The Annals of Statistics
1982, Vol. 10, No. 2, 415–428

ROBUST ESTIMATION IN MODELS FOR INDEPENDENT NON-IDENTICALLY DISTRIBUTED DATA[1]

BY RUDOLF BERAN

University of California, Berkeley

This paper concerns robust estimation of the parameter θ which indexes a parametric model for independent non-identically distributed data. For reasonable choices of contamination neighborhood and of what is to be estimated when the parametric model does not hold, we characterize asymptotically minimax robust estimates of θ. When applied to the normal regression model, the theory yields recipes for the influence curves of optimal robust regression and scale estimates. The contamination neighborhood does not assume regression plus error structure, the regression and scale parameters are estimated simultaneously, and the theory establishes ròles for estimates with redescending influence curves as well as for those with monotone influence curves. When applied to the logit and probit models, the theory recommends influence curves which differ markedly from those of the maximum likelihood estimates except in the i.i.d. case.

1. Introduction. A major advance in statistical analysis has been the development of robust procedures for fitting linear regression models. Despite considerable interest in this development (review papers include Huber, 1973; Hogg, 1974; Bickel, 1976), the available theory for robust regression remains unsatisfactory in the following important respects:

(i) The contamination neighborhoods considered are too restricted. Regression plus error structure is assumed, complete knowledge of the regressors is often assumed, and symmetry (or a requirement nearly as strong) is often imposed upon the possible error distributions.

(ii) Unresolved by available theory is the role of regression estimates with redescending influence curves versus those with monotone influence curves.

(iii) Theoretical results concerning robust estimation of the scale parameter are less complete than those regarding the regression parameters.

(iv) Available robust regression theory does not extend very helpfully to exponential linear models or to other models for independent non-identically distributed (i.n.i.d.) observations.

It has become increasingly evident that the basic questions for any robustness study are:

(a) What are rich, technically workable, contamination neighborhoods about the postulated model?

(b) What is being estimated (or tested) when the postulated model holds and when it does not hold?

(c) What are the criteria for assessing performance of a robust procedure?

The importance of questions (a) and (c) was recognized early in robustness investigations. Question (b) has come to prominence more gradually. It has frequently been replaced by the simpler question (b′) What procedures are being considered? Cf. Bickel's (1976) review and the literature on M, L, or R estimates. Reluctance to tackle question (b) head-on has contributed, no doubt, to the over-emphasis on symmetric error distributions in

Received October 1980; revised October 1981.

[1] Supported in part by the National Science Foundation Grant MCS-75-10376.

AMS 1980 *subject classifications.* Primary 62F35; secondary 62C20.

Key words and phrases. Robust estimates, independent non-identically distributed, parametric model, robust regression, logit model, asymptotic minimax.

robust regression and to the vague dissatisfaction with robust methods which is felt by some statisticians.

Two very different approaches to question (b) have been attempted:

Idealism. There exists a true value of θ which is to be estimated. Whether or not a postulated parametric model involving θ fits the experimental data does not affect this goal (Bickel, 1976, page 147; Rieder, 1980).

Empiricism. What is to be estimated is a particular functional of the distribution of the data; the value of the functional at the postulated parametric model is the parameter θ. Question (b) becomes: which functional? See Hampel (1974), Holm (1976), Beran (1981a), Millar (1981), and, for a related variant, Bickel and Lehmann (1975).

Although both points of view can be criticized, the second reflects the appealing idea that basic concepts in science should be defined operationally.

Robust estimation in general parametric models for i.n.i.d. data is the theme of this paper. For such models, the paper will propose specific reasonable answers to questions (a), (b), (c), and will then develop the associated robustness theory. Optimal robust estimates will be characterized essentially through their influence curves. It is one of the merits of the principles adopted here that they yield quantitative answers to the problem of robust estimation in general parametric models for i.n.i.d. data.

Suppose that the parametric model for the sample $\{X_i : 1 \leq i \leq n\}$ asserts that the distribution of X_i is $P_{\theta,i}$, where $\theta \in \Theta \subset R^k$, and that the joint distribution of the sample is $P_\theta^n = P_{\theta,1} \times P_{\theta,2} \times \cdots \times P_{\theta,n}$. Suppose that the actual distribution of the data is $Q^n = Q_1 \times Q_2 \times \cdots \times Q_n$. Technicalities aside, we might define the aim of robust estimation of θ to be the estimation of that value $t \in \Theta$ which minimizes $\sum_{i=1}^n \| P_{t,i} - Q_i \|^2$, where $\| \cdot \|$ is distance on probabilities. This operational minimum distance answer to question (b) rests on the idea that, in fitting the parametric model P_θ^n to the sample, we seek to estimate the actual distribution Q^n by a member of the parametric family. For i.i.d. models, various forms of the idea have been considered by Holm (1976), Beran (1977, 1981), Millar (1981), and Parr and Schucany (1980).

Certain technical difficulties arise with a minimum distance definition of what is to be estimated. Does the object in question exist? Is it unique? What does it look like? A modified definition which partly avoids these problems will be given in Section 2.1. Section 2.2 will describe the contamination neighborhoods used in this paper. Salient features of these neighborhoods are their richness and their local character. For example, the contamination neighborhood about a standard linear model does not only contain distributions having regression plus error structure. However, the "diameter" of the contamination neighborhood is limited.

Good local asymptotic performance is essential in robust estimation because local departures from the parametric model cannot be detected reliably by any test. While local asymptotic robustness need not ensure global robustness (cf. Beran, 1981), it often does. The examples in Section 4 illustrate this phenomenon. For the normal linear model, our local optimality theory recommends (as natural special cases) two influence curves whose empirical value has been demonstrated by recent Monte Carlo studies.

A less justifiable aspect of the contamination neighborhoods is the assumption that the observations are independent. Weakening this would certainly be desirable.

Performance of an estimate of θ will be measured by its minimax risk over contamination neighborhoods of the parametric model. The pessimism inherent in the minimax approach is desirable here because the contamination neighborhoods are local; no test can distinguish between distributions in these neighborhoods with asymptotic power one. Asymptotically minimax estimates, which are the goal of the theory, will be characterized in Section 2.3.

In the past, robustness studies have sometimes emphasized the efficiency of a robust estimate at the ideal parametric model. However, efficiency at the parametric model is logically important only if one believes that the parametric model holds *strictly* for most samples observed. Our assumption is weaker: that the actual distribution of the sample is near a member of the parametric family but is not otherwise specified. It is this realistic lack of knowledge which makes maximum risk over neighborhoods of the parametric model the preferred measure of performance.

Assumptions and the main robust estimation results are presented in Section 2. Section 3 pursues the implications for robust estimation in the normal regression model, both linear and nonlinear. The robust estimation of scale is linked to the estimation of the regression parameters, the roles of monotone and redescending influence curves are clarified, and certain influence curves are recommended on asymptotic minimax grounds. Also treated in Section 3 is robust estimation in the logit and probit models. Interestingly, the theory recommends influence curves which differ, in a conservative direction, from those of the maximum likelihood estimates. Section 4 defines and studies Hellinger differentiable functionals of product measures, then draws on this material to prove the theorems in Section 2.

It has been suggested by some that robust procedures should adapt themselves to the sample, treating "good" samples less cautiously than clearly "contaminated" samples; cf. Hogg's (1974) review paper. The robust estimates developed in this paper do not have this property. For the i.i.d. case, it is known (Beran, 1981a, 1981b) that suitably constructed adaptive robust estimates and tests are, in fact, asymptotically minimax over contamination neighborhoods smaller than those considered in this paper. We expect that an analogous result holds for general i.n.i.d. models. Considerable care is needed in constructing adaptive robust procedures, precisely because local departures from the parametric model cannot be reliably detected by any test.

2. Main results.

2.1. *What is being estimated*? Let $\mathscr{X}$ be a finite dimensional Euclidean space with Borel sets $\mathscr{A}$. Suppose that the parametric model for the sample $\{X_i : 1 \le i \le n\}$ postulates that the distribution of X_i is $P_{\theta,i}$ and that the $\{X_i\}$ are independent. Each $P_{\theta,i}$ is a probability on $(\mathscr{X}, \mathscr{A})$ and $\theta \in \Theta$, an open subset of R^k. Suppose that the actual distribution of the sample is $Q^n = Q_1 \times Q_2 \times \cdots \times Q_n$, each factor being a probability on $(\mathscr{X}, \mathscr{A})$.

Choose a, b in $\mathscr{X}$ such that $a < b$ (component-wise) and set $w(x) = 1$ if $a \le x \le b$, $w(x) = 0$ otherwise. Let $F_{\theta,i}$, G_i be the cdf's of $P_{\theta,i}$, Q_i respectively and let

$$(2.1) \qquad f_{\theta,i}(x) = \int w(x-t)\, dF_{\theta,i}(t), \qquad g_i(x) = \int w(x-t)\, dG_i(t).$$

The distance between $P_{\theta,i}$ and Q_i will be defined by

$$(2.2) \qquad \| Q_i - P_{\theta,i} \|_i = \left[\int \{g_i(x) - f_{\theta,i}(x)\}^2 \, d\mu_{\theta,\iota} \right]^{1/2},$$

where $\mu_{\theta,i}$ is a probability on $(\mathscr{X}, \mathscr{A})$ which is allowed to vary with θ and i.

This distance compares probabilities assigned to rectangles by Q_i with the corresponding probabilities assigned by $P_{\theta,i}$. When $a = 0$ and $b = \infty$, $\| Q_i - P_{\theta,i} \|_i$ is the $L^2(\mu_{\theta,i})$ distance between the cdf's G_i, $F_{\theta,i}$. Millar (1981) has made extensive use of this distance between cdf's in an analysis of robust estimation for i.i.d. parametric models. When $a = -\varepsilon$ and $b = \varepsilon$, with ε small and positive, $\| Q_i - P_{\theta,i} \|_i$ is effectively the $L^2(\mu_{\theta,i})$ distance between the Lebesgue densities, should they exist in smooth versions. Whatever the choice of $a < b$, the ball $\{Q: \| Q - P_{\theta,i} \|_i \le c\}$ contains the Kolmogorov-Smirnov ball $\{G: \sup_x | G(x) - F_{\theta,i}(x) | \le c/2\}$.

As was suggested in the Introduction, we might reasonably define the aim of robust estimation of θ to be the estimation of that value $t \in \Theta$ which minimizes $\sum_{i=1}^n \| Q_i - P_{t,i} \|_i^2$. The values of a, b, $\mu_{\theta,i}$ determine the metric $\| \cdot \|_i$ and hence, the minimum distance functional of Q^n to be estimated when the actual distribution Q^n of the sample does not belong to the parametric model $\{P_\theta^n : \theta \in \Theta\}$. Choosing a, b, $\mu_{\theta,i}$ requires the statistician to clarify her or his goals in fitting an approximately correct parametric model to data. For instance, is the fitted parametric distribution to best approximate the center, one tail, or both tails of the actual distribution? Are the two distributions to be compared through their cdf's or densities? For a more concrete discussion of these issues, see the linear model example of Section 3.1.

Unfortunately, minimizing values of t need not exist or be unique. One way out of the theoretical difficulty is to adopt *local* definitions of what is to be estimated: for every $\theta \in \Theta$, we give a definition, valid for all distributions in a small neighborhood of P_θ^n. The existence of a *global* definition matching the local definitions in each of the small neighborhoods becomes a separate issue.

The following smoothness assumption will be made on the parametric model P_θ^n.

ASSUMPTION A1. There exist column vector functions $\{\gamma_{\theta,i} \in L_2^k(\mu_{\theta,i})\}$ such that, for every $\theta \in \Theta$, the matrices

$$V_n(\theta) = \textstyle\sum_{i=1}^n \displaystyle\int \gamma_{\theta,i}(x)\gamma'_{\theta,i}(x)\, d\mu_{\theta,i}, \qquad n \geq 1 \tag{2.3}$$

are non-singular and

$$\lim_{n\to\infty} \textstyle\sum_{i=1}^n \displaystyle\int \{f_{\theta+V_n^{-1}(\theta)h,i}(x) - f_{\theta,i}(x) - V_n^{-1}(\theta)h'\gamma_{\theta,i}(x)\}^2\, d\mu_{\theta,i} = 0 \tag{2.4}$$

for every column vector $h \in R^k$.

Fix $\theta \in \Theta$ and suppose that $\sum_{i=1}^n \| Q_i - P_{\theta,i}\|_i^2$ is small, Q^n being the actual distribution of the sample. Keeping the linearization (2.4) in mind, define the local functional to be estimated as

$$T_n(\theta, Q^n) = \theta + V_n^{-1}(\theta)h_0, \tag{2.5}$$

where h_0 is the value of $h \in R^k$ which minimizes the quantity

$$\textstyle\sum_{i=1}^n \displaystyle\int \{g_i(x) - f_{\theta,i}(x) - V_n^{-1}(\theta)h'\gamma_{\theta,i}(x)\}^2\, d\mu_{\theta,i}. \tag{2.6}$$

Here $f_{\theta,i} + V_n^{-1}(\theta)h'\gamma_{\theta,i}$ serves as a local approximation to $f_{\theta+V_n^{-1}(\theta)h,i}$, and $T_n(\theta, Q^n)$ is simply a local version of the minimum distance functional proposed earlier and shares the same intuitive statistical rationale as that functional. It is easily checked that

$$T_n(\theta, Q^n) = \theta + V_n^{-1}(\theta) \textstyle\sum_{i=1}^n \displaystyle\int \gamma_{\theta,i}(x)[g_i(x) - f_{\theta,i}(x)]\, d\mu_{\theta,i}. \tag{2.7}$$

Alternatively, if

$$\delta_{\theta,i}(t) = \int w(x-t)\gamma_{\theta,i}(x)\, d\mu_{\theta,i}, \qquad \rho_{\theta,i}(t) = \delta_{\theta,i}(t) - \int \delta_{\theta,i}(x)\, dP_{\theta,i}, \tag{2.8}$$

then

$$T_n(\theta, Q^n) = \theta + V_n^{-1}(\theta) \textstyle\sum_{i=1}^n \displaystyle\int \rho_{\theta,i}(x)\, dG_i. \tag{2.9}$$

As expected, $T_n(\theta, P_\theta^n) = \theta$ for every $\theta \in \Theta$. It is not immediately clear, of course, that the local functionals $\{T_n(\theta, Q^n) : \theta \in \Theta\}$ can be matched with a global functional $T_n(Q^n)$ which, for every $\theta \in \Theta$, approximates $T_n(\theta, Q^n)$ whenever Q^n is close to P_θ^n. We will not tackle this question here. Instead, we will seek global estimates $\hat{T}_n$ (depending on the sample but not on θ) which are asymptotically minimax estimates of $T_n(\theta, Q^n)$ in a neighborhood of P_θ^n, whatever the choice of $\theta \in \Theta$.

2.2. *The contamination neighborhoods and the risk.* The performance of any estimate $\hat{T}_n$ will be examined for every distribution Q^n in the contamination neighborhoods

$\{B_n(\theta, c) : \theta \in \Theta, c > 0\}$, where

$$(2.10)\qquad B_n(\theta, c) = \{Q^n \colon \textstyle\sum_{i=1}^n \displaystyle\int [G_i(x) - F_{\theta,i}(x)]^2 \, d\mu_{\theta,i}(x + h) \le c^2, h = a, h = b\}.$$

The constants a, b are those appearing in the definition of w. Chosen in part for its mathematical tractability, $B_n(\theta, c)$ contains a rich variety of distributions Q^n near P_θ^n. Whatever the choice of $a < b$, $B_n(\theta, c)$ contains $\{Q^n \colon \sum_{i=1}^n \sup_x |G_i(x) - F_{\theta,i}(x)|^2 \le c^2\}$. In particular, $B_n(\theta, c)$ allows mixture contamination of the form

$$(2.11)\qquad G_i(x) = (1 - \varepsilon_i) F_{\theta,i}(x) + \varepsilon_i H_i(x),$$

where $\{H_i \colon 1 \le i \le n\}$ are arbitrary cdf's on $\mathscr{X}$ and $\sum_{i=1}^n \varepsilon_i^2 \le c^2$.

Let

$$(2.12)\qquad C_n(\theta) = \textstyle\sum_{i=1}^n \displaystyle\int \rho_{\theta,i}(x)\rho'_{\theta,i}(x) \, dP_{\theta,i},$$

and assume the following.

Assumption A2. For every $\theta \in \Theta$, the matrices $\{C_n(\theta) \colon n \ge 1\}$ are non-singular and

$$(2.13)\qquad \lim_{n\to\infty} \sup_{1 \le i \le n} \sup_x |C_n^{-1/2}(\theta) \rho_{\theta,i}(x)| = 0.$$

Let $A_n(\theta) = C_n^{-1/2}(\theta) V_n(\theta)$. We define the risk of any estimate $\hat{T}_n$ to be

$$(2.14)\qquad R_n(\hat{T}_n, Q^n) = E_{Q^n} u[|A_n(\theta)\{\hat{T}_n - T_n(\theta, Q^n)\}|^2],$$

where u is any monotone increasing bounded function mapping $R^+ \to R^+$ and $|\cdot|$ is any metric on R^k. The matrix $A_n'(\theta)A_n(\theta)$ plays the role of a generalized information matrix. The performance of $\hat{T}_n$ will be measured by $\sup_{Q^n \in B_n(\theta,c)} R_n(\hat{T}_n, Q^n)$, calculated for every $\theta \in \Theta$. The underlying idea is that $\hat{T}_n$ should estimate $T_n(\theta, Q^n)$ wherever Q^n is near P_θ^n, whatever the choice of $\theta \in \Theta$.

Remarks. In defining the risk, $\{A_n(\theta)\}$ might be replaced by other sequences of matrices $\{D_n\}$. Under some assumptions on $\{D_n\}$, a more complicated version of Theorem 1 (Section 2.3) remains valid and the identification of asymptotically minimax estimates in Theorem 2 is unaffected.

The contamination described by $B_n(\theta, c)$ is asymptotically local. For fixed $c > 0$, increasing n forces most of the cdf's G_i to approach the corresponding $F_{\theta,i}$. Important for the asymptotics, this feature of the contamination model is tolerable because c can be taken arbitrarily large and because even a small amount of contamination can be dangerous (e.g., one sufficiently extreme outlier in a million observations destroys the usefulness of the sample mean). See also the discussion in the Introduction.

2.3. *Asymptotically minimax estimates.* Let ϕ_k be the standard k-dimensional normal density and let

$$(2.15)\qquad r_0(u) = \int u(|z|^2)\phi_k(z) \, dz.$$

The following result, a consequence of the Hájek-LeCam asymptotic minimax theorem, indicates how well we may robustly estimate $T_n(\theta, Q^n)$. The infimum on the left side of (2.16) is taken over all possible estimates $\hat{T}_n$.

Theorem 1. *Suppose Assumptions* A1 *and* A2 *are satisfied. Then, for every* $\theta \in \Theta$,

$$(2.16)\qquad \lim_{c\to\infty} \lim\inf_n \inf_{\hat{T}_n} \sup_{Q^n \in B_n(\theta,c)} R_n(\hat{T}_n, Q^n) \ge r_0(u).$$

Attainability of this lower bound is not immediately evident. Under the following additional assumption, sufficient conditions can be given for an estimator sequence $\{\hat{T}_n : n \geq 1\}$ to be asymptotically minimax.

ASSUMPTION A3. The vector functions $\{\gamma_{\theta,i}\}$ are such that, for every $\theta \in \Theta$,

$$\lim_{n\to\infty}\sup_{1\leq i\leq n} C_n^{-1/2}(\theta) \int \gamma_{\theta,i}(x)\gamma'_{\theta,i}(x)\, d\mu_{\theta,i} C_n^{-1/2}(\theta) = 0 \tag{2.17}$$

and

$$\sup_n C_n^{-1/2}(\theta)\, V_n(\theta)\, C_n^{-1/2}(\theta) < \infty. \tag{2.18}$$

THEOREM 2. *Suppose Assumptions* A1, A2, A3 *are satisfied. Let* $\{\hat{T}_n : n \geq 1\}$ *be any sequence of estimates which has the property that, for every* $\theta \in \Theta$ *and every* $c > 0$,

$$V_n^{1/2}(\theta)(\hat{T}_n - \theta) - V_n^{-1/2}(\theta) \textstyle\sum_{i=1}^n \rho_{\theta,i}(X_i) \rightarrow_{Q_n^n} 0. \tag{2.19}$$

under every sequence of product measures $\{Q_n^n \in B_n(\theta, c) : n \geq 1\}$. *Then*

$$\lim_{n\to\infty}\sup_{Q^n \in B_n(\theta,c)} R_n(\hat{T}_n, Q^n) = r_0(u), \tag{2.20}$$

for every $\theta \in \Theta$ *and every* $c > 0$.

Of course, (2.20) implies that equality holds in (2.16) and that estimates satisfying (2.19) are asymptotically minimax robust estimates. From the expression (2.9) for $T_n(\theta, Q^n)$, it is evident that (2.19) is equivalent to requiring

$$V_n^{1/2}(\theta)\{\hat{T}_n - T_n(\theta, Q^n)\} - V_n^{-1/2}(\theta) \textstyle\sum_{i=1}^n \{\rho_{\theta,i}(X_i) - \displaystyle\int \rho_{\theta,i}(t)\, dG_{n,i}(t)\} \rightarrow_{Q_n^n} 0 \tag{2.21}$$

for every sequence $\{Q_n^n \in B_n(\theta, c) : n \geq 1\}$. Here $Q_n^n = Q_{n,1} \times Q_{n,2} \times \cdots \times Q_{n,n}$ and $G_{n,i}$ is the cdf of $Q_{n,i}$.

A plausible one-step construction of estimates $\{\hat{T}_n\}$ that satisfy (2.19) runs as follows. First, find initial estimates $\{\hat{\theta}_n\}$ such that $\{V_n^{1/2}(\theta)(\hat{\theta}_n - \theta) : n \geq 1\}$ is tight under every sequence $\{Q_n^n \in B_n(\theta, c) : n \geq 1\}$, whatever the choice of $\theta \in \Theta$ and $c > 0$. In other words, $\{\hat{\theta}_n\}$ should have the right rate of convergence to θ and should not misbehave over the contamination neighborhoods; the tightness requirement is an expression of qualitative robustness under local contamination. Then, define

$$\hat{T}_n = \hat{\theta}_n + V_n^{-1}(\hat{\theta}_n) \textstyle\sum_{i=1}^n \rho_i(X_i, \hat{\theta}_n), \tag{2.22}$$

where $\rho_i(x, \theta)$ is another notation for $\rho_{\theta,i}(x)$.

To see why this construction should work, under regularity conditions, observe that

$$V_n^{1/2}(\theta)(\hat{T}_n - \theta) = V_n^{1/2}(\theta)(\hat{\theta}_n - \theta) + V_n^{1/2}(\theta)\, V_n^{-1}(\hat{\theta}_n)\, V_n^{1/2}(\theta) \textstyle\sum_{j=1}^3 A_{j,n}, \tag{2.23}$$

where

$$\begin{aligned} A_{1,n} &= V^{-1/2}(\theta) \textstyle\sum_{i=1}^n \{\rho_i(X_i, \hat{\theta}_n) - \displaystyle\int \rho(x, \hat{\theta}_n)\, dG_{n,i}\} \\ A_{2,n} &= V_n^{-1/2}(\theta) \textstyle\sum_{i=1}^n \displaystyle\int \rho_i(x, \hat{\theta}_n)\, d(G_{n,i} - F_{\theta,i}) \\ A_{3,n} &= V_n^{-1/2}(\theta) \textstyle\sum_{i=1}^n \displaystyle\int \rho_i(x, \hat{\theta}_n)\, d(F_{\theta,i} - F_{\hat{\theta}_n,i}). \end{aligned} \tag{2.24}$$

We expect that, under every sequence $\{Q_n^n \in B_n(\theta, c)\}$,

$$A_{1,n} = V_n^{-1/2}(\theta) \textstyle\sum_{i=1}^n \{\rho_i(X_i, \theta) - \displaystyle\int \rho(x, \theta)\, dG_{n,i}\} + o_p(1)$$

$$(2.25) \qquad A_{2,n} = V_n^{-1/2}(\theta) \textstyle\sum_{i=1}^n \displaystyle\int \rho(x, \theta)\, dG_{n,i} + o_p(1)$$

$$A_{3,n} = V_n^{1/2}(\theta)(\hat{\theta}_n - \theta) + o_p(1)$$

and $V_n(\theta) V_n^{-1}(\hat{\theta}_n)$ converges in probability to the identity matrix. From this, (2.19) would follow.

The argument can be made rigorous by assuming that (a) the $\{\gamma_{\theta,i}\}$ and $\{\mu_{\theta,i}\}$ are sufficiently smooth as functions of θ; (b) the $\{\hat{\theta}_n : n \geq 1\}$ are discretized estimates, constructed as follows. Suppose the $\{\theta_n^* : n \geq 1\}$ are estimates such that $\{V_n^{1/2}(\theta)(\theta_n^* - \theta)\}$ is tight under every sequence $\{Q_n^n \in B_n(\theta, c)\}$, whatever the choice of $\theta \in \Theta$ and $c > 0$. Pave Θ with parallelepipeds centered at the points $\{V^{-1/2}(\theta_n^*)h : h$ a $k \times 1$ vector with integer components$\}$. Set $\hat{\theta}_n$ equal to the center of the parallelepiped which contains θ_n^*.

A related argument appears in Beran (1981a); also noted in that paper are the similarities with LeCam's classical study of one-step maximum likelihood estimates.

3. Examples.

This section describes implications of the preceding theory for robust estimation in the normal regression model, both linear and nonlinear, and in the logit and probit models. The aim is to illustrate the fruitfulness of the general approach which we have followed.

3.1. *Normal linear model.* The parametric model specifies that the X_i are independent $N(\sum_{j=1}^r c_{ij}\beta_j, \sigma^2)$, where $r < n$. Let $\beta = (\beta_1, \beta_2, \cdots, \beta_r)'$, let C_n be the $n \times r$ matrix with components $\{c_{ij} : 1 \leq i \leq n,\ 1 \leq j \leq r\}$ and assume that rank $(C_n) = r$. Let $G_n = C_n(C_n'C_n)^{-1}C_n'$ and write $G_n = \{g_{ij,n}\}$. The vector parameter to be estimated robustly is $\theta = (\beta_1, \beta_2, \cdots, \beta_r, \sigma^2)'$.

For this example, we suppose that

$$(3.1) \qquad d\mu_{\theta,i}(x) = d\lambda(\sigma^{-1}(x - c_i'\beta)),$$

where c_i' is the ith row of C_n and λ is a probability on the real line, symmetric about zero. We further assume that the kernel w is one of two possible functions: either $a = 0$ and $b = \infty$, or $a = -\varepsilon$ and $b = \varepsilon$ with ε positive and small. In the first case, w is a translated odd function; in the second instance w is even. The key assumption on the matrices $\{G_n : n \geq 1\}$ is

$$(3.2) \qquad \lim_{n\to\infty} \max_{1\leq i\leq n} g_{ij,n} = 0.$$

Under these conditions, Assumptions A1, A2, A3 of Section 2 are satisfied. The importance of (3.2) in the asymptotics of ordinary least squares estimates has been pointed out by Huber (1973).

Let $r_{\theta,i}(x) = \sigma^{-1}(x - c_i'\beta)$. For the normal linear model under consideration

$$(3.3) \qquad \gamma_{\theta,i}(x) = \sigma^{-1}\begin{pmatrix} c_i\gamma_1(r_{\theta,i}(x)) \\ \gamma_2(r_{\theta,i}(x)) \end{pmatrix},$$

with

$$(3.4) \qquad \begin{aligned} \gamma_1(x) &= \int (x - u)\phi(x - u) w(\sigma u)\, du, \\ \gamma_2(x) &= \int \{-1 + (x - u)^2\}\phi(x - u) w(\sigma u)\, du, \end{aligned}$$

ϕ being the $N(0, 1)$ density. Hence

$$V_n(\theta) = \sigma^{-2}\begin{pmatrix} C_n'C_n \int \gamma_1^2(x)\, d\lambda & 0 \\ 0 & n \int \gamma_2^2(x)\, d\lambda \end{pmatrix},$$

the off-diagonal elements vanishing because of the assumptions on λ and w. For $i = 1, 2$ let

$$\delta_i(x) = \int \gamma_i(x + u) w(\sigma u)\, d\lambda(x + u)$$

and let $\rho_i(x) = \delta_i(x) - \int \delta_i(t)\phi(t)\, dt$. Then

$$\rho_{\theta,i}(x) = \sigma^{-1} \begin{pmatrix} c_i\rho_1(r_{\theta,i}(x)) \\ \rho_2(r_{\theta,i}(x)) \end{pmatrix}.$$

According to Theorem 2, a sequence of estimates $\{(\hat{\beta}_n, \hat{\sigma}_n)\}$ is asymptotically minimax if, under every sequence $\{Q_n^n \in B_n(\theta, c)\}$,

$$\begin{aligned} (C_n'C_n)^{1/2}(\hat{\beta}_n - \beta) &= \sigma\left\{ \int \gamma_1^2(x)\, d\lambda \right\}^{-1} (C_n'C_n)^{-1/2} \textstyle\sum_{i=1}^n c_i\rho_1(r_{\theta,i}(X_i)) + o_p(1) \\ n^{1/2}(\hat{\sigma}_n - \sigma) &= \sigma\left\{ \int \gamma_2^2(x)\, d\lambda \right\}^{-1} n^{-1/2} \textstyle\sum_{i=1}^n \rho_2(r_{\theta,i}(X_i)) + o_p(1) \end{aligned} \tag{3.5}$$

Special cases. (i) When $a = 0$ and $b = \infty$, (3.4) simplifies to $\gamma_1(x) = -\phi(x)$ and $\gamma_2(x) = -x\phi(x)$. If λ is effectively Lebesgue measure (i.e., λ is uniform on a symmetric interval large enough to contain any actual observation), then, to an adequate approximation,

$$\rho_1(x) = \Phi(x) - 2^{-1}, \qquad \rho_2(x) = 2^{-1}\pi^{-1/2} - \phi(x), \tag{3.6}$$

Φ being the $N(0, 1)$ cdf, and

$$\int \gamma_1^2(x)\, d\lambda = 2^{-1}\pi^{-1/2}, \qquad \gamma_2^2(x)\, d\lambda = 4^{-1}\pi^{-1/2}.$$

The function $\rho_1(x)$ is strictly monotone increasing in x, while $\rho_2(x)$ is strictly monotone in $|x|$. The i.i.d. sub-case of this example was treated by Millar (1981), who also examined the effects of varying λ and of replacing the normal by other distributions. Parr and Schucany (1980) report favorable Monte Carlo results for minimum distance location estimates having the influence curve ρ_1 defined in (3.6).

(ii) When $a = -\varepsilon$ and $b = \varepsilon$ with ε every small and positive, and λ has continuous Lebesgue density λ', reasonable approximations are $\gamma_1(x) = x\phi(x)$, $\gamma_2(x) = (x^2 - 1)\phi(x)$ and $\rho_1(x) = x\phi(x)\lambda'(x)$, $\rho_2(x) = (x^2 - 1)\phi(x)\lambda'(x) - \int(t^2 - 1)\phi^2(t)\lambda'(t)\, dt$. If λ is effectively Lebesgue measure, then

$$\rho_1(x) = x\phi(x), \qquad \rho_2(x) = (x^2 - 1)\phi(x) + 4^{-1}\pi^{-1/2}$$

$$\int \gamma_1^2(x)\, d\lambda = 4^{-1}\pi^{-1/2}, \qquad \int \gamma_2^2(x)\, d\lambda = 3(8\pi^{1/2})^{-1}.$$

Both $\rho_1(x)$ and $\rho_2(x)$ redescend to zero as $|x|$ increases. The score function ρ_1 has previously been considered for robust regression by Holland and Welsch (1977), who note its fine empirical performance in a Monte Carlo study.

Of course, different functionals are being estimated in (i) and (ii). When $a = 0$ and $b = \infty$, the distance $\|\cdot\|_i$ is more sensitive to discrepancies in the tails of distributions than when $a = -\varepsilon$ and $b = \varepsilon$. A small probability mass that is moved from one tail to the other affects cdf comparisons more than density comparisons, in a given L^2-norm. Since the functional being estimated is, roughly speaking, the value of $t \in \Theta$ minimizing $\sum_{i=1}^n \|P_{t,i} - Q_i\|_i^2$, it is not surprising that the estimates $(\hat{\beta}_n, \hat{\sigma}_n)$ in case (ii) should discount the tails of the sample more than in case (i). In the robustness framework of this paper, both estimates can be optimal.

The asymptotics used here rest on (3.2), which limits possible behavior of the regression matrices $\{C_n : n \geq 1\}$. Under (3.2), no single observation would be given disproportionate weight in the calculation of the least squares estimate of β. A striking consequence of assuming (3.2) is the appearance in (3.5) of C_n itself, rather than some robust modification of C_n. Even though most probabilities Q^n in the contamination neighborhood $B_n(\theta, c)$ do not have strict linear model structure, the asymptotically minimax estimate $\hat{\beta}_n$ acts as if they did.

Global robust estimates $(\hat{\beta}_n, \hat{\sigma}_n^2)$ satisfying (3.5) for every sequence $\{Q_n^n \in B_n(\theta, c)\}$, whatever the choice of $\theta \in \Theta$ and $c > 0$, can be obtained by the one-step construction (2.22). M-estimates with smooth, monotone, bounded score functions can serve as suitable initial estimates.

The size of the contamination neighborhood $B_n(\theta, c)$ depends considerably upon the choice of λ, as does the functional being estimated. Extreme cases include λ atomic at the origin and λ effectively Lebesgue.

The example extends easily to linear regression models with non-normal error density f. Only the definitions of ρ_i and γ_i change in (3.5).

3.2. *Normal non-linear regression.* The previous example can be generalized to non-linear regression models which are locally linear. Suppose that the X_i are independent and that under the parametric model, the distribution of X_i is $N(g_i(\beta),\sigma^2)$, where $g_i(\beta)$ is differentiable in β. The parameter to be estimated is again $\theta = (\beta_1, \beta_2, \cdots, \beta_r, \sigma^2)$.

Make the same assumptions regarding $\mu_{\theta,i}$ and w as in Section 3.1. Let $r_{\theta,i}(x) = \sigma^{-1}(x - g_i(\beta))$ and let $c_i(\beta)$ be the $r \times 1$ vector whose jth component is $\partial g_i(\beta)/\partial\beta_j$. Let $C_n(\beta)$ be the $n \times r$ matrix whose ith row is $c_i'(\beta)$. With $C_n(\beta)$, $c_i(\beta)$ in place of C_n, c_i respectively, equation (3.5) characterizes asymptotically minimax estimates of β and σ^2. However, checking the assumptions A1 to A3 and devising suitable initial estimates is typically much more difficult in nonlinear regression models; cf. the classical treatment of i.n.i.d. parametric models by Ibragimov and Khasminskii (1975).

3.3. *Logit model.* Each observation X_i is necessarily either 0 or 1. The parametric model asserts that the X_i are independent Binomial $(1, \pi_i(\theta))$, where $\text{logit}\{\pi_i(\theta)\} = \sum_{j=1}^{r} c_{ij}\theta_j$. Let $\theta = (\theta_1, \theta_2, \cdots, \theta_r)'$ and define C_n as in Section 3.1, assuming rank $(C_n) = r$. A reasonable choice for $\mu_{\theta,i}$ is the probability which assigns mass ½ to $x = 0$ and mass ½ to $x = 1$.

Setting $a = 0$ and $b = \infty$ in defining w yields

$$\gamma_{\theta,i}(x) = \begin{cases} -c_i\pi_i(\theta)\{1 - \pi_i(\theta)\} & \text{if } \ x = 0, \\ \qquad 0 & \text{if } \ x = 1 \end{cases}$$

c_i' being the ith row of C_n. Thus, $V_n(\theta) = 2^{-1}C_n'D_n^2(\theta)C_n$, where $D_n(\theta) = \text{diag}[\pi_i(\theta)\{1 - \pi_i(\theta)\}]$, and

$$\rho_{\theta,i}(x) = \begin{cases} -2^{-1}c_i\pi_i^2(\theta)\{1 - \pi_i(\theta)\} & \text{if } \ x = 0, \\ 2^{-1}c_i\pi_i(\theta)\{1 - \pi_i(\theta)\}^2 & \text{if } \ x = 1. \end{cases}$$

Checking Assumptions A1 to A3 is cumbersome. The matrix $C_n(\theta)$ defined in (2.12) here becomes $C_n(\theta) = 4^{-1}C_n'D_n^3(\theta)C_n$.

According to Theorem 2, a sequence of estimates $\{\hat{T}_n\}$ is asymptotically minimax in this model if, under every sequence $\{Q_n^n \in B_n(\theta, c)\}$,

$$\{C_n'D_n^2(\theta)C_n\}^{1/2}(\hat{T}_n - \theta) = \{C_n'D_n^2(\theta)C_n\}^{-1/2}C_n'D_n(\theta)y_n(\theta) + o_p(1), \tag{3.7}$$

where $y_n(\theta)$ is the $n \times 1$ vector whose ith component is $X_i - \pi_i(\theta)$. Thus, the one-step construction (2.22) of $\hat{T}_n$ here becomes

$$\hat{T}_n = \hat{\theta}_n + \{C_n'D_n^2(\hat{\theta}_n)C_n\}^{-1}C_n'D_n(\hat{\theta}_n)y_n(\hat{\theta}_n), \tag{3.8}$$

with $\hat{\theta}_n$ a suitable initial estimate.

On the other hand, a one-step version of the classical maximum likelihood estimate for θ in the logit model would be

$$T_n^* = \hat{\theta}_n + \{C_n'D_n(\hat{\theta}_n)C_n\}^{-1}C_n'y_n(\hat{\theta}_n).$$

The estimate $\hat{T}_n$ puts relatively less weight on observations associated with very large or very small values of $\pi_i(\theta)$ than does the classical estimate. When the X_i are i.i.d. (i.e., $r = 1$ and C_n is the $n \times 1$ vector of ones), both $\hat{T}_n$ and T_n^* reduce to the same estimate.

The alternate choice $a = -\varepsilon$ and $b = \varepsilon (0 < \varepsilon < 2^{-1})$ in defining w yields the same functional $T_n(\theta, Q^n)$ as did $a = 0$ and $b = \infty$, and so yields the same recipe (3.7) for an asymptotically minimax estimate. The reason for coincidence is simple: apart from a factor of $2^{1/2}$, the distance $\| \cdot \|_i$ is the same for both choices of (a, b).

3.4. *Probit model.* The probit model differs from the logit model in only one respect: $\pi_i(\theta) = \Phi(\sum_{j=1}^r c_{ij}\theta_j)$ where Φ is the standard normal cdf. For the choices of a, b, $\mu_{\theta,i}$ already described in Section 3.3, the one-step construction (2.22) of an asymptotically minimax robust estimate $\hat{T}_n$ becomes

$$\hat{T}_n = \hat{\theta}_n + \{C_n' E_n^2(\hat{\theta}_n) C_n\}^{-1} C_n' E_n(\hat{\theta}_n) y_n(\hat{\theta}_n), \tag{3.9}$$

where $E_n(\theta) = \text{diag}\{\phi(c_i'\theta)\}$, ϕ is the standard normal density, $\hat{\theta}_n$ is a suitable initial estimate, and other notation is as in Section 3.3. The derivation of (3.9) strictly parallels that of (3.8).

On the other hand, a one-step version of the maximum likelihood estimate for θ in the probit model would be

$$T_n^* = \hat{\theta}_n + \{C_n' E_n^2(\hat{\theta}_n) D_n^{-1}(\hat{\theta}_n) C_n\}^{-1} C_n' E_n(\hat{\theta}_n) D_n^{-1}(\hat{\theta}_n) y_n(\hat{\theta}_n),$$

where $D_n(\theta) = \text{diag}[\pi_i(\theta)\{1 - \pi_i(\theta)\}]$.

The estimates $\hat{T}_n$ and T_n^* differ strikingly, except in the i.i.d. case. $\hat{T}_n$ assigns bounded weight to every observation and relatively little weight to observations associated with large or small values of $\pi_i(\theta)$. However, the classical estimate T_n^* gives relatively large weights to observations associated with extreme values of $\pi_i(\theta)$, because $\phi(x)/[\Phi(x)\{1 - \Phi(x)\}]$ becomes infinite as $x \to \pm\infty$. Thus, T_n^* is very sensitive to departures from the assumed probit model in the extreme probabilities $\pi_i(\theta)$.

The discussion in Sections 3.3 and 3.4 may be summarized as follows: (a) Classical maximum likelihood estimates are qualitatively robust in the logit model but not in the probit model; (b) the quantitative robustness theory of this paper recommends nonclassical estimates for both models, except in the i.i.d. case. These new estimates treat observations associated with very small or very large probabilities $\pi_i(\theta)$ more cautiously than do the classical estimates.

4. Proofs.

The proofs of Theorems 1 and 2 are organized around a concept of Hellinger differentiability for functionals of product measures. The key technical result is an asymptotic minimax lower bound for estimates of such functionals.

4.1. *Hellinger differentiable functionals.* Define a set H as follows (cf. Neveu, 1965, page 112, and Koshevnik and Levit, 1976): A typical element of H is a pair (ξ, P), usually written $\xi(dP)^{1/2}$, such that P is a probability on $(\mathscr{X}, \mathscr{A})$ and ξ is a random variable in $L_2(P)$. For simplicity, the element $1(dP)^{1/2}$ is written as $(dP)^{1/2}$. Suppose $\xi(dP)^{1/2}$ and $\eta(dQ)^{1/2}$ are elements of H and that $\nu = 2^{-1}(P + Q)$. Define the inner product

$$\langle \xi(dP)^{1/2}, \eta(dQ)^{1/2} \rangle = \int \xi\eta(dP/d\nu)^{1/2}(dQ/d\nu)^{1/2}\, d\nu \tag{4.1}$$

and, for arbitrary real a, b, the linear combination

$$a\xi(dP)^{1/2} + b\eta(dQ)^{1/2} = \{a\xi(dP/d\nu)^{1/2} + b\eta(dQ/d\nu)^{1/2}\}(d\nu)^{1/2}. \tag{4.2}$$

The choice of dominating probability ν does not affect these definitions. The corresponding norm $\| \cdot \|_H$ on H is given by

$$\| \xi(dP)^{1/2} \|_H^2 = \langle \xi(dP)^{1/2}, \xi(dP)^{1/2} \rangle = \int \xi^2\, dP. \tag{4.3}$$

In particular, $\| (dP)^{1/2} - (dQ)^{1/2} \|_H$ is the Hellinger distance between the probabilities P and Q.

Suppose $\phi = (\phi_1, \phi_2, \cdots, \phi_k)'$ is a random column vector whose components lie in $L_2(P)$. Then $\phi(dP)^{1/2}$ represents the vector $(\phi_1(dP)^{1/2}, \cdots, \phi_k(dP)^{1/2})'$. If $(dQ)^{1/2}$ belongs to H, $\langle \phi(dP)^{1/2}, \eta(dQ)^{1/2} \rangle$ represents the column vector of componentwise inner products.

Let $P^n = P_1 \times P_2 \times \cdots \times P_n$ be a product measure whose factors are probabilities on $(\mathscr{X}, \mathscr{A})$. Let Π^n be the set of all product measures and let

$$(4.4) \qquad H_n(P^n, c) = \{Q^n \in \Pi^n : \textstyle\sum_{i=1}^n \| (dQ_i)^{1/2} - (dP_i)^{1/2} \|_H^2 \le c^2\}.$$

Definition. A sequence of vector-valued functionals $\{S_n : \Pi^n \to R^n; n \ge 1\}$ is Hellinger differentiable at $\{P^n \in \Pi^n\}$ if there exist a triangular array of random $k \times 1$ vectors $\{\zeta_{i,n} : 1 \le i \le n;\ n \ge 1\}$ and a sequence of $k \times k$ matrices $\{A_n : n \ge 1\}$ which have the following properties:

(i) For every $n \ge 1$,

$$(4.5) \qquad \zeta_{i,n} \in L_2^k(P_i), \qquad \int \zeta_{i,n}\, dP_i = 0, \qquad 1 \le i \le n$$

and

$$(4.6) \qquad \textstyle\sum_{i=1}^n \displaystyle\int \zeta_{i,n} \zeta_{i,n}'\, dP_i = I_k,$$

the $k \times k$ identity matrix.

(ii) For every finite $c > 0$,

$$(4.7) \qquad \lim_{n\to\infty} \sup_{Q^n \in H_n(P^n,c)} | A_n\{S_n(Q^n) - S_n(P^n)\} - 2 \textstyle\sum_{i=1}^n \langle \zeta_{i,n}(dP_i)^{1/2}, (dQ_i)^{1/2} - (dP_i)^{1/2} \rangle | = 0.$$

(iii) The triangular array $\{\zeta_{i,n} : 1 \le i \le n;\ n \ge 1\}$ satisfies a Lindeberg condition: for every $\varepsilon > 0$ and every $d \in R^k$ of unit length,

$$(4.8) \qquad \lim_{n\to\infty} \textstyle\sum_{i=1}^n \displaystyle\int (d'\zeta_{i,n})^2 I(|d'\zeta_{i,n}| > \varepsilon)\, dP_i = 0.$$

Parts (i) and (ii) of this definition express the differentiability idea while part (iii) is important for asymptotic theory, describing how well such functionals can be estimated.

Suppose $X_1, X_2, \cdots, X_n$ are independent and that the distribution of X_i is Q_i. Let $\hat{S}_n = \hat{S}_n(X_1, X_2, \cdots, X_n)$ be any estimate of $S_n(Q^n)$ based on these random variables. As risk function take

$$(4.9) \qquad R_n(\hat{S}_n, Q^n) = E_{Q^n} u[|A_n\{\hat{S}_n - S_n(Q^n)\}|^2],$$

where A_n is the matrix appearing in (4.7), u is any monotone increasing bounded function mapping $R \to R$, and $|\cdot|$ is any metric on R^k.

Proposition 1. *If $\{S_n : \Pi^n \to R^k;\ n \ge 1\}$ is a sequence of functionals Hellinger differentiable at $\{P^n \in \Pi^n\}$, then*

$$(4.10) \qquad \lim_{c\to\infty} \liminf_n \inf_{\hat{S}_n} \sup_{Q^n \in H_n(P^n,c)} R_n(\hat{S}_n, Q^n) \ge r_0(u),$$

with $r_0(u)$ defined in (2.15).

Proposition 1 extends to i.n.i.d. sampling a result by Koshevnik and Levit (1976) on estimation of functionals in the i.i.d. case. The proof rests, in part, on the following fact.

Lemma 1. *Let $\{\zeta_{i,n} : 1 \le i \le n;\ n \ge 1\}$ be any triangular array satisfying* (4.5), (4.6) *and* (4.8). *For every $h \in R^k$, there exists a sequence of product probabilities $\{Q_n^n(h) =$*

$Q_{n,i}(h) \times \cdots \times Q_{n,n}(h) \in \Pi^n\}$ *such that*

$$(4.11) \qquad \lim_{n\to\infty} \sup_{|h|\le b} \textstyle\sum_{i=1}^n \| \{dQ_{n,i}(h)\}^{1/2} - (dP_i)^{1/2} - 2^{-1}h'\zeta_{i,n}(dP_i)^{1/2}\|_H^2 = 0$$

for every finite $b > 0$.

Proof. Let $\zeta_{i,n,j}$ be the jth component of $\zeta_{i,n}$, $1 \le j \le k$. The Lindeberg condition (4.8) implies that there exists a sequence $\{\varepsilon_n\}$ decreasing to 0 such that

$$(4.12) \qquad \lim_{n\to\infty} \max_{1\le j\le k} \textstyle\sum_{i=1}^n E_{P_i}\{\zeta_{i,n,j}^2 I(|\zeta_{i,n,j}| > \varepsilon_n)\} = 0.$$

For every possible i, j, n define

$$(4.13) \qquad \zeta_{i,n,j}^* = \begin{cases} \zeta_{i,n,j} & \text{if } \ |\zeta_{i,n,j}| \le \varepsilon_n, \\ 0 & \text{otherwise,} \end{cases}$$

and let $\bar\zeta_{i,n,j} = \zeta_{i,n,j}^* - \int \zeta_{i,n,j}^* \, dP_i$. Evidently, $\int \bar\zeta_{i,n,j}\, dP_i = 0$ and $|\bar\zeta_{i,n,j}| \le 2\varepsilon_n$.

Define the factors of the product probability $Q_n^n(h)$ by setting

$$(4.14) \qquad \frac{dQ_{n,i}(h)}{dP_i} = \begin{cases} 1 + h'\bar\zeta_{i,n} & \text{if } \ \varepsilon_n < (2bk)^{-1}, \\ 1 & \text{otherwise,} \end{cases}$$

where $\bar\zeta_{i,n} = (\bar\zeta_{i,n,1}, \bar\zeta_{i,n,2}, \cdots, \bar\zeta_{i,n,k})'$. It is easily checked that (4.14) does, in fact, describe probability densities on $(\mathscr{X}, \mathscr{A})$ whenever $|h| \le b$.

Verification of (4.11) for $\{Q_n^n(h)\}$ as defined above rests on the Taylor expansion of $(1 + z)^{1/2}$ to a linear term plus remainder and on the following facts:

$$(4.15) \qquad \textstyle\sum_{i=1}^n \displaystyle\int |\bar\zeta_{i,n}|^2\, dP_i \le \textstyle\sum_{i=1}^n \displaystyle\int |\zeta_{i,n,}|^2\, dP_i = k,$$

the last equality a consequence of (4.6), and

$$(4.16) \qquad \lim_{n\to\infty} \textstyle\sum_{i=1}^n \displaystyle\int |\bar\zeta_{i,n} - \zeta_{i,n}|^2\, dP_i = 0$$

because of (4.12).

Proof of Proposition 1. Lemma 1 and (4.6) imply that

$$(4.17) \qquad \lim_{n\to\infty} \sup_{|h|\le b} \textstyle\sum_{i=1}^n \| \{dQ_{n,i}(h)\}^{1/2} - (dP_i)^{1/2}\|_H^2 = 4^{-1}b^2.$$

Consequently,

$$(4.18) \qquad \liminf_n \inf_{\hat S_n} \sup_{Q^n\in H_n(P^n,c)} R_n(\hat S_n, Q^n) \ge \liminf_n \inf_{\hat S_n} \sup_{|h|\le c} R_n(\hat S_n, Q_n^n(h)).$$

Moreover, Lemma 1, (4.6) and (4.7) entail

$$(4.19) \qquad \lim_{n\to\infty} \sup_{|h|\le c} |A_n\{S_n(Q_n^n(h)) - S_n(P^n)\} - h| = 0.$$

Without loss of generality, we may assume that the function u is uniformly continuous as well as monotone increasing and bounded. Thus,

$$(4.20) \qquad \liminf_n \inf_{\hat S_n} \sup_{|h|\le c} R_n(\hat S_n, Q_n^n(h)) = \liminf_n \inf_{\hat V_n} \sup_{|h|\le c} E_{Q_n^n(h)}\, u(|\hat V_n - h|^2),$$

where $\hat V_n = A_n\{\hat S_n - S_n(P^n)\}$.

Let $L_n(h) = \log\{\Pi_{i=1}^n dQ_{n,i}(h)/dP_i\}$. Under $\{P^n\}$,

$$(4.21) \qquad L_n(h) = h'Z_n - 2^{-1}|h|^2 + o_p(1),$$

where $Z_n = \sum_{i=1}^n \zeta_{i,n}$; moreover $\{Z_n\}$ converges weakly under $\{P^n\}$ to the standard k-dimensional normal distribution; cf. Ibragimov and Khasminskii (1975) for both assertions. It follows by the Hájek-LeCam asymptotic minimax theorem (Hájek, 1972; Le Cam, 1972) that

$$(4.22) \qquad \lim_{c\to\infty} \liminf_n \inf_{\hat V_n} \sup_{|h|\le c} \mathrm{E}_{Q_n^n(h)}\, u(|\hat V_n - h|^2) \ge r_0(u).$$

Combining (4.18), (4.20),and (4.22) yields Proposition 1.

4.2. *Proof of Theorem* 1. Evidently, the neighborhood $B_n(\theta, c)$ defined in (2.10) contains the Hellinger ball $H_n(P_\theta^n, 2^{-1}c)$. In view of Proposition 1, it suffices to show that the functionals $\{T_n(\theta, Q^n)\}$ defined in (2.7) are Hellinger differentiable at $\{P^n\}$, with

$$\zeta_{i,n,}(x) = C_n^{-1/2}(\theta)\rho_{\theta,i}(x), \qquad A_n = A_n(\theta) = C_n^{-1/2}(\theta)V_n(\theta). \tag{4.23}$$

Let $\nu_i = 2^{-1}(P_{\theta,i} + Q_i)$ and set $p_{\theta,i} = dP_{\theta,i}/d\nu_i$, $q_i = dQ_i/d\nu_i$. From (2.8) and (2.9), it follows that

$$A_n(\theta)\{T_n(\theta, Q^n) - \theta\} = 2\textstyle\sum_{i=1}^n \langle \zeta_{i,n}(dP_{\theta,i})^{1/2}, (dQ_i)^{1/2} - (dP_{\theta,i})^{1/2}\rangle + r_n(\theta, Q^n), \tag{4.24}$$

where

$$r_n(\theta, Q^n) = \textstyle\sum_{i=1}^n \displaystyle\int \zeta_{i,n}(q_i^{1/2} - p_{\theta,i}^{1/2})^2\, d\nu_i. \tag{4.25}$$

Under Assumption A2,

$$\lim_{n\to\infty} \sup_{Q^n\in H_n(P_\theta^n, c)} |r_n(\theta, Q^n)| = 0 \tag{4.26}$$

and the triangular array $\{\zeta_{i,n} : 1 \le i \le n;\, n \ge 1\}$ satisfies the Lindeberg condition (4.8) at $\{P_\theta^n\}$. Thus the functionals $\{T_n(\theta, Q^n)\}$ are Hellinger differentiable at $\{P_\theta^n\}$ as asserted.

4.3. *Proof of Theorem* 2. From (2.18), (2.21), and (4.23), it follows that

$$A_n(\theta)\{\hat{T}_n - T_n(\theta, Q_n^n)\} - \textstyle\sum_{i=1}^n \left(\zeta_{i,n}(x_i) - \displaystyle\int \zeta_{i,n}\, dQ_{n,i}\right) \to_{Q_n^n} 0 \tag{4.27}$$

under every sequence $\{Q_n^n \in B_n(\theta, c) : n \ge 1\}$. Since u is bounded and continuous a.e., it suffices to show that the sum in (4.27) converges weakly, under every sequence $\{Q_n^n \in B_n(\theta, c)\}$, to the standard k-dimensional normal distribution. Let $\xi_{i,n} = d'\zeta_{i,n}$ where d is an arbitrary unit vector in R^k. Let $s_n^2 = \sum_{i=1}^n \mathrm{Var}_{Q_{n,i}}(\xi_{i,n})$. The desired weak convergence will be proved by verifying the appropriate Lindeberg condition

$$\lim_{n\to\infty} s_n^{-2} \textstyle\sum_{i=1}^n \displaystyle\int \left(\xi_{i,n} - \int \xi_{i,n}\, dQ_{n,i}\right)^2 I\left(\left|\xi_{i,n} - \int \xi_{i,n}\, dQ_{n,i}\right| > \varepsilon s_n\right) dQ_{n,i} = 0. \tag{4.28}$$

The term inside the limit in (4.28) is bounded above by

$$\begin{aligned} &s_n^{-2} \sup_i \sup_x \left\{\xi_{i,n}(x) - \int \xi_{i,n}\, dQ_{n,i}\right\}^2 \textstyle\sum_{i=1}^n Q_{n,i}\left(\left|\xi_{i,n} - \displaystyle\int \xi_{i,n}\, dQ_{n,i}\right| > \varepsilon s_n\right) \\ &\qquad \le 2\varepsilon^{-2}s_n^{-2}\left\{\sup_i \sup_x \xi_{i,n}^2(x) + \sup_i\left(\int \xi_{i,n}\, dQ_{n,i}\right)^2\right\}. \end{aligned} \tag{4.29}$$

In view of this bound and Assumption A2, (4.28) can be established by showing that

$$\lim_{n\to\infty} \textstyle\sum_{i=1}^n \left(\displaystyle\int \xi_{i,n}\, dQ_{n,i}\right)^2 = 0 \tag{4.30}$$

and

$$\lim_{n\to\infty} \textstyle\sum_{i=1}^n \displaystyle\int \xi_{i,n}^2\, dQ_{n,i} = 1. \tag{4.31}$$

From the definitions (4.23) and (2.8) of $\zeta_{i,n}$ and $\rho_{\theta,i}$, it is apparent that

$$\begin{aligned} \int \xi_{i,n}\, dQ_{n,i} = &\int \{G_{n,i}(x) - F_{\theta,i}(x)\}\, d'C_n^{-1/2}\, \gamma_{\theta,i}(x + a)\, d\mu_{\theta,i}(x + a) \\ &- \int \{G_{n,i}(x) - F_{\theta,i}(x)\}\, d'C_n^{-1/2}(\theta)\gamma_{\theta,i}(x + b)\, d\mu_{\theta,i}(x + b). \end{aligned} \tag{4.32}$$

Moreover,

$$\sum_{i=1}^{n}\left[\int \{G_{n,i}(x) - F_{\theta,i}(x)\}\, d'C_n^{-1/2}(\theta)\gamma_{\theta,i}(x+h)\, d\mu_{\theta,i}(x+h)\right]^2$$

$$\leq \sup_i\left\{d'C_n^{-1/2}(\theta)\int \gamma_{\theta,i}\gamma'_{\theta,i}\, d\mu_{\theta,i}C_n^{-1/2}(\theta)d\right\}\sum_{i=1}^{n}\int \{G_{n,i}(x) - F_{\theta,i}(x)\}^2\, d\mu_{\theta,i}(x+h), \tag{4.33}$$

which tends to zero, for $h = a$ or b, as $n \to \infty$ because of (2.10) and Assumption A3. This proves (4.30). The argument for (4.31) is similar.

REFERENCES

Beran, R. J. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* **5** 445–463.

Beran, R. J. (1981a). Efficient robust estimates in parametric models *Z. Wahrsch. verw. Gebiete* **55** 91–108.

Beran, R. J. (1981b). Efficient robust tests in parametric models.`*Z. Wahrsch. verw. Gebiete* **57** 73–86.

Bickel, P. J. (1976). Another look at robustness: a review of reviews and some new developments. *Scand. J. Statist.* **3** 145–168.

Bickel, P. J. and Lehmann, E. L. (1975). Descriptive statistics for nonparametric models, I, II. *Ann. Statist.* **1** 597–616.

Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. *Proc. Sixth Berkeley Symp. Math. Statist. Probability* **1** 175–194. University of California Press.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393.

Holland, P. W. and Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Commun. Statist.-Theor. Meth.* **A6** 813–827.

Holm, S. (1976). Discussion of paper by P. J. Bickel. *Scand. J. Statist.* **3** 158–161.

Hogg, R. G. (1974). Adaptive robust procedures; a partial review and some suggestions for future applications and theory. *J. Amer. Statist. Assoc.* **69** 909–925.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures, and Monte Carlo. *Ann. Statist.* **1** 799–821.

Ibragimov, I. A. and Khas'Minskii, R. Z. (1975). Local asymptotic normality for non-identically distributed observations. *Theory. Probability Appl.* **20** 246–260.

Koshevnik, Yu. A. and Levit, B. Ya. (1976). On a nonparametric analogue of the information matrix. *Theory Probability Appl.* **21** 738–753.

LeCam, L. (1972). Limits of experiments. *Proc. Sixth Berkeley Symp. Math. Statist. Probability* **1** 245–261. University of California Press.

Millar, P. W. (1981). Robust estimation via minimum distance methods. *Z. Wahrsch verw. Gebiete* **55** 73–89.

Parr, W. C. and Schucany, W. R. (1980). Minimum distance and robust estimation. *J. Amer. Statist. Assoc.* **75** 616–624.

Rieder, H. (1980). Estimates derived from robust tests. *Ann. Statist.* **8** 106–115.

Department of Statistics
University of California, Berkeley
Berkeley, California 94720