{\theta,g}$ and let $\hat{F}_n$ be their empirical distribution. One can estimate θ by $\hat{\theta}_n = T(\hat{F}_n)$. Heuristically,

$$\begin{aligned}\hat{\theta}_n &= T\big(P_{\theta,g} + \hat{F}_n - P_{\theta,g}\big)\\ &= T(P_{\theta,g}) + \int \underline{\psi}(X, P_{\theta,g})\, d\big(\hat{F}_n - P_{\theta,g}\big) + \text{Remainder}\big(\hat{F}_n - P_{\theta,g}\big) \qquad (3)\\ &= \theta + \frac{1}{n}\sum_{i=1}^{n} \underline{\psi}(x_i, P_{\theta,g}) + \text{Remainder}\big(\hat{F}_n - P_{\theta,g}\big).\end{aligned}$$

Fernholz (1983) provided conditions under which $\text{Remainder}(\hat{F}_n - P_{\theta,g}) = o_{P_{\theta,g}}(n^{-1/2})$.

We do not restrict ourselves to estimates which can only be expressed as functionals. In general, an estimate $\hat{\theta}_n$ is called (*locally*) *asymptotically linear* if it satisfies (3) for a $\underline{\psi}$ and if $\text{Remainder}(\hat{F}_n - P_{\theta,g}) = o_{P_{\theta,g}}(n^{-1/2})$. Correspondingly, the function $\underline{\psi}$ is again called the influence function of $\hat{\theta}_n$.

One important feature of an influence function is that it enables us to calculate the asymptotic distribution of the corresponding estimate. Indeed, by the central limit theorem,

$$\sqrt{n}\,\big(\hat{\theta}_n - \theta\big) \to_{P_{\theta,g}} N\big(0, V(\underline{\psi}, \theta, g)\big),$$

where $V(\underline{\psi}, \theta, g) = \int \underline{\psi}(X, \theta, g)\underline{\psi}(X, \theta, g)^T\, dP_{\theta,g}$ is the asymptotic covariance matrix.

The influence function of an estimate also indicates the sensitivity of the estimate to the observations. If the influence function is unbounded, then a single outlier could totally distort the estimate. For parametric models, Hampel (1968) called an estimate *bounded robust*, or B-robust, if its influence function is bounded. The idea is to restrict the sensitivity of the estimate to outliers. The use of estimates with bounded influence functions has become a standard approach in robust estimation. However, a bounded influence function usually forces a compromise between efficiency and robustness. Attention is then restricted to a subclass of estimates with influence functions that are bounded by a constant, and we will find the best one within this class. Based

on this idea, Hampel (1968) and Hampel, Ronchetti, Rousseeuw and Stahel (1986) introduced the following criterion of optimality. For $C > 0$, find an influence function $\underline{\psi}_0$ to solve the problem

$$\text{(V)} \qquad \begin{aligned} &\text{minimize } \int \underline{\psi}^T \underline{\psi}\, dP_{\theta,g} \text{ over all influence functions } \underline{\psi} \\ &\text{subject to } \sup_x \|\underline{\psi}(x)\| \le C. \end{aligned}$$

DEFINITION 1. An influence function $\underline{\psi}$ solving problem (V) is called the *optimal* (*B-robust*) influence function corresponding to C (or simply *B-optimal*). An estimate is said to be *B-optimal* if its influence function is B-optimal.

In this paper, we first identify the lowest bound that an influence can have. Then we establish the existence and uniqueness of the optimal influence functions in general semiparametric models. The optimal influence functions are given explicitly for a special case which includes many interesting models. In the examples we consider, optimal estimators can be constructed from the optimal influence functions. Although it is difficult to provide explicit optimal influence functions for general semiparametric models, we are able to approximate the optimal influence function by a series of functions with explicit expressions.

2. Preliminaries. Let us introduce notation and review some basic concepts related to semiparametric models. We essentially follow the lines of Chapters 2 and 3 of BKRW. For the semiparametric model defined by (1), assume that $P_{\theta,g}$ has density $p(x, \theta, g)(x)$ with respect to Lebesgue measure (μ). Throughout this paper the models are assumed to be *regular*, the definition of which may be found in BKRW. Define the score function for θ by

$$\text{(4)} \qquad \underline{\dot{\mathbf{l}}}_\theta(x, \theta, g) = \left(\dot{\mathbf{l}}_1(x, \theta, g), \ldots, \dot{\mathbf{l}}_d(x, \theta, g)\right)^T,$$

with

$$\dot{\mathbf{l}}_1 = \frac{\partial}{\partial \theta_1} \log p(x, \theta, g), \ldots, \dot{\mathbf{l}}_d = \frac{\partial}{\partial \theta_d} \log p(x, \theta, g).$$

Let $L_2(P_{\theta,g})$ be the Hilbert space for which the inner product is generated by $P_{\theta,g}$. We define the *tangent space* for the nonparametric component. First, however, introduce the set

$$\mathbf{M}_0 = \Big\{ t(x) : t(x) \text{ is a score function for any submodel } \left\{P_{\theta, g_\eta} : \eta \in (0,1),\, g_\eta \in \mathbf{G}\right\} \Big\}.$$

The tangent space $\dot{P}_2$ is defined as the L_2-closure of the linear span of M_0. Usually $\dot{P}_2$ is an infinite dimensional space for a semiparametric model. Let

$\Pi(\cdot|\dot{P}_2)$ be the projection operator from $L_2(P_{\theta,g})$ onto its subspace $\dot{P}_2$. The *efficient score function* for θ is

$$\underline{\dot{l}}^*_\theta = \underline{\dot{l}}_\theta - \Pi\left(\underline{\dot{l}}_\theta|\dot{P}_2\right). \tag{5}$$

Correspondingly, $\tilde{\underline{l}}_\theta := A^{-1}\underline{\dot{l}}^*_\theta$ is the *efficient influence function*, where $A = \int \underline{\dot{l}}^*_\theta \underline{\dot{l}}^T_\theta \, dP_{\theta,g}$. An estimate of θ is *efficient* if it has an efficient influence function. Begun, Hall, Huang and Wellner (1983) established the optimality of efficient estimates using both convolution and minimax arguments. It can be shown that the efficient influence function minimizes the trace of $V(\underline{\psi}, \theta, g)$ and thus it solves problem (V) in the case $C = \infty$.

In order to solve problem (V) for $C < \infty$, one needs to propose candidates for influence functions. Assume that an estimate $\hat{\theta}_n$ of θ is Gaussian regular and is asymptotically linear with influence function $\underline{\psi}$. It is known from Corollary 3.3.4 of BKRW that $\underline{\psi}(\cdot, \theta, g) - \tilde{\underline{l}}_\theta$ is orthogonal to both $\{1\}$ and $\dot{P}_2$. Thus, the influence function $\underline{\psi}(\cdot, \theta, g)$ must satisfy the following conditions:

Consistency.

$$\text{(i)} \qquad \int \underline{\psi}(X, \theta, g)\, dP_{\theta,g} = 0,$$

$$\text{(ii)} \qquad \int \underline{\psi}(X, \theta, g)\underline{\dot{l}}^T_\theta \, dP_{\theta,g} = I_{d\times d},$$

$$\text{(iii)} \qquad \int \underline{\psi}(X, \theta, g)t(X)\, dP_{\theta,g} = \underline{0}, \qquad \forall\, t(x) \in \dot{P}_2.$$

Sometimes there may be a nuisance parameter η in addition to the nonparametric component. Denote the score function for η by $\underline{\dot{l}}_\eta$. The influence function $\underline{\psi}$ will satisfy one additional condition:

$$\text{(iv)} \qquad \int \underline{\psi}(X, \theta, g)\underline{\dot{l}}^T_\eta \, dP_{\theta,g} = 0.$$

Convention. Even though an influence function is associated with an estimate, a function satisfying the consistency conditions (i)–(iii) [and (iv) when there is a nuisance parameter] is still treated as an influence function.

Hampel's problem (V) may be reformulated as follows: minimize $\int \|\underline{\psi}\|^2 \, dP_{q,g}$ among all functions $\underline{\psi}$ satisfying the consistency conditions (i)–(iii) [and (iv) when there is a nuisance parameter] and $\sup_x \|\underline{\psi}(x)\| \le C$.

Since $\dot{P}_2$ may be of infinite dimension, it is difficult to solve problem (V) directly. As is the case for parametric model, it is expected that there exists a value C_0 such that problem (V) has a solution only when C is greater than C_0. In the next section we identify this lowest bound and establish the existence and uniqueness of the optimal influence functions in general semiparametric models. In Section 3, we derive explicit expressions of the optimal influence functions. Finally, in Section 4, we calculate the optimal influence

functions and construct the corresponding optimal estimates for the examples.

3. The lowest bound. Throughout this section, we let (θ, g) be fixed so that we can suppress the dependence of P and $\underline{\psi}$ on (θ, g).

PROPOSITION 1. *Let* $S_c = \{\underline{\psi}: \underline{\psi}$ *satisfies conditions* (i)–(iii) *and* $\operatorname{ess\,sup}_x \|\underline{\psi}\| \le c\}$. *If the set* S_c *is nonempty, then there is a solution to problem* (V). *If both* $\underline{\psi}_1$ *and* $\underline{\psi}_2$ *solve* (V), *then they differ from each other at most on a* P*-null set.*

PROOF. It is clear that the set S_c is convex and closed. While the convexity is straightforward, closedness follows from the fact that for any $\underline{\psi}_n \in S_c$ such that $\|\underline{\psi}_n - \underline{\psi}\|_{L_2} \to 0$, there is a subsequence of $\underline{\psi}_{n_k}$ converging to $\underline{\psi}$ almost surely, implying $\operatorname{ess\,sup}_x |\underline{\psi}(x)| \le c$. By Theorem 3.12.1 in Luenberger (1963), there exists a $\underline{\psi}_0$ minimizing $\|\underline{\psi}\|_{L_2}$ within S_c. However, such a $\underline{\psi}_0$ may not be bounded by c everywhere; $\underline{\psi}_0$ can be modified on a P-null set such that $\sup_x \|\underline{\psi}_0\| \le c$ and $\|\underline{\psi}_0\|_{L_2}$ remains the same. Therefore, the modified $\underline{\psi}_0$ solves problem (V). The existence follows.

To prove uniqueness, suppose that problem (V) has two solutions, $\underline{\psi}_1$ and $\underline{\psi}_2$. Introduce $A = \{x: \underline{\psi}_1(x) = \underline{\psi}_2(x)\}$. Define the function $\underline{\psi}_0(x) = \{\underline{\psi}_1(x) + \underline{\psi}_2(x)\}/2$. Note that $\underline{\psi}_0(x)$ is also in S_c and

$$\int \|\underline{\psi}_0\|^2 \, dP = \int \|\tfrac{1}{2}\underline{\psi}_1 + \tfrac{1}{2}\underline{\psi}_2\|^2 \, dP = \int_A + \int_{A^C},$$

where A^C is the complement of the set A. By the strict convexity of $\|\cdot\|^2$, for any $x \in A^C$,

$$\|\underline{\psi}_0(x)\|^2 < \tfrac{1}{2}\|\underline{\psi}_1(x)\|^2 + \tfrac{1}{2}\|\underline{\psi}_2(x)\|^2.$$

Therefore, $\underline{\psi}_1$ would not be the solution of problem (V) had it been that $P(A) < 1$. Thus $\underline{\psi}_1(x)$ and $\underline{\psi}_2(x)$ are the same almost everywhere. □

Since $S_{c_1} \subset S_{c_2}$ whenever $c_1 < c_2$, S_{c_2} is nonempty as long as S_{c_1} is not empty. We would like to find the lowest c that makes S_c nonempty. In other words, we need to identify the lowest bound that an influence function can have.

For simplicity, we temporarily assume that the parameter θ is of one dimension. Let $\dot{\mathbf{l}}_\theta$ be the score function for θ and let us introduce the space $H = \dot{P}_2 \oplus \{1\} \subset L_2(P)$, where $\oplus$ is the linear sum of two linear spaces. Denote by $\overline{H}$ the L_1-closure of H. We need the following assumption for the existence of the optimal influence function:

(S) $$\dot{\mathbf{l}}_\theta \notin \overline{H}.$$

Condition (S) is to distinguish θ from the nonparametric component and can often be verified by checking the structures of $\dot{\mathbf{l}}_\theta$ and $\overline{H}$. For instance, in

the symmetric location model, the score function $\dot{\mathbf{l}}_\theta$ is antisymmetric, while $\overline{H}$ consists of symmetric functions. Therefore, condition (S) holds readily.

LEMMA 1. *Under condition* (S), *there always exists a function* ψ_0^* *satisfying the consistency conditions* (i)–(iii), *which minimizes* $\sup_x |\psi(x)|$ *among all influence functions.*

PROOF. Let $\overline{H}^\perp = \{\psi \in L_\infty(P)$: $\int \psi h\, dP = 0,\ \forall\ h \in \overline{H}\}$ be the orthogonal complement of $\overline{H}$ and $B = \{\psi \in \overline{H}^\perp,\ \sup_x |\psi(x)| \le 1\}$ be the unit ball in $\overline{H}^\perp$. By Theorem 1 in Luenberger [(1963), page 119],

$$0 < \min_{h \in \overline{H}} \int |\dot{\mathbf{l}}_\theta + h|\, dP = \max_{\psi \in B} \int \psi \dot{\mathbf{l}}_\theta\, dP. \tag{6}$$

Furthermore, the maximum on the right-hand side of (6) can always be achieved by a function $\psi^* \in B$. We assume that $\sup_x |\psi^*(x)| = 1$, or else we would consider $\psi^*(x)/\sup_x |\psi^*(x)|$ instead.

Note that $\int \psi^* \dot{\mathbf{l}}_\theta\, dP \neq 0$, which enables us to define $\psi_0^* = (\int \psi^* \dot{\mathbf{l}}_\theta\, dP)^{-1} \psi^*$. Then

$$\sup_x |\psi_0^*(x)| = \left(\int \psi^* \dot{\mathbf{l}}_\theta\, dP\right)^{-1} \sup_x |\psi^*(x)| = \left(\int \psi^* \dot{\mathbf{l}}_\theta\, dP\right)^{-1}. \tag{7}$$

Suppose ψ is another bounded influence function. Then ψ satisfies consistency conditions (i) and (iii) and thus $\psi \in \overline{H}^\perp$. By (7) and the definition of $\psi^*(x)$,

$$\int \frac{\psi}{\sup_x |\psi(x)|} \dot{\mathbf{l}}_\theta\, dP \le \int \psi^*(x) \dot{\mathbf{l}}_\theta\, dP = \frac{1}{\sup_x |\psi_0^*(x)|}.$$

Since ψ satisfies consistency condition (ii), $\sup_x |\psi_0^*(x)| \le \sup_x |\psi(x)|$. □

Next we shall extend the above lemma to the case in which θ is of high dimension. Before doing this, we would like to review some basic concepts in functional analysis. Let B be a Banach space and B^* be its dual space. Let a_n and a be vectors in B. We say a_n *converges weakly* to a if for any functional f in B^*, $f(a_n) \to f(a)$. We write $a_n \to_w a$ for the weak convergence. Correspondingly, suppose $f_n(\cdot)$ and $f(\cdot)$ are vectors in B^*. We say f_n converges to f in the weak* topology if for any $a \in B$, $f_n(a) \to f(a)$.

Let $\underline{\dot{\mathbf{l}}}_\theta$ be the vector of score functions defined by (4). Condition (S) needs to be modified:

$$a^T \underline{\dot{\mathbf{l}}}_\theta \notin \overline{H}, \qquad \forall\ a \in R^d,\ a \neq 0. \tag{S′}$$

It can be seen that condition (S′) ensures nonsingularity of the information matrix for θ.

THEOREM 1. *Under condition* (S′), *there always exists a function* $\underline{\psi}_0^*$ *satisfying the consistency conditions* (i)–(iii), *which minimizes* $\sup_x \|\underline{\psi}(x)\|$

among all influence functions. Conversely, if there exists a bounded influence function, then condition (S′) *holds.*

PROOF. Denote by M the set of all bounded functions satisfying consistency conditions (i)–(iii). One needs to show that M is nonempty. Condition (S′) implies that

$$\dot{\mathbf{l}}_i \notin H_i := \overline{H} \oplus \left\{\dot{\mathbf{l}}_j : j = 1, \ldots, d, j \neq i\right\}, \qquad i = 1, \ldots, d.$$

According to Lemma 1, for any i there exists a bounded function ψ_i^* which is orthogonal to H_i and satisfies $\int \psi_i^* \dot{\mathbf{l}}_i \, dP = 1$. This shows that $\underline{\psi}^* := (\psi_1^*, \ldots, \psi_d^*)^T$ belongs to M and thus M is nonempty.

Let $\lambda_0 = \min_{\underline{\psi} \in M} \|\underline{\psi}\|_\infty$ and $\underline{\psi}_n \in M$ such that $\lim_n \|\psi_n\|_\infty = \lambda_0$. Then $\{\underline{\psi}_n : n \geq 1\}$ is a bounded set in $L_\infty(P)$. The weak* compactness of the unit ball in $L_\infty(P)$ implies the existence of $\underline{\psi}_0^* \in M$ and a subsequence of $\{\underline{\psi}_n : n \geq 1\}$ (still denoted by $\underline{\psi}_n$), such that $\underline{\psi}_n$ tends to $\underline{\psi}_0$ in the weak* topology. It is easy to show that $\lambda_0 = \lim_n \|\underline{\psi}_n\|_\infty = \|\underline{\psi}_0\|_\infty$ and thus $\underline{\psi}_0^*$ minimizes $\sup_x \|\underline{\psi}(x)\|$ over all influence functions.

Conversely, assume that M contains a bounded influence function $\underline{\psi}_0$. Let $a \in R^d$ be a nonzero vector. Without loss of generality we can even assume that $\|a\| = 1$. Suppose $a^T \dot{\mathbf{l}}_\theta \in \overline{H}$. Then there is a series $\{h_n \in H : n \geq 1\}$ such that $h_n \to_{L_1} a^T \dot{\mathbf{l}}_\theta$. On the other hand, $\underline{\psi}_0$ satisfies conditions (i)–(iii) and therefore $\int (a^T \underline{\psi}_0) h_n \, dP = 0$ and $\int (a^T \underline{\psi}_0)(a^T \dot{\mathbf{l}}_\theta) \, dP = 1$. A contradiction comes from the fact that $0 = \int (a^T \underline{\psi}_0) h_n \, dP \to \int (a^T \underline{\psi}_0)(a^T \dot{\mathbf{l}}_\theta) \, dP = 1$. □

DEFINITION 2. An asymptotically linear estimate $\tilde{\theta}_n$ of θ is called the *most B-robust* estimate if its influence function minimizes the sup norm among all influence functions.

COROLLARY 1. *Let* $C_0 = \sup_x \|\psi_0^*(x)\|$. *Problem* (V) *always has a solution when the bound* $c \geq C_0$. *For any* $c > C_0$, *let* $\psi_c(\cdot)$ *be the optimal influence function corresponding to bound* c. *Then* $\int \psi_c^2(x) \, dP$ *is a decreasing function of* c.

PROOF. The first conclusion of the corollary follows directly from the above lemma and Proposition 1. The second conclusion is evident because the set S_c is increasing with c. □

4. Explicit optimal influence functions. Explicit expression of optimal influence functions is essential for the construction of the optimal estimates. It is usually difficult to achieve this by solving problem (V) directly. However, we can show that an influence function of a special form solves problem (V).

In one-dimensional parametric models, let $\dot{\mathbf{l}}_\theta$ be the score function. For $c > 0$, introduce the Huber truncation function $h_c(x) = \max(-c, \min(x, c))$. Hampel (1968) showed that the optimal influence function must be of the

form $h_c(a\dot{\mathbf{l}}_\theta + b)$, where a and b are constants such that $h_c(a\dot{\mathbf{l}}_\theta + b)$ satisfies consistency conditions (i) and (ii).

As for multidimensional parametric models, let $H_c(\underline{x})$: $R^d \to R^+$ be the multidimensional version of the Huber function defined by

$$H_c(\underline{x}) = \frac{h_c(\|\underline{x}\|)}{\|\underline{x}\|}\underline{x} \equiv \min\left(1, \frac{c}{\|\underline{x}\|}\right)\underline{x}.$$

Hampel, Ronchetti, Rousseeuw and Stahel (1986) expected the optimal influence function to be of a similar expression: $H_c(A\dot{\underline{\mathbf{l}}}_\theta + \underline{b})$ for some matrix A and a vector $\underline{b}$. Shen (1994) established the existence of A and $\underline{b}$ in general multidimensional parametric models.

In semiparametric models we would also expect the optimal influence functions to be of the above form. In the following we generalize Theorem 4.3.1 of Hampel, Ronchetti, Rousseeuw and Stahel (1986) to semiparametric models.

PROPOSITION 2. *Denote by $\bar{P}_2$ the L_1-closure of $\dot{P}_2$. If for $c > 0$, there exists a $d \times d$ matrix A, a d-vector $\underline{b}$ and a d-vector of functions $\underline{t}(x)$ in $\bar{P}_2$, such that the function $\underline{\psi}_0 := H_c(A\dot{\underline{\mathbf{l}}}_\theta + \underline{t}(x) + \underline{b})$ satisfies conditions* (i)–(iii), *then $\underline{\psi}_0$ solves problem* (V).

In the presence of a nuisance parameter $\eta \in R^{d'}$, the optimal influence function is expected to be of a similar form. In fact, if $\dot{\mathbf{l}}_\eta$ is the score function for η and if there exists a $d \times d'$ matrix B such that $\underline{\psi}_0 := H_c(A\dot{\underline{\mathbf{l}}}_\theta + B\dot{\underline{\mathbf{l}}}_\eta + \underline{t}(x) + \underline{b})$ satisfies conditions (i)–(iv), then $\underline{\psi}_0$ is an optimal influence function for θ. The proof is similar to the above Proposition 2. This can be compared with the optimal influence functions for partitioned parameters discussed in Section 4.4 of Hampel, Ronchetti, Rousseeuw and Stahel (1986).

Note that $\underline{t}(x)$ need not be a vector of functions in the L_2-space. Therefore, Proposition 2 narrows down the scope of functions we will consider.

4.1. *A special case.* We are able to provide explicit optimal influence functions for a special case. Following Huber (1964), we introduce a function $\rho_c(\cdot)$: $R \to R^+$ by

$$\rho_c(x) = \begin{cases} x^2/2, & \text{if } |x| \le c, \\ c^2/2 + c(|x| - c), & \text{if } |x| > c, \end{cases}$$

for $0 < c < \infty$, and by $\rho_0(x) = |x|$ and $\rho_\infty(x) = x^2$. It can be seen that $\rho_c'(x) = h_c(x)$.

LEMMA 2. *Suppose a random vector $\underline{\xi} \in R^d$ satisfies $E\|\underline{\xi}\| < \infty$. For any $c > 0$, define a function $f(\underline{a}) = E\rho_c(\|\underline{\xi} - \underline{a}\|)$. Let $\underline{a}_0$ minimize $f(\underline{a})$. Then $\|\underline{a}_0\| \le 2c + 3E\|\underline{\xi}\|$.*

PROOF. First we assume that $E\underline{\xi} = 0$. In this case, we shall show that $\|\underline{a}_0\| \le 2c + E\|\underline{\xi}\|$. If this were not true, we would have $\|\underline{a}_0\| > 2c + E\|\underline{\xi}\|$. Then by Jensen's inequality,

$$\begin{aligned} f(\underline{a}_0) &= E\rho_c\big(\|\underline{\xi} + \underline{a}_0\|\big) \\ &\ge \rho_c\big(\|E\underline{\xi} + \underline{a}_0\|\big) = \rho_c(\|\underline{a}_0\|) \\ &= c^2/2 + c\big(\|\underline{a}_0\| - c\big) \\ &> c^2/2 + c\big(E\|\underline{\xi}\| + c\big). \end{aligned}$$

However,

$$f(0) = E\rho_c\big(\|\underline{\xi}\|\big) \le c^2/2 + c\big(E\|\underline{\xi}\| + c\big) < f(\underline{a}_0).$$

This contradicts the fact that $\underline{a}_0$ minimizes $f(\cdot)$. If $E\underline{\xi} \ne 0$, we consider $\underline{\xi} - E\underline{\xi}$. From what we have just shown, $\|\underline{a}_0 + E\underline{\xi}\| \le 2c + E\|\underline{\xi} - E\underline{\xi}\|$, which implies $\|\underline{a}_0\| \le 2c + 3E\|\underline{\xi}\|$. □

Let $\{P_{\theta,\eta,g}\}$ be a semiparametric model, with $\theta \in R^d$ being the main parameter and $\eta \in R^{d'}$ being the nuisance parameter. Assume that the tangent space for $g(\cdot)$ is of the form

$$\dot{P}_2 = \{\nu(T)\colon \nu(T) \in L_2(P),\ E\nu(T) = 0\}, \tag{8}$$

where T is a fixed measurable function. In other words, $\dot{P}_2$ is generated by a fixed function T. As will be seen later, the examples that we consider are of this type. Let $\dot{\underline{\mathbf{l}}}_\theta$ and $\dot{\underline{\mathbf{l}}}_\eta$ be the score functions for θ and η, respectively. Again define $\overline{H}$ to be the L_1-closure of $\dot{P}_2 \oplus \{1\}$. It is evident that $\overline{H} = \{\nu(T)\colon \nu(T) \in L_1(P)\}$. Introduce the following condition:

$$(\mathrm{S}'') \qquad a^T\dot{\underline{\mathbf{l}}}_\theta + b^T\dot{\underline{\mathbf{l}}}_\eta \notin \overline{H} \qquad \forall\, a \in R^d,\ \forall\, b \in R^{d'},\ \|a\| + \|b\| \ne 0.$$

Define a $d \times d$ matrix $A = (a_{ij})$, with its off-diagonal elements a_{ij} varying freely and its diagonal elements satisfying the constraint

$$\sum_{i=1}^{d} a_{ii} = 1. \tag{9}$$

Therefore, the matrix A has $(d \times d - 1)$ free variables. Let (b_{ij}) be the elements of the $d \times d'$ matrix B and be allowed to vary freely. Denote by M the set of triples $(A, B, \underline{\nu}(T))$ such that A satisfies (9) and $\underline{\nu}(\cdot)$ belongs to $\overline{H}$. Define a functional from M to R^+ by

$$f(A, B, \underline{\nu}(\cdot)) = \int \rho_c\Big(\|A\dot{\underline{\mathbf{l}}}_\theta + B\dot{\underline{\mathbf{l}}}_\eta + \underline{\nu}(T)\|\Big)\, dP.$$

THEOREM 2. *Assume conditions* (S″) *and* (8) *hold. Then for any* $c > 0$ *there exist a scalar* $\lambda \ne 0$ *and a triple* $(A_0, B_0, \underline{\nu}(T)) \in M$ *that minimize* $f(\cdot)$ *over* M. *Moreover, the function* $\underline{\psi} \equiv H_c(A_0\dot{\underline{\mathbf{l}}}_\theta + B_0\dot{\underline{\mathbf{l}}}_\eta + \underline{\nu}(T))$ *satisfies conditions* (i), (iii), (iv) *and*

$$\int \underline{\psi}\dot{\underline{\mathbf{l}}}_\theta^T\, dP = \lambda I_{d\times d}. \tag{10}$$

Defining $\underline{\psi}_0 = \lambda^{-1}\psi$, *then* $\underline{\psi}_0$ *is the optimal influence function corresponding to bound* $\overline{C}_0 = \lambda^{-1}c$.

For some models it is relatively easy to calculate the conditional distribution of X given T. The above theorem provides a heuristic way to search for A, B and $\nu(\cdot)$. For instance, when both θ and η are of one dimension, one can start with a fixed b and find $\nu(b,t)$ to minimize $E\{\rho_c(\dot{\mathbf{l}}_\theta + b\dot{\mathbf{l}}_\eta + \nu)|T = t\}$ and then find b_0 to minimize

$$E\rho_c\left\{\dot{\mathbf{l}}_\theta + b\dot{\mathbf{l}}_\eta + \nu(b,T)\right\} \tag{11}$$

over all possible values of b. In fact, b_0 can be found by numerical calculation.

COROLLARY 2. *When* $d = d' = 1$, *there always exists* b_0 *to minimize* (11). *Define* $\lambda = \int h_c\{\dot{\mathbf{l}}_\theta + b_0\dot{\mathbf{l}}_\eta + \nu(b_0,T)\}\dot{\mathbf{l}}_\theta\, dP$. *Then* $\psi := \lambda^{-1}h_c\{\dot{\mathbf{l}}_\theta + b_0\dot{\mathbf{l}}_\eta + \nu(b_0,T)\}$ *is the optimal influence function with bound* $\lambda^{-1}c$.

4.2. *General cases.* For a general semiparametric model, it is possible to approximate the optimal influence function by a series of functions of those forms.

Let $\overline{H}$ be defined as in the previous section. Then $\overline{H}$ is a complete and separable space. Thus, $\overline{H}$ has a countable basis which can be denoted by $\{e_1(x),\ldots,e_n(x)\ldots\}$. Let H_n be the linear space spanned by $\{1,e_1,\ldots,e_n\}$. In the following theorem θ is assumed to be of one dimension, although this requirement is not essential.

THEOREM 3. *Let* $C_0 > 0$ *be defined as in Corollary* 1. *For any constant* $c \le C_0$, *let* ψ_c^* *be the influence function solving problem* (V) *corresponding to bound* c. *Then there exist a triangular array of real values* $\{d_i^n,\ i = 0,\ldots,n,\ n = 1,2\ldots\}$ *and a scalar* λ_n *such that the function* $\psi_n(x) = h_c(\lambda_n\dot{\mathbf{l}}_\theta + d_0^n + d_1^n e_1 + \cdots + d_n^n e_n)$ *satisfies*: *consistency conditions* (i) *and* (ii), $\psi_n(\cdot) \perp H_n$ *and* $\psi_n(\cdot)$ *tends to* ψ_c^* *in the* L_2*-norm.*

When $c = \infty$, $a_n(x) := \lambda_n^{-1}(d_0^n + d_1^n e_1 + \cdots + d_n^n)$ represents the projection of $\dot{\mathbf{l}}_\theta$ to H_n. Since H_n tends to $\overline{H}$ and $\dot{\mathbf{l}}_\theta$ is orthogonal to $\{1\}$, $a_n(\cdot)$ tends to $\pi(\dot{\mathbf{l}}_\theta|\overline{H})$ as n increases. The λ_n and d_i^ns can be calculated numerically, and therefore Theorem 3 gives an approximation to the efficient influence function.

5. Optimal estimates. Let $x_1,\ldots,x_n$ be i.i.d. samples from $P_{\theta,g}$. In order to construct the optimal estimates, it is necessary to estimate the optimal influence functions. In particular, when the bound $c = \infty$, the problem becomes constructing the efficient estimates. There is a general approach in BKRW to constructing the efficient estimates, although no explicit estimates are provided for general semiparametric models.

When the bound $c < \infty$, one needs explicit optimal influence functions before being able to construct the optimal estimates. We believe that satisfactory results can only be given for special models. As a general rule, we can apply the one-step procedure explained in Chapter 7 of BKRW. Specifically, we start with a preliminary $\sqrt{n}$-consistent estimate $\tilde{\theta}_n$, and then estimate the score function $\underline{\mathbf{l}}_\theta$. The next step is to estimate the optimal influence function by, say, $\hat{\psi}_n$. This is possible when the optimal influence functions have explicit forms. In some cases we can show that $\hat{\psi}_n$ satisfies conditions (1.4) and (1.5) of Klaassen (1987). The one-step estimate is then defined by

$$\hat{\theta}_n = \tilde{\theta}_n + \frac{1}{n}\sum_{i=1}^{n} \hat{\psi}_n(x_i).$$

Klaassen (1987) provided conditions under which $\hat{\theta}_n$ is the optimal estimate.

When the explicit forms of the optimal influence functions are not available, which may happen not only to the case $c < \infty$, but also to the case $c = \infty$, it is usually difficult, if not impossible, to estimate the optimal influence functions. Theorem 3 enables us to work on the subspaces of $\dot{P}_2$. We can estimate the optimal influence function by estimating λ_n and d_i^n's. The one-step estimates can again be calculated from the estimated optimal influence functions. It is evident that a successful construction requires careful selection of the dimensions of the subspaces. We have not attempted a rigorous approach in this direction. However, Shen and Wong (1994) were able to construct the nonparametric maximum likelihood estimates by the method of sieves. They used the maximum likelihood estimates over parametric submodels to approximate the nonparametric maximum likelihood estimate for a semiparametric model. They also calculated the rate of convergence for the approximation. We hope that their approach could provide insight to solving our problem.

6. Examples (Continued). We revisit the examples introduced in the Introduction and calculate optimal influence functions for those models.

EXAMPLE 1. *Symmetric location model.* The underlying density is denoted by $g(x - \theta)$; the score function for θ is $\dot{\mathbf{l}}_\theta = -g'(x - \theta)/g(x - \theta)$; the tangent space consists of symmetric functions of $(x - \theta)$ that have zero expectations. For $c > 0$, define $\lambda_c = \int h_c(\dot{\mathbf{l}}_\theta)\dot{\mathbf{l}}_\theta \, dP_{\theta, g}$. The optimal influence function for this model has the following simple form:

$$\psi_c(x, \theta, g) = \lambda_c^{-1} h_c\left(-\frac{g'}{g}(x - \theta)\right).$$

The one-step procedure can be used to construct the optimal estimator. One can start with a preliminary estimate, say, the median $\tilde{\theta}_n$. The $\tilde{\theta}_n$ can be discretized in the following way: form a grid of cubes with sides of length $n^{-1/2}$ over R^d and the discretized $\tilde{\theta}_n$ is the midpoint of the cube into which $\tilde{\theta}_n$ has fallen. The discretized median may be again denoted by $\tilde{\theta}_n$. When n is

large, $\tilde{\theta}_n$ will fall into a compact set in which there are only a finite number of cubes with sides of length $n^{-1/2}$. Hence $\tilde{\theta}_n$ may be considered deterministic for the purpose of applying asymptotic theories. Having done this, we calculate the residuals $\varepsilon_i = x_i - \tilde{\theta}_n$, $i = 1, \ldots, n$. Following Bickel (1982) or Klaassen (1987), we can split the residuals into $\varepsilon_1, \ldots, \varepsilon_{n_1}$ and $\varepsilon_{n_1+1}, \ldots, \varepsilon_{n_1+n_2}$, where $n_1 \approx \alpha n$, α is a constant and $n_2 = n - n_1$. Note that Bickel (1982) used an n_1 that tends to infinity at a lower rate than n.

One can estimate the score function from $\varepsilon_1, \ldots, \varepsilon_{n_1}$ by $\hat{\mathbf{I}}_n(x) = \{\tilde{\mathbf{I}}_n(x) - \tilde{\mathbf{I}}_n(-x)\}/2$, where $\tilde{\mathbf{I}}_n$ is calculated in the same way as in (6.3) of Bickel (1982). For $c > 0$, we estimate the optimal influence function by $\hat{\psi}_n(x) = \lambda_n^{-1}(c) h_c(\hat{\mathbf{I}}_n(x))$ with

$$\lambda_n(c) = \frac{1}{n_2} \sum_{i=n_1+1}^{n} h_c\big(\hat{\mathbf{I}}_n(\varepsilon_i)\big)\hat{\mathbf{I}}_n(\varepsilon_i).$$

Following the approach of Bickel (1982), it can be shown that the one-step estimate

$$\hat{\theta}_n = \tilde{\theta}_n + \frac{1}{n_2} \sum_{i=n_1+1}^{n} \hat{\psi}_n(\varepsilon_i)$$

is the optimal estimate with influence function ψ_c.

EXAMPLE 2. *Heteroscedastic regression model.* Let

$$f(x, y; \theta, \eta, g) = \exp(-r(x, \eta))g(\varepsilon)k(x)$$

be the joint density of (X, Y), where $\varepsilon = \exp(-r(x, \eta))(y - \theta^T x)$ and $k(x)$ is the density of X. The score function for θ is given by

$$\underline{\dot{\mathbf{I}}}_\theta(x, y; \theta, \eta, g) = \nabla_\theta \log(f) = -\frac{g'}{g}(\varepsilon)x \exp(-r(x, \eta)).$$

Similarly, the score function for η is

$$\underline{\dot{\mathbf{I}}}_\eta(x, y; \theta, \eta, g) = \nabla_\eta \log(f) = -\left\{\frac{g'}{g}(\varepsilon)\varepsilon + 1\right\}\nabla_\eta r(x, \eta).$$

The tangent space for $g(\cdot)$ can be written as

$$\dot{P}_2 = \{a(\varepsilon): a(\cdot) \text{ is symmetric about 0 and } Ea(\varepsilon) = 0\}.$$

We expect the optimal influence function for θ to be of the form

$$\underline{\psi} = H_c\left\{A\underline{\dot{\mathbf{I}}}_\theta + B\underline{\dot{\mathbf{I}}}_\eta + \underline{a}(\varepsilon)\right\} \tag{12}$$

for some matrices A, B and for a vector of symmetric functions $\underline{a}(\cdot)$. For this model an influence function $\underline{\psi}$ will satisfy consistency conditions (i)–(iv). One

needs the following condition for the existence of the optimal influence function of the form (12):

(R) For any vector $\underline{\alpha} = (\alpha_1, \ldots, \alpha_d)^T$,
$$P\left(\underline{\alpha}^T \dot{\underline{\mathbf{l}}}_\theta = 0\right) = 1 \text{ implies that } \underline{\alpha} = 0.$$

Note that condition (R) implies condition (S″) for this model. For any $c > 0$, we try to find a matrix $A_0(\theta, \eta, g)$ to minimize

$$\int \rho_c\left(\|A\dot{\underline{\mathbf{l}}}_\theta(x, y; \theta, \eta, g)\|\right) dP_{(\theta,\eta,g)}(x, y) \tag{13}$$

over all matrices $A = (a_{ij})$ satisfying constraint (9).

THEOREM 4. *Under condition* (R), *for any* $c > 0$ *there exists a matrix* $A_0(\theta,\eta, g)$ *minimizing* (13) *over all* A *satisfying constraint* (9). *Moreover*, $H_c(A_0(\theta, \eta, g)\dot{\underline{\mathbf{l}}}_\theta)$ *satisfies conditions* (i), (iii) *and* (iv), *and for some nonzero scalar* $\lambda_0(\theta, \eta, g)$,

$$\int H_c\left\{A_0(\theta, \eta, g)\dot{\underline{\mathbf{l}}}_\theta^T\right\} dP_{(\theta,\eta,g)} = \lambda_0(\theta, \eta, g) I_{d\times d}.$$

Therefore, $\psi_0(\cdot; \theta, \eta, g) := \lambda_0(\theta, \eta, g)^{-1} H_c\{A_0(\theta, \eta, g)\dot{\underline{\mathbf{l}}}_\theta\}$ *is the optimal influence function corresponding to bound* $\lambda_0(\theta, \eta, g)^{-1}c$.

PROOF. The proof follows the same lines as those of Theorem 2. □

The one-step procedure can also be used to construct the optimal estimates. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be i.i.d. random samples from model (2). Assume that $\tilde{\theta}_n$ and $\tilde{\eta}_n$ are the preliminary $\sqrt{n}$-consistent estimates of θ and η, respectively, that is, $\sqrt{n}(\tilde{\theta}_n - \theta) = O_P(1)$ and $\sqrt{n}(\tilde{\eta}_n - \eta) = O_P(1)$. The existence of $\tilde{\theta}_n$ and $\tilde{\eta}_n$ will not be discussed here. The estimates $\tilde{\theta}_n$ and $\tilde{\eta}_n$ can also be discretized so that they may be considered deterministic.

One can compute the residuals: $\varepsilon_i = (y_i - \tilde{\theta}_n^T x_i)\exp\{-r(x_i, \tilde{\eta}_n)\}$, $i = 1, \ldots, n$. As in Example 1, split the residuals into $\{\varepsilon_1, \ldots, \varepsilon_{n_1}\}$ and $\{\varepsilon_{n_1+1}, \ldots, \varepsilon_{n_1+n_2}\}$, where $n = n_1 + n_2$ and $n_1 \approx \alpha n$ with α being a constant. The first part of the ε's can be used to estimate $(g'/g)(x)$ in the same way as in (6.3) of Bickel (1982). Assume that $\hat{q}_n(x)$ is an antisymmetric version of the estimate. We can estimate the score function $\dot{\underline{\mathbf{l}}}_\theta(x, y; \theta, \eta, g)$ by

$$\hat{\mathbf{l}}_n(x, y; \theta, \eta) = \hat{q}_n\{(y - \theta^T x)\exp(-r(x, \eta))\}x \exp(-r(x, \eta)).$$

For $c > 0$, let matrix A_n minimize among all A satisfying (9) the following expression:

$$\begin{aligned}&\int\int \rho_c\left\{\left\|A\hat{\mathbf{l}}_n\left(x, y; \tilde{\theta}_n, \tilde{\eta}_n\right)\right\|\right\}\exp(-r(x, \tilde{\eta}_n)) \\ &\quad\times \hat{g}_n\{(y - \theta^T x)\exp(-r(x, \tilde{\eta}_n))\}\, dy\, k(x)\, dx \\ &\quad\equiv \int\int \rho_c\{\|A\hat{q}_n(y)\, x \exp(-r(x, \tilde{\eta}_n))\|\}\hat{g}_n(y)\, dy\, k(x)\, dx.\end{aligned} \tag{14}$$

The same approach as in Theorem 4 shows that for some nonzero scalar λ_n, A_n will satisfy

$$\int\int H_c\{A_n\hat{q}_n(y)x\exp(-r(x,\tilde{\eta}_n))\}\hat{q}_n(y)x^T \times \exp(-r(x,\tilde{\eta}_n))\hat{g}_n(y)\,dy\,k(x)\,dx = \lambda_n I_{d\times d}.$$

The optimal influence function $\psi_0(x, y; \beta, \theta, g)$ can be estimated by

$$\hat{\psi}_n(x, y; \theta, \eta) = \lambda_n^{-1}H_c\{A_n\hat{\mathbf{I}}_n(x, y; \theta, \eta)\}.$$

Finally, apply the one-step procedure to obtain the optimal robust estimate

$$\hat{\theta}_n = \tilde{\theta}_n + \frac{1}{n_2}\sum_{i=n_1+1}^{n}\hat{\psi}_n(x_i, y_i; \tilde{\theta}_n, \tilde{\eta}_n). \tag{15}$$

Under certain conditions, it can be shown that $\hat{\theta}_n$ is an asymptotically linear estimate of θ, with influence function ψ_0 defined by Theorem 4.

EXAMPLE 3. *Semiparametric mixture model.* We consider a special kind of mixture models, that is, the exponential mixture models, in which, for given η,

$$f(x, \theta, \eta) = \exp\{\eta^T T(x, \theta) + S(x, \theta) - b(\theta, \eta)\}. \tag{16}$$

The incidental parameter η has a distribution $G(\cdot)$.

For simplicity, assume that θ is of one dimension. Let $\dot{T}(x, \theta)$, $\dot{S}(x, \theta)$ and $\dot{b}(\theta, \eta)$ be the partial derivatives of $T(x, \theta)$, $S(x, \theta)$ and $b(\theta, \eta)$ with respect to θ. Theorem 4.5.1 and Corollary 4.5.1 of BKRW show that the score function for θ is

$$\dot{\mathbf{I}}_\theta(X, \theta, G) = \dot{T}(X, \theta)E(\eta|T) + \dot{S}(X, \theta) - E\{\dot{b}(\theta, \eta)|T\} \tag{17}$$

and that the tangent space for G is

$$\dot{\mathbf{P}}_2 = \{w(T(X, \theta)): w(T(X, \theta)) \in L_2(P) \text{ and } Ew(T(X, \theta)) = 0\}.$$

The optimal influence function is expected to be of the form

$$\psi = \lambda^{-1}h_c\{\dot{\mathbf{I}}_\theta + w_0(T) + \alpha_0\}, \tag{18}$$

where $w_0(T) \in \dot{P}_2$ and λ and a_0 are constants.

THEOREM 5. *For any $c > 0$, let $w(t, \theta, G)$ be the solution of the equation*

$$E\{h_c(\dot{\mathbf{I}}_\theta(X, \theta, G) + w)|T(X, \theta) = t\} = 0. \tag{19}$$

Define $\psi_1 = h_c\{\dot{\mathbf{I}}_\theta(x, \theta, G) + w(T(x, \theta), \theta, G)\}$ and $\lambda = \int\psi_1\dot{\mathbf{I}}_\theta\,dP_{\theta,G}$. Then $\psi := \lambda^{-1}\psi_1$ is the B-optimal influence function with bound $\lambda^{-1}c$.

PROOF. It is evident that ψ can be written in the form of (18) with

$$a_0 = \lambda^{-1}Ew(T(x, \theta), \theta, G) \quad \text{and} \quad w_0(T) = w(T) - Ew(T) \in \dot{P}_2.$$

Consistency conditions (i) and (iii) follow from the fact that $E(\psi|T(x,\theta)=t)=0$ for any t, and condition (ii) is straightforward. □

A special case of model (16) is when $T(x,\theta)\equiv T(x)$ and $S(x,\theta)\equiv S(x)\theta$. The score function in (17) becomes

$$\dot{\mathbf{l}}_\theta(x,\theta,G)=S(x)-E\big(\dot{b}(\theta,\eta)|T\big).$$

The optimal influence function reduces to

$$\psi(x)=\lambda^{-1}h_c\{S(x)+w(T(x),\theta)\},$$

where $w(t,\theta)$ solves the equation

$$E\{h_c(S(x)+w)|T=t\}=0. \tag{20}$$

In this situation, the problem is equivalent to robust estimation for the parametric model derived by conditioning X on T. Since the conditional density of X on T does not depend on the unknown mixing distribution G, the calculation of the optimal influence function becomes much simpler in this case.

A famous example of a mixture model is the Neyman–Scott (scale) model, in which a random vector $X=(X^{(1)},\dots,X^{(k)})^T$ has components that are independently and identically distributed according to $\mathrm{N}(\nu,\sigma^2)$. The main parameter is $\theta=-(2\sigma^2)^{-1}$; the nuisance parameter is $\eta=\nu\sigma^{-2}$. The conditional density of X for given η is

$$f(x,\theta,\eta)=\exp\left\{\eta\sum_{j=1}^{k}x^{(j)}+\theta\sum_{j=1}^{k}(x^{(j)})^2-\frac{k}{2}\left(-\frac{\eta^2}{2\theta}+\ln\left(-\frac{\pi}{\theta}\right)\right)\right\}.$$

The nuisance parameter η is assumed to follow a distribution $G(\cdot)$. Then the class of (unconditional) distributions of X constitutes a mixture model. Define

$$T(X)=\sum_{j=1}^{k}X^{(j)},\qquad S(X)=\sum_{j=1}^{k}(X^{(j)})^2\quad\text{and}$$

$$b(\theta,\eta)=\frac{k}{2}\left\{-\frac{\eta^2}{2\theta}+\ln\left(-\frac{\pi}{\theta}\right)\right\}.$$

The score function for θ is $\dot{\mathbf{l}}_\theta=S(x)-E(\dot{b}(\theta,\eta)|T)$. For $c>0$, choose $w(T)$ such that

$$E\big\{h_c\big(\dot{\mathbf{l}}_\theta+w\big)|T\big\}=0. \tag{21}$$

Let $U(X)=-2\theta\sum_{j=1}^{k}(X^{(j)}-T(X)/k)^2$. Then $S(X)=-U(X)/2\theta+T^2(X)/k$. It is well known that $U(X)$ is independent of $T(X)$ and has a χ^2_{k-1} distribution. Write $g_{k-1}(u)$ for the χ^2_{k-1} distribution density. Then (21) is equivalent to

$$\int h_c\left(-\frac{1}{2\theta}u+w_1\right)g_{k-1}(u)\,du=0.$$

Note that $w_1(\cdot)$ is a function depending only on θ. Since g_{k-1} is a nonatomic density, it is easy to show that $w_1(\theta)$ is a continuous and strictly monotonic function of θ.

Let $\lambda(\theta) = \int h_c\{-U(X)/2\theta + w_1(\theta)\}S(X)\, dP_{\theta,G}$ and $\psi_c := \lambda^{-1}(\theta) \times h_c\{-U(X)/2\theta + w_1(\theta)\}$. Theorem 5 implies that ψ_c is an optimal influence function corresponding to bound $\lambda^{-1}(\theta)c$.

We shall construct the optimal estimates corresponding to the optimal influence function just derived. Suppose we have i.i.d. samples $X_i = (X_i^{(1)}, \ldots, X_i^{(k)})$, $i = 1, \ldots, n$, from the Neyman–Scott (scale) model. Let $\hat{\theta}_n$ be the M-estimate solving equation

$$\sum_{i=1}^{n} h_c\left(\sum_{j=1}^{k} \left(X_i^{(j)} - T(X_i)/k\right)^2 + w_1(\theta) \right) = 0.$$

Using the standard approach for parametric models, we can easily show that $\hat{\theta}_n$ is the optimal estimate corresponding to ψ_c.

In an extreme case, $c = \infty$, it can be seen that $w_1(\theta) = -(k-1)\sigma^2$ and $\hat{\theta}_n$ becomes the efficient estimate of θ, which in turn gives the efficient estimate of σ^2:

$$\hat{\sigma}_n^2 = \frac{1}{n(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{k} \left(X_i^{(j)} - T(X_i)/k\right)^2.$$

Another example is the two sample exponential mixture model where, for given η, a bivariate random vector $X = (x^{(1)}, x^{(2)})^T$ has a conditional density

$$f(x_1, x_2, \theta, \eta) = \exp\{-\eta(x_1 + \theta x_2) + \ln(\eta^2\theta)\}, \qquad x_1, x_2 > 0.$$

The nuisance parameter η is distributed according to $G(\cdot)$. Define $T(X, \theta) = -(X^{(1)} + \theta X^{(2)})$ and $b(\theta, \eta) = -\ln(\eta^2\theta)$. The score function for θ is

$$\dot{\mathbf{l}}_\theta(X, \theta, G) = -X^{(2)}E(\eta|T) - E\left(\dot{b}(\theta, \eta)|T\right).$$

The marginal distribution of T is

$$p_T(t, G) = -\text{Constant} \times t \int \eta^2 \exp(\eta t)\, dG(\eta), \quad \text{for } t < 0.$$

The conditional density of $X^{(2)}$ given T is $-(T/\theta)\text{Beta}(1, 1)$, where Beta(1, 1) is the beta distribution. Chapter 4 of BKRW shows $E(\eta|T) = (p_T'/p_T)(t) - T^{-1}$. To obtain the optimal influence function, we solve w from equation

$$E\{h_c(-X^{(2)}E(\eta|T) + w)|T\} = 0.$$

By the symmetry of the beta distribution, $w(T) = -(T/2\theta)E(\eta|T)$. Then the optimal influence function is given by

$$\psi_c(x^{(1)}, x^{(2)}; \theta, G) = \lambda_c^{-1} h_c\left\{ \left(\frac{x^{(1)} - \theta x^{(2)}}{2\theta} \right) E(\eta|T) \right\}, \tag{22}$$

where λ_c is the normalizing constant.

Let $(x_1^{(1)}, x_1^{(2)}), \ldots, (x_n^{(1)}, x_n^{(2)})$ be i.i.d. samples from $P_{\theta,G}$. We construct the optimal estimate for the extreme case $c = 0$. In this case, the optimal influence function becomes $\psi_0 = \lambda^{-1}\,\text{sgn}(X^{(1)} - \theta X^{(2)})$. Since the conditional

expectation $E(\eta|T)$ is always nonnegative, it disappears from (22). The estimate $\hat{\theta}_n := \text{med}\{X_i^{(1)}/X_i^{(2)}: i = 1, \ldots, n\}$ solves equation

$$\sum_{i=1}^{n} \text{sgn}\left(x_i^{(1)} - \theta x_i^{(2)}\right) = 0.$$

Thus, $\hat{\theta}_n$ is the most robust estimate with influence function ψ_0.

7. Technical details.

PROOF OF PROPOSITION 2. For any function $\underline{\psi}$,

$$\|\underline{\psi}\|^2 = \|\underline{\psi} - A\dot{\underline{\mathbf{l}}}_\theta - \underline{t} - \underline{b}\|^2 + 2\underline{\psi}^T\left(A\dot{\underline{\mathbf{l}}}_\theta + \underline{t} + \underline{b}\right) - \|A\dot{\underline{\mathbf{l}}}_\theta + \underline{t} + \underline{b}\|^2. \tag{23}$$

Let $\underline{t}_n(\cdot)$ be a series of functions in $\dot{P}_2$ converging to $\underline{t}(\cdot)$ in the L_1-norm and satisfying $\|t_n(x)\| \le \|t(x)\|$, for all x. Define $\underline{\psi}_n = H_c(A\dot{\underline{\mathbf{l}}}_\theta + \underline{t}_n + \underline{b})$. Then by (23),

$$\begin{aligned}\int\|\underline{\psi}_n\|^2\,dP &= \int\|\underline{\psi}_n - A\dot{\underline{\mathbf{l}}}_\theta - \underline{t}_n - \underline{b}\|^2\,dP \\ &\quad + 2\int\underline{\psi}_n^T\left(A\dot{\underline{\mathbf{l}}}_\theta + \underline{t}_n + \underline{b}\right)dP - \int\|A\dot{\underline{\mathbf{l}}}_\theta + \underline{t}_n + \underline{b}\|^2\,dP.\end{aligned} \tag{24}$$

Note that $\underline{\psi}_n$ minimizes $\int\|\underline{\psi} - A\dot{\underline{\mathbf{l}}}_\theta - \underline{t}_n - \underline{b}\|^2\,dP$ among all functions bounded by c. Thus if $\underline{\psi}$ satisfies consistency conditions (i)–(iii) and is bounded by c, then (24) implies

$$\begin{aligned}&\int\|\underline{\psi}_n\|^2\,dP \\ &\quad\le \int\|\underline{\psi} - A\dot{\underline{\mathbf{l}}}_\theta - \underline{t}_n - \underline{b}\|^2\,dP + 2\int\underline{\psi}_n^T\left(A\dot{\underline{\mathbf{l}}}_\theta + \underline{t}_n + \underline{b}\right)dP \\ &\qquad - \int\|A\dot{\underline{\mathbf{l}}}_\theta + \underline{t}_n + \underline{b}\|^2\,dP \\ &\quad= \int\|\underline{\psi}\|^2\,dP + 2\int\underline{\psi}_n^T\left(A\dot{\underline{\mathbf{l}}}_\theta + \underline{t}_n + \underline{b}\right)dP - 2\int\underline{\psi}^T\left(A\dot{\underline{\mathbf{l}}}_\theta + \underline{t}_n + \underline{b}\right)dP.\end{aligned} \tag{25}$$

Since $\underline{\psi}_n$ is bounded by c and $\underline{t}_n(\cdot) \to \underline{t}(\cdot)$ in the L_1-norm, the dominated convergence theorem yields $\underline{\psi}_n \to \underline{\psi}_0$ in probability P and thus in the L_2-norm as well. Because $\underline{\psi}_0$ satisfies consistency conditions (i)–(iii),

$$\begin{aligned}&\lim_n \int\underline{\psi}_n^T\left(A\dot{\underline{\mathbf{l}}}_\theta + \underline{t}_n + \underline{b}\right)dP \\ &\quad= \int\underline{\psi}_0^T\left(A\dot{\underline{\mathbf{l}}}_\theta + \underline{t} + \underline{b}\right)dP \\ &\quad= \text{tr}\int\underline{\psi}_0\left(A\dot{\underline{\mathbf{l}}}_\theta + \underline{t} + \underline{b}\right)^T dP = \text{tr}\int\underline{\psi}\left(A\dot{\underline{\mathbf{l}}}_\theta + \underline{t} + \underline{b}\right)^T dP \\ &\quad= \int\underline{\psi}^T\left(A\dot{\underline{\mathbf{l}}}_\theta + \underline{t} + \underline{b}\right)dP.\end{aligned}$$

By (25), $\int \|\underline{\psi}_0\|^2\, dP = \liminf_n \int \|\underline{\psi}_n\|^2\, dP \le \int \|\underline{\psi}\|^2\, dP$, which shows that $\underline{\psi}_0$ solves problem (V). □

PROOF OF THEOREM 2. Let $(A_n, B_n, \underline{\nu}_n(\cdot)) \in M$ be such that

$$\lim_n f(A_n, B_n, \underline{\nu}_n(\cdot)) = m := \min_{(A,B,\underline{\nu})\in M} f(A, B, \underline{\nu}(\cdot)).$$

Condition (S″) implies that all the elements of A_n and B_n have to be bounded. There exist subsequences of A_n and B_n (still denoted by A_n and B_n) which converge to some matrices A_0 and B_0, respectively. It can be shown that A_0 continues to satisfy constraint (9). For any fixed t, let $\underline{\nu}_n'(t)$ minimize $E\{\rho_c(\|A_n\underline{\dot{\mathbf{i}}}_\theta + B_n\underline{\dot{\mathbf{i}}}_\eta + \underline{\nu}\|)|T = t\}$. Then

$$E\Big\{\rho_c\Big(\|A_n\underline{\dot{\mathbf{i}}}_\theta + B_n\underline{\dot{\mathbf{i}}}_\eta + \underline{\nu}_n'(T)\|\Big)|T = t\Big\}$$
$$\le E\Big\{\rho_c\Big(\|A_n\underline{\dot{\mathbf{i}}}_\theta + B_n\underline{\dot{\mathbf{i}}}_\eta + \underline{\nu}_n(T)\|\Big)|T = t\Big\}.$$

Hence $E\{\rho_c(\|A_n\underline{\dot{\mathbf{i}}}_\theta + B_n\underline{\dot{\mathbf{i}}}_\eta + \underline{\nu}_n'(T)\|)\} \to m$. On the other hand, Lemma 2 yields

$$\|\underline{\nu}_n'(t)\| \le 2c + 3E\Big(\|A_n\underline{\dot{\mathbf{i}}}_\theta + B_n\underline{\dot{\mathbf{i}}}_\eta\|\,|T = t\Big).$$

Thus, $\underline{\nu}_n'(T) \in L_2(P)$ and $\{\underline{\nu}_n'(T)\colon n \ge 1\}$ is a bounded set in $L_2(P)$. By weak compactness, there is a subsequence of $\{\underline{\nu}_n'(T)\}$ [still denoted by $\nu_n'(T)$] and a function $\underline{\nu}_0(T) \in \overline{H}$, such that $\underline{\nu}_n'(T) \to_w \underline{\nu}_0(T)$ in $L_2(P)$. Since the functional $f(A, B, \underline{\nu}(\cdot))$ is convex and is strongly continuous in the sense of the L_2-norm, it follows from Proposition 38.7 in Zeidler (1985) that

$$f(A_0, B_0, \underline{\nu}_0(T)) \le \liminf_n f(A_n, B_n, \underline{\nu}_n'(T)).$$

Thus the triple $(A_0, B_0, \underline{\nu}_0(\cdot))$ belongs to M and minimizes $f(\cdot)$ over all $(A, B, \underline{\nu}(\cdot)) \in M$. Let $\underline{\psi} = H_c(A_0\underline{\dot{\mathbf{i}}}_\theta + B_0\underline{\dot{\mathbf{i}}}_\eta + \underline{\nu}_0(T))$. Taking the partial derivatives of $f(\cdot)$ with respect to elements of B_0 shows that $\underline{\psi}$ is orthogonal to $\mathring{\underline{\mathbf{i}}}_\eta$ and therefore condition (iv) follows. Taking the Gateaux derivatives of $f(\cdot)$ with respect to $\underline{\nu}_0$ yields conditions (i) and (iii). Finally it remains to establish (10). Write $\underline{\psi} = (\psi_1, \ldots, \psi_d)^T$ and $\underline{\dot{\mathbf{i}}}_\theta = (\dot{\mathbf{i}}_1, \ldots, \dot{\mathbf{i}}_d)^T$. We first take the partial derivatives of $f(\cdot)$ with respect to the off-diagonal elements of A and obtain

$$\int \psi_i\dot{\mathbf{i}}_j\, dP = 0, \qquad i = 1, \ldots, d, j = 1, \ldots, d, i \ne j.$$

By constraint (9), $a_{dd} = 1 - a_{11} - \cdots - a_{(d-1)(d-1)}$. Taking the partial derivatives of $f(\cdot)$ with respect to $a_{11}, \ldots, a_{(d-1)(d-1)}$ gives

$$\int \psi_i\dot{\mathbf{i}}_i\, dP - \int \psi_d\dot{\mathbf{i}}_d\, dP = 0, \qquad i = 1, \ldots, (d-1).$$

Define $\lambda = \int \psi_d\dot{\mathbf{i}}_d\, dP$. Then $\int \psi_i\dot{\mathbf{i}}_i\, dP = \lambda$, $i = 1, \ldots, d$, which completes the proof. □

PROOF OF THEOREM 3. Repeating the proof of Lemma 1 we can show that there is the lowest bound C_n, for functions satisfying conditions (i) and (ii) and being orthogonal to H_n. Clearly, C_n is less than or equal to C_0 and must in turn be less than or equal to c. From Corollary 4 of Shen (1994), for any $c > C_n$ there exist a scalar λ_n and a function

$$a_n(x) := d_0^n + d_1^n e_1 + \cdots + d_n^n e_n \in H_n$$

such that $\psi_n(x) := h_c(\lambda_n \dot{\mathbf{l}}_\theta + a_n(x))$ is orthogonal to H_n and $\int \psi_n \dot{\mathbf{l}}_\theta \, dP = 1$.

Since $\sup_x |\psi_n(x)| \le c$, $\{\psi_n : n \ge 1\}$ is a bounded set in $L_\infty(P)$. By Alaoglu's theorem [see, e.g., Dunford and Schwartz (1966)], there is a subsequence of ψ_n, which can still be denoted by ψ_n, and a function $\psi_0 \in L_\infty(P)$, such that ψ_n tends to ψ_0 in the weak* topology. We shall show that ψ_0 is the optimal influence function corresponding to bound c. In fact, as a result of weak* convergence,

$$\int \psi_0 \dot{\mathbf{l}}_\theta \, dP = \lim_n \int \psi_n \dot{\mathbf{l}}_\theta \, dP = 1, \tag{26}$$

and for any $k \ge 1$ and any $t(\cdot) \in H_k$,

$$\int \psi_0 t(x) \, dP = \lim_n \int \psi_n t(x) \, dP = 0.$$

This implies that $\psi_0 \perp H_k$, $\forall\, k$, and, in turn, $\psi_0 \perp \overline{H}$. Therefore, ψ_0 satisfies consistency conditions (i)–(iii). Next we need to show that ψ_0 minimizes $\int \psi^2 \, dP$ among all influence functions bounded by c. Suppose ψ_1 is another influence function also bounded by c. Then since ψ_n is given by truncating $\lambda_n \dot{\mathbf{l}}_\theta + a_n(x)$ at $\pm c$,

$$\int \left(\psi_n - \lambda_n \dot{\mathbf{l}}_\theta - a_n(x)\right)^2 dP \le \int \left(\psi_1 - \lambda_n \dot{\mathbf{l}}_\theta - a_n(x)\right)^2 dP.$$

Since both ψ_n and ψ_1 satisfy consistency conditions (i) and (ii) and are orthogonal to $a_n(x)$, the above inequality implies

$$\int \psi_n^2 \, dP \le \int \psi_1^2 \, dP + 2\lambda_n \int \psi_n \dot{\mathbf{l}}_\theta \, dP - 2\lambda_n \int \psi_1 \dot{\mathbf{l}}_\theta \, dP = \int \psi_1^2 dP.$$

The weak* convergence argument and the Cauchy–Schwarz inequality lead to

$$\int \psi_0^2 \, dP = \lim_n \int \psi_0 \psi_n \, dP \le \liminf_n \left(\int \psi_0^2 \, dP\right)^{1/2} \left(\int \psi_n^2 \, dP\right)^{1/2}, \tag{27}$$

implying

$$\int \psi_0^2 \, dP \le \liminf_n \int \psi_n^2 \, dP \le \int \psi_1^2 \, dP. \tag{28}$$

It thus remains to prove that $\operatorname{ess\,sup}_x |\psi_0(x)| \le c$. However, this follows from

$$\left|\int \psi_0 f dP\right| = \left|\lim_n \int \psi_n f dP\right| \le c \int |f| \, dP, \qquad \forall\, f \in L_1(P).$$

Therefore, ψ_0 solves problem (V) corresponding to bound c. By uniqueness of the optimal influence function, $\psi_0 \equiv \psi_c^*$.

Finally, we derive the strong convergence of ψ_n in the L_2-norm. Suppose that ψ_n' is another subsequence of $\{\psi_n\}$, converging weakly to ψ_0'. Then ψ_0' also solves problem (V) with bound c. By unicity, $\psi_c^* = \psi_0'$ and therefore the original sequence ψ_n converges weakly to the unique ψ_c^*. Hence, $\int \psi_n^2\, dP \leq \int (\psi_c^*)^2\, dP$ since $\psi_c^* \perp H_n$. It follows from (26) and (28) that $\int (\psi_c^*)^2\, dP \leq \lim_n \int \psi_n^2\, dP$. Therefore, $\lim_n \int \psi_n^2\, dP = \int (\psi_c^*)^2\, dP$. By Vitali's theorem (see, for example, Section A.5 of BKRW), $\int (\psi_n - \psi_0)^2\, dP \to 0$ and thus ψ_n converges to ψ_c^* in the L_2-norm, finishing the proof. □

Acknowledgments. This work is part of my Ph.D. dissertation under the supervision of Professor P. J. Bickel at the University of California, Berkeley. I am extremely grateful to his generous guidance and constant encouragement. I want to thank Professor L. R. Haff for correction of the language. I would also like to thank an Associate Editor and a referee for their detailed and helpful comments.

REFERENCES

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

BERAN, R. (1978). An efficient and robust adaptive estimator of location. *Ann. Statist.* **6** 292–313.

BICKEL, P. J. (1978). Using residuals robustly I: Tests for heteroscedasticity, nonlinearity. *Ann. Statist.* **6** 266–291.

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

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

CARROLL, R. J. and RUPPERT, D. (1982). Robust estimation in heteroscedastic linear models. *Ann. Statist.* **10** 429–441.

CHEN, S. (1990). Generalization of influence functions and their applications. Ph.D. dissertation, Univ. Rochester.

CHENG, C. L. and VAN NESS, J. W. (1992). Generalized M-estimators for errors-in-variables regression. *Ann. Statist.* **20** 385–397.

DUNFORD, N. and SCHWARTZ, J. (1966). *Linear Operators*. Interscience, New York.

FERNHOLZ, L. T. (1983). *von Mises' Calculus for Statistical Functionals*. Springer, New York.

HAMPEL, F. R. (1968). Contribution to the theory of robust estimation. Ph.D dissertation, Univ. California, Berkeley.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

JOBSON, J. D. and FULLER, W. A. (1980). Least squares estimation when the covariate matrix and parameter are functionally related. *J. Amer. Statist. Assoc.* **75** 176–181.

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

LINDSAY, B. G. (1980). Nuisance parameters, mixture models and the efficiency of partial likelihood estimators. *Philos. Trans. Roy. Soc. London Ser. A* **296** 639–665.

LUENBERGER, D. G. (1963). *Optimization by Vector Space Methods*. Wiley, New York.

NEYMAN, J. and SCOTT, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32.

SASIENI, P. (1993). Maximum weighted partial likelihood estimates for the Cox model. *J. Amer. Statist. Assoc.* **88** 144–152.

SHEN, L. Z. (1994). Explicit optimal B-robust influence functions in multidimensional parametric models. *Comm. Statist. Theory Methods* **23**. 1103–1122.

SHEN, X. and WONG, W. (1991). Convergence rate of sieves estimates. *Ann. Statist.* **22** 580–615.

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

VON MISES, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.* **18** 309–348.

WU, C. O. (1990). Asymptotically efficient robust estimates in some semiparametric models. Ph.D. Dissertation, Univ. California, Berkeley.

ZEIDLER, E. (1985). *Nonlinear Functional Analysis and Its Applications III: Variational Methods and Optimization* (L. F. Boron, transl.). Springer, New York.

BIOMETRICS AND STATISTICAL SCIENCES
THE PROCTER & GAMBLE COMPANY
11262 CORNELL PARK DRIVE
CINCINNATI, OHIO 45242