Z. Wahrscheinlichkeitstheorie verw. Gebiete
55, 91–108 (1981)

Zeitschrift für
Wahrscheinlichkeitstheorie
und verwandte Gebiete
© Springer-Verlag 1981

Efficient Robust Estimates in Parametric Models*

Rudolf Beran

Dept. of Statistics, University of California, Berkeley, CA 94720, USA

Summary. Let $\{P_\theta^n\colon \theta\in\Theta\}$, Θ an open subset of R^k, be a regular parametric model for a sample of n independent, identically distributed observations. This paper describes estimates $\{T_n;\, n\geqq 1\}$ of θ which are asymptotically efficient under the parametric model and are robust under small deviations from that model. In essence, the estimates are adaptively modified, one-step maximum likelihood estimates, which adjust themselves according to how well the parametric model appears to fit the data. When the fit seems poor, T_n discounts observations that would have large influence on the value of the usual one-step MLE. The estimates $\{T_n\}$ are shown to be asymptotically minimax, in the Hájek-LeCam sense, for a Hellinger ball contamination model. An alternative construction of robust asymptotically minimax estimates, as modified MLE's, is described for canonical exponential families.

1. Introduction

The most satisfactory account to date of what constitutes optimal estimation in large samples from a parametric model is provided by the Hájek (1972) and LeCam (1972) asymptotic minimax theorem. Foreshadowed by earlier work, notably LeCam (1953), the paper by Dvoretzky, Kiefer, Wolfowitz (1956) on the empirical c.d.f., and a result attributed to C. Stein and to H. Rubin in Chernoff (1956), the asymptotic minimax theorem has several desirable features. Asymptotic optimality is established over the class of all procedures; for the regular models of classical parametric estimation theory, the asymptotic minimax theorem leads to the most general form of the Fisher information bound; parametric models which are not regular in this sense can also be treated. Offsetting these virtues is the fact that the theorem refers only to the local behavior of procedures in large samples. Nevertheless, asymptotic minimax is a powerful criterion for identifying good statistical methods.

* This research was supported in part by National Science Foundation Grant MCS 75-10376

0044-3719/81/0055/0091/$03.60

Recent work has explored asymptotic minimax ideas in more complicated statistical contexts, including nonparametrics and robust estimation in parametric models. (For examples of the former, see Koshevnik and Levit (1976), Moussatat (1976), Millar (1979); for the latter, Beran (1980).) In some problems, such as nonparametric estimation of a c.d.f., familiar procedures turn out to be asymptotically minimax. However, in other areas, such as robust estimation of a parameter or nonparametric density estimation, asymptotically minimax procedures are not obvious and may depend strongly upon the mathematical formulation of the problem. The understanding of possibilities to be gained in such cases makes the search for asymptotically minimax procedures particularly interesting.

Robust estimation in parametric models is our present theme. As in an earlier paper (Beran, 1980), we take the fundamental goal of robust estimation to be the estimation of the actual distribution of the sample. A parameter estimate determines a fitted parametric distribution, that member of the parameter model which is identified by the parameter estimate. The fitted parametric distribution is regarded as an estimate of the actual distribution. Questions of efficiency and robustness are addressed by comparing the fitted parametric distribution with the actual distribution. Where this paper and the earlier one differ importantly is in the contamination model - the set of distributions outside the parametric model which are regarded as possible for the data. The larger set treated in this paper leads to a more interesting and, we hope, more useful theory of robust estimation.

Let $\{P_\theta^n\colon \theta\in\Theta\}$ denote the parametric model for a sample of size n; let T_n be an estimate of θ based upon this sample; and let Q^n be the actual distribution of the sample. The adequacy of $P_{T_n}^n$ as an estimate of Q^n can be measured by the maximum risk, $\sup_{Q^n} E_{Q^n} l[d(P_{T_n}^n, Q^n)]$, attained over all probabilities Q^n in some neighborhood of P_θ^n. Here d is a metric on probabilities and l is a reasonable loss function. The goal is to find an estimate T_n which minimizes the maximum risk, at least for all sufficiently large sample sizes.

Further developments in this approach to robust estimation depend on several choices. Apart from some technicalities, it will be assumed in this paper that d is Hellinger metric; l is non-negative and monotone increasing; the probabilities P_θ^n and Q^n are product measures; the marginal probabilities $\{P_\theta\colon \theta\in\Theta\}$ are quadratic mean differentiable in θ; and the neighborhood of P_θ^n to which Q^n belongs is a Hellinger ball. The most significant of these choices, from the viewpoint of robustness, are the first and last.

As will be shown, these specific assumptions result in an optimistic robustness theory under which adaptively modified, one-step maximum likelihood estimates of θ are asymptotically minimax. These optimal robust estimates $\{T_n\}$ adjust themselves according to how well the parametric model $\{P_\theta^n\colon \theta\in\Theta\}$ appears to fit the data. When the fit seems good, T_n is virtually identical with the usual (LeCam, 1956) one-step MLE of θ; however, when the fit seems poor, T_n discounts observations that would have unduly large influence upon the value of the one-step MLE. The transition in the behavior of T_n occurs continuously according to the estimated degree-of-fit of the parametric model. Moreover, the $\{T_n\}$ are asymptotically efficient as estimates of θ within the parametric model $\{P_\theta^n\colon \theta\in\Theta\}$.

It is interesting that relatively simple decision theoretic considerations based upon a Hellinger ball contamination model should lead to estimates $\{T_n\}$ which embody two older themes in refined form: the idea that robustness can be achieved through sensible adaptation (cf. the review paper by Hogg (1974)); and the idea that asymptotic efficiency at the parametric model is compatible with a reasonable amount of robustness (Beran, 1977, 1978).

The Hellinger ball contamination model might be questioned on the grounds that it fails to allow for round-off errors in the data and, therefore, must be too restricted. Actually, rounding-off of observations causes no conceptual difficulty if, in modelling, densities with respect to Lebesgue measure are re-interpreted (after rescaling) as densities with respect to a more realistic counting measure. We would also argue by results. The non-robustness of the sample mean is already evident under the Hellinger ball contamination model (Example 1 of Sect. 2); the corresponding asymptotically minimax estimates $\{T_n\}$ appear robust. Fuller discussion awaits practical experience with these estimates and the investigation of asymptotically minimax estimates for other, more severe, contamination neighborhoods.

The technical results are organized as follows. Section 2 gives an asymptotic lower bound for the maximum risk achieved by any estimate of θ over a Hellinger ball contamination neighborhood. The existence of robust estimates $\{T_n\}$ whose maximum risk attains the lower bound asymptotically is demonstrated by the constructions in Sects. 3 and 4. Section 3 handles the general case of quadratic mean differentiable models, while Sect. 4 describes adaptively modified maximum likelihood estimates which work in canonical exponential families.

2. Asymptotic Minimax Bound

The following notation is useful in expressing the assumptions and results of this paper. Let Π be the set of all probabilities on a measurable space. Define a set H as follows (cf. Neveu (1965), p. 112; Koshevnik and Levit (1976)): A typical element of H is a pair (ξ, P), usually written $\xi(dP)^{1/2}$, such that $P \in \Pi$ 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 that $\xi(dP)^{1/2}$ and $\eta(dQ)^{1/2}$ are elements of H and that $\mu = 2^{-1}(P + Q)$. Define the inner product

$$\langle \xi(dP)^{1/2}, \eta(dQ)^{1/2} \rangle = \int \xi\eta \left(\frac{dP}{d\mu}\right)^{1/2} \left(\frac{dQ}{d\mu}\right)^{1/2} d\mu \tag{2.1}$$

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

$$a\xi(dP)^{1/2} + b\eta(dQ)^{1/2} = \left[a\xi \left(\frac{dP}{d\mu}\right)^{1/2} + b\eta \left(\frac{dQ}{d\mu}\right)^{1/2}\right] (d\mu)^{1/2}. \tag{2.2}$$

The corresponding norm $\|\cdot\|$ on H is given by

$$\begin{aligned} \|\xi(dP)^{1/2}\|^2 &= \langle \xi(dP)^{1/2}, \xi(dP)^{1/2} \rangle \\ &= \int \xi^2 \, dP. \end{aligned} \tag{2.3}$$

In particular, $\|(dP)^{1/2}-(dQ)^{1/2}\|$ is the Hellinger distance between the probabilities $P, Q \in \Pi$. The elements $\xi(dP)^{1/2}$ and $\eta(dQ)^{1/2}$ of H are said to be equivalent if $\|\xi(dP)^{1/2}-\eta(dQ)^{1/2}\|=0$. The set $\bar{H}$ of equivalence classes in H forms a Hilbert space with the above inner product and norm.

Suppose $\phi=(\phi_1, \phi_2, \ldots, \phi_n)'$ is a random vector whose components lie in $L_2(P)$. Then $\phi(dP)^{1/2}$ represents the vector $(\phi_1(dP)^{1/2}, \phi_2(dP)^{1/2}, \ldots, \phi_n(dP)^{1/2})'$ and $\|\phi(dP)^{1/2}\|^2$ means $\int |\phi|^2 dP$. Similarly, if $\eta(dQ)^{1/2}$ belongs to H, $\langle\phi(dP)^{1/2}, \eta(dQ)^{1/2}\rangle$ denotes the column vector of componentwise inner products.

Let $\{P^n=P_\theta \times P_\theta \times \ldots \times P_\theta\, n\text{-times}: \theta \in \Theta\}$ be the parametric model for a sample of size n. Suppose that $Q^n=Q \times Q \times \ldots \times Q$ n-times is the actual distribution of the sample. Both $\{P_\theta: \theta \in \Theta\}$ and Q are probabilities on a Euclidean space $\mathscr{X}$ with Borel σ-algebra $\mathscr{B}$. The $\{P_\theta: \theta \in \Theta\}$ are assumed to satisfy the following regularity conditions. The parameter space Θ is an open subset of R^k. The mapping $\theta \to P_\theta$ has the property that

(i) for every $\theta \in \Theta$, there exists $\eta_\theta \in L_2^k(P_\theta)$ such that

$$\lim_{t \to 0} |t|^{-1} \|(dP_{\theta+t})^{1/2}-(dP_\theta)^{1/2}-t'\eta_\theta(dP_\theta)^{1/2}\|=0; \tag{2.4}$$

(ii) for every $\theta \in \theta$, the Fisher information matrix

$$I(\theta)=4\int \eta_\theta \eta'_\theta \, dP_\theta \tag{2.5}$$

is non-singular. It is clear that the quadratic mean derivative η_θ is unique, up to equivalence in $L_2^k(P_\theta)$, and that $\int \eta_\theta \, dP_\theta=0$.

Let T_n be an estimate of θ based upon the sample of size n. The risk of T_n discussed in the introduction is a function of the squared Hellinger distance $\|(dP_{T_n}^n)^{1/2}-(dQ^n)^{1/2}\|^2$. Since

$$\|(dP_{T_n}^n)^{1/2}-(dQ^n)^{1/2}\|^2=2-2\{1-2^{-1}\|(dP_{T_n})^{1/2}-(dQ)^{1/2}\|^2\}^n, \tag{2.6}$$

it is equally reasonable, in the i.i.d. sampling situation, to consider risks which depend on T_n through $\|(dP_{T_n})^{1/2}-(dQ)^{1/2}\|^2$. Suppose Q^n lies in a Hellinger ball centered at P_θ^n and of radius less than $\sqrt{2}$. Replacing $(dP_{T_n})^{1/2}$ with the hyperplane approximation implied by (2.4) suggests risks which are functions of

$$\|(T_n-\theta)'\eta_\theta(dP_\theta)^{1/2}+(dP_\theta)^{1/2}-(dQ)^{1/2}\|^2. \tag{2.7}$$

The difference $(dQ)^{1/2}-(dP_\theta)^{1/2}$ has an orthogonal decomposition in H, one component of which is the projection into the subspace spanned by the components of the vector $\eta_\theta(dP_\theta)^{1/2}$. Thus, (2.7) can be expressed as the sum of two terms, only one of which depends on T_n. This term equals $4^{-1}(T_n - T(\theta, Q))' I(\theta)(T_n-T(\theta, Q))$, where

$$T(\theta, Q)=\theta+4I^{-1}(\theta)\langle\eta_\theta(dP_\theta)^{1/2}, (dQ)^{1/2}-(dP_\theta)^{1/2}\rangle. \tag{2.8}$$

For every $c \in (0, 2)$, let $S_n(\theta, c)$ be the set of all probabilities Q on $(\mathscr{X}, \mathscr{B})$ such that $\|(dQ^n)^{1/2}-(dP_\theta^n)^{1/2}\|^2 \leqq 2-c$. In view of the preceding paragraph, the adequacy of T_n as a robust estimate can be measured through the risk

$$R_n(T_n, Q) = E_{Q^n} u[n(T_n - T(\theta, Q))' I(\theta)(T_n - T(\theta, Q))], \tag{2.9}$$

where u is a monotone loss function. Our goal is to choose T_n so as to minimize $\sup\{R_n(T_n, Q)\colon Q \in S_n(\theta, c)\}$, whatever the value of $\theta \in \Theta$. The function u will be required to have the following statistically reasonable properties: u is real-valued, monotone increasing on $[0, \infty)$ with $u(0) = 0$, and

$$R_0(u) = \int u(|z|^2)\, \phi_k(z)\, dz < \infty, \tag{2.10}$$

ϕ_k being the standard k-dimensional normal density.

It is clear from (2.4) and (2.8) that

$$T(\theta, P_{\theta + n^{-1/2}h}) = \theta + n^{-1/2} h + o(n^{-1/2}) \tag{2.11}$$

for every $h \in R^k$. Under some further assumptions on $\{P_\theta\colon \theta \in \Theta\}$, it can be shown that $T(\theta, Q)$ approximates that value of $t \in \Theta$ which minimizes $\|(dP_t)^{1/2} - (dQ)^{1/2}\|$; for details see Beran (1977).

The following theorem gives an asymptotic minimax bound for the robust estimation problem just described.

Theorem 1. *Under the assumptions described above,*

$$\lim_{c \to 0} \liminf_{n} \inf_{T_n} \sup_{Q \in S_n(\theta, c)} R_n(T_n, Q) \geqq R_0(u) \tag{2.12}$$

for every $\theta \in \Theta$.

Proof. It suffices to prove (2.12) under the additional assumption that u is bounded and continuous, since the theorem, as stated, then follows. Fix $\theta \in \Theta$ and let $\theta_n = \theta + n^{-1/2} h$, where $h \in R^k$. Hájek (1972) and Le Cam (1972) have shown that

$$\lim_{b \to \infty} \liminf_{n} \inf_{T_n} \sup_{|h| \leqq b} E_{P^n_{\theta_n}} u[n(T_n - \theta_n)' I(\theta)(T_n - \theta_n)] \geqq R_0(u). \tag{2.13}$$

This inequality remains valid under the following substitutions: first replace the risk with $R_n(T_n, P_{\theta_n})$, using (2.11); then replace $R_n(T_n, P_{\theta_n})$ with $R_n(T_n, Q)$ and the supremum over $|h| \leqq b$ with the supremum over all probabilities Q such that $n\|(dQ)^{1/2} - (dP_\theta)^{1/2}\|^2 \leqq b^2$. This step makes use of (2.4) and the nonsingularity of $I(\theta)$. Because of the identity (2.6), with T_n replaced by θ, the inequality $n\|(dQ)^{1/2} - (dP_\theta)^{1/2}\|^2 \leqq b^2$ implies $\|(dQ^n)^{1/2} - (dP_\theta^n)^{1/2}\| \leqq 2 - 2\exp(-b^2/2) + o(1)$ for all large n. The theorem follows from (2.13) and these considerations.

The main result of this paper is the asymptotic attainability of the lower bound in (2.12).

Theorem 2. *In addition to the assumptions for Theorem* 1, *suppose that u is bounded and that the mapping $\theta \to P_\theta$ is one-to-one. Then, there exist robust estimates $\{T_n;\, n \geqq 1\}$ such that*

$$\lim_{n \to \infty} \sup_{Q \in S_n(\theta, c)} R_n(T_n, Q) = R_0(u) \tag{2.14}$$

for every $\theta \in \Theta$ and every $c \in (0, 2)$.

Such asymptotically minimax estimates $\{T_n\}$ will be constructed in Sect. 3. The example below, exhibiting the misbehaviour of the sample mean, supports our view that the Hellinger ball contamination model is useful in robustness studies. Interestingly, the proof of Theorem 1 implies that the parametric family $\{P_\theta\colon \theta\in\Theta\}$ itself contains a least favorable distribution. Thus, the minimax criterion can hardly be regarded as too pessimistic in this context.

Example 1. Suppose Θ is the real line, P_θ is $N(\theta, 1)$, the loss function u is as in Theorem 1, and M_n is the sample mean. Then, for every $\theta\in\Theta$,

$$\lim_{c\to 0}\lim_{n\to\infty}\sup_{Q\in S_n(\theta,c)} R_n(M_n, Q)=u(\infty), \tag{2.15}$$

which exceeds $R_0(u)$ unless u is trivially constant.

To verify (2.15), fix $\theta\in\Theta$ and let $\{a_n; n\geqq 1\}$ be a sequence of positive constants such that $\lim\limits_{n\to\infty} n^{-1/2}a_n=\infty$. Let $D(a)$ be the atomic probability supported on the point a and define

$$Q_n=(1-bn^{-1})P_\theta+bn^{-1}D(a_n+\theta) \tag{2.16}$$

for some $b>0$. It is clear that $\lim\limits_{n\to\infty}\|(dQ_n^n)^{1/2}-(dP_\theta^n)^{1/2}\|^2=2-2\exp(-b/2)$ and that the probability sequences $\{Q_n^n; n\geqq 1\}$ and $\{P_\theta^n; n\geqq 1\}$ are not contiguous (because the log-likelihood ratio of Q_n^n with respect to P_θ^n converges in P_θ^n-probability to $-b$).

Observe that $T(\theta, Q_n)=\theta$, because $D(a_n+\theta)$ and P_θ are mutually singular, and that

$$\begin{aligned}
R_n(M_n, Q_n)&=E_{Q_n^n}u[n(M_n-\theta)^2]\\
&=\sum_{r=0}^{n}\binom{n}{r}(bn^{-1})^r(1-bn^{-1})^{n-r}Eu[n^{-1}|(n-r)^{1/2}Z+ra_n|^2]\\
&\geqq(1-bn^{-1})^n Eu(|Z|^2)\\
&\quad+[1-(1-bn^{-1})^n]\,Eu[|\max\{0, n^{-1/2}a_n-|Z|\}|^2],
\end{aligned}$$

where Z has a $N(0, 1)$ distribution. Hence,

$$\liminf_n R_n(M_n, Q_n)\geqq\exp(-b)R_0(u)+[1-\exp(-b)]\,u(\infty), \tag{2.18}$$

which implies (2.15). The sample mean fails to be asymptotically minimax, even for bounded u, because it is too sensitive to distant outliers.

Remark. The asymptotic minimax approach to robustness is inherently local: for every $Q\in S_n(\theta, c)$, $n\|(dQ)^{1/2}-(dP_\theta)^{1/2}\|^2\leqq -2\log(c/2)$, so that the lack-of-fit of the parametric model to the actual marginal distribution Q decreases as sample size increases. This feature of the contamination model is advantageous for assessing the effect of small but potentially troublesome departures from the parametric model (Example 1 is an instance). However, by confounding sample size with contamination amount, it is possible to devise trivial estimates $\{T_n\}$ which satisfy

(2.14) without being robust globally (see the remarks in Sect. 3.3). Theorem 2 asserts the existence of estimates $\{T_n\}$ which satisfy (2.14) and behave sensibly under non-local contamination.

3. Asymptotically Minimax Estimates

Let $\{P_\theta: \theta \in \Theta\}$ be a parametric model which has properties (2.4) and (2.5). We will prove Theorem 2 in Subsect. 3.4 by constructing robust estimates $\{T_n; n \geqq 1\}$ which satisfy (2.14). Notation and lemmas needed for the construction are described in Subsects. 3.1 to 3.4.

3.1 Initial Estimates

The first lemma asserts in a formal way the existence of estimates $\{\theta_n^*; n \geqq 1\}$ for θ which are robust under the Hellinger ball contamination model. The estimates $\{\theta_n^*\}$, which are functions of the sample $(x_1, x_2, \ldots, x_n)$, provide the starting point for the construction of asymptotically minimax estimates $\{T_n\}$.

Lemma 1. *Suppose $\{P_\theta: \theta \in \Theta\}$ satisfies* (2.4), (2.5) *and the mapping $\theta \to P_\theta$ is one-to-one. Then there exist estimates $\{\theta_n^*; n \geqq 1\}$ which have the following property for every $\theta \in \Theta$. Suppose $\{Q_n; n \geqq 1\}$ is any sequence of probabilities on $(\mathscr{X}, \mathscr{B})$ such that $\{n^{1/2}[(dQ_n)^{1/2} - (dP_\theta)^{1/2}]; n \geqq 1\}$ converges weakly to an element in H. The distributions of $\{n^{1/2}(\theta_n^* - \theta); n \geqq 1\}$ under $\{Q_n^n\}$ are tight.*

Extension of the argument on pp. 104–107 of Le Cam (1969) shows that a modified minimum Kolomogorov distance estimate of θ will serve as the θ_n^* described in this lemma. Other constructions of such $\{\theta_n^*\}$ are possible in particular parametric models $\{P_\theta: \theta \in \Theta\}$, as will be illustrated for exponential families in Sect. 4.

A discretized version $\hat{\theta}_n$ of θ_n^* is defined as follows (cf. Le Cam (1969)). Let d be an arbitrary positive constant. Cover the parameter space $\Theta \subset R^k$ with disjoint semi-closed hypercubes of side length $n^{-1/2}d$. Set $\hat{\theta}_n$ equal to the center of the hypercube which contains θ_n^*. Tightness of $\{n^{1/2}(\theta_n^* - \theta)\}$ under $\{Q_n^n\}$ clearly implies tightness of $\{n^{1/2}(\hat{\theta}_n - \theta)\}$.

3.2 Pseudo-Derivatives

Under the assumptions on $\{P_\theta: \theta \in \Theta\}$, the quadratic mean derivative η_θ appearing in (2.4) need not be continuous in θ. To avoid technical difficulties, we will replace η_θ with a sequence of related difference quotients, defined in (3.5) below.

Separability of Θ and quadratic mean continuity of P_θ in θ ensure the existence of a probability μ on $(\mathscr{X}, \mathscr{B})$ such that $P_\theta \ll \mu$ for every $\theta \in \Theta$. Let $p_\theta(x) = \frac{dP_\theta}{d\mu}(x)$. Let $\{e_i; 1 \leqq i \leqq k\}$ be the usual orthonormal basis for R^k; e_j is a column vector whose j^{th} component is 1 and whose other components are 0. For $x \in \mathscr{X}$ and $\theta \in \Theta$ define

$$\psi_{n,\theta}(x) = \begin{cases} \sum_{j=1}^{k} n^{1/2}[p_\theta^{-1/2}(x)\, p_{\theta+n^{-1/2}e_j}^{1/2}(x) - 1]\, e_j & \text{if } p_\theta(x) > 0 \\ 0 & \text{if } p_\theta(x) = 0. \end{cases} \tag{3.1}$$

Lemma 2. *Suppose the assumptions of Lemma* 1 *are satisfied. Let* $\{Q_n; n \geqq 1\}$ *be any sequence of probabilities on* $(\mathscr{X}, \mathscr{B})$ *such that* $\{n^{1/2}[(dQ_n)^{1/2} - (dP_\theta)^{1/2}]; n \geqq 1\}$ *converges weakly to an element in H. Then*

$$\|\psi_{n,\hat\theta_n}(dP_{\hat\theta_n})^{1/2} - \eta_\theta(dP_\theta)^{1/2}\| \xrightarrow{Q_n^n} 0 \tag{3.2}$$

as $n \to \infty$.

Proof. From (2.4) and the property of $\hat\theta_n$ described in Lemma 1,

$$\|n^{1/2}[(dP_{\hat\theta_n})^{1/2} - (dP_\theta)^{1/2}] - n^{1/2}(\hat\theta_n - \theta)'\, \eta_\theta(dP_\theta)^{1/2}\| \xrightarrow{Q_n^n} 0 \tag{3.3}$$

and

$$\|n^{1/2}[(dP_{\hat\theta_n+n^{-1/2}e_j})^{1/2} - (dP_\theta)^{1/2}] - [n^{1/2}(\hat\theta_n - \theta) + e_j]'\, \eta_\theta(dP_\theta)^{1/2}\| \xrightarrow{Q_n^n} 0. \tag{3.4}$$

Let $B_n = \{x : p_{\hat\theta_n}(x) > 0\}$. Then

$$\begin{aligned} &\int_{B_n} |\psi_{n,\hat\theta_n}(x)\, p_{\hat\theta_n}^{1/2}(x) - \eta_\theta(x)\, p_\theta^{1/2}(x)|^2\, d\mu \\ &\quad = \sum_{j=1}^{k} \int_{B_n} [n^{1/2}(p_{\hat\theta_n+n^{-1/2}e_j}^{1/2}(x) - p_{\hat\theta_n}^{1/2}(x)) - e_j'\eta_\theta(x)\, p_\theta^{1/2}(x)]^2\, d\mu \\ &\quad \leqq \sum_{j=1}^{k} \|n^{1/2}[(dP_{\hat\theta_n+n^{-1/2}e_j})^{1/2} - (dP_{\hat\theta_n})^{1/2}] - e_j'\eta_\theta(dP_\theta)^{1/2}\|^2 \xrightarrow{Q_n^n} 0 \end{aligned} \tag{3.5}$$

by (3.3) and (3.4).

On the other hand, $P_\theta(\bar B_n) \xrightarrow{Q_n^n} 0$ because $P_{\hat\theta_n}(\bar B_n) = 0$ and the variation norm $\|P_\theta - P_{\hat\theta_n}\|_{\text{var}}$ is bounded from above by $2\|(dP_\theta)^{1/2} - (dP_{\hat\theta_n})^{1/2}\|$, which converges to zero under Q_n^n because of (3.3). Thus

$$\begin{aligned} &\int_{\bar B_n} |\psi_{n,\hat\theta_n}(x)\, p_{\hat\theta_n}^{1/2}(x) - \eta_\theta(x)\, p_\theta^{1/2}(x)|^2\, d\mu \\ &\quad = \int_{\bar B_n} |\eta_\theta(x)|^2\, dP_\theta \xrightarrow{Q_n^n} 0. \end{aligned} \tag{3.6}$$

The lemma follows from (3.5) and (3.6).

3.3 Random Windows

The asymptotically minimax estimate T_n constructed in Subsect. 3.4 depends upon a random window w_n which has the technical properties described in Lemma 3 below. The role of w_n is to discount observations that would otherwise have undue influence upon the value of T_n.

Let m be an absolutely continuous function mapping R^+ into $[0, 1]$ such that $m(0)=1$, $\sup_{x>0}[xm(x)]<\infty$, and the derivative m' is bounded. Let $\{Q_n; n \geqq 1\}$ be any sequence of probabilities on $(\mathscr{X}, \mathscr{B})$ such that $\{n^{1/2}[(dQ_n)^{1/2}-(dP_\theta)^{1/2}]; n \geqq 1\}$ converges weakly to an element in H. Let $\{c_n^*; n \geqq 1\}$ be positive, real-valued statistics such that c_n^* is a function of the sample $(x_1, x_2, \dots, x_n)$ and $\{n^{1/4-\delta} c_n^*\}$ is tight under $\{Q_n^n\}$ for some $\delta \in (0, 1/4)$. Discretize c_n^* as follows. Let b be an arbitrary positive constant. Starting with the origin as an endpoint, cover the positive real axis with disjoint semi-closed intervals of length $n^{-1/4+\delta} b$. Set $\hat{c}_n$ equal to the center of the interval which contains c_n^*. Tightness of $\{n^{1/4-\delta} c_n^*\}$ under $\{Q_n^n\}$ implies tightness of $\{n^{1/4-\delta} \hat{c}_n\}$. Note that $\hat{c}_n \geqq 2^{-1} n^{-1/4+\delta} b$.

Define the random window w_n by

$$w_n(x; w_1, x_2, \dots, x_n)=m(\hat{c}_n |\psi_{n, \hat\theta_n}(x)|) \tag{3.7}$$

for $x \in \mathscr{X}$. We will write $w_n(x)$ in place of $w_n(x; x_1, x_2, \dots, x_n)$.

Lemma 3. *Suppose the assumptions of Lemma 1 are satisfied. Then*

$$\int [w_n(x)-1]^2 \, dP_{\hat\theta_n} \xrightarrow{Q_n^n} 0 \tag{3.8}$$

and

$$n^{-1/4} \sup_x |\psi_{n, \hat\theta_n}(x) \, w_n(x)| \xrightarrow{Q_n^n} 0. \tag{3.9}$$

Proof. The validity of (3.9) rests upon the inequality

$$n^{-1/4} \sup_x |\psi_{n. \hat\theta_n}(x) \, w_n(x)| \leqq (n^{1/4} \hat{c}_n)^{-1} \sup_{z>0} [z m(z)]. \tag{3.10}$$

On the other hand, (3.8) follows because of Lemma 2 and the inequality

$$\int [w_n(x)-1]^2 \, dP_{\hat\theta_n} \leqq \hat{c}_n^2 \, \|\psi_{n, \hat\theta_n} (dP_{\hat\theta_n})^{1/2}\|^2 \sup_x |m'(x)|^2. \tag{3.11}$$

As will become clearer in Subsect. 3.4, the statistic $\hat{c}_n$ should measure the lack-of-fit between the parametric model $\{P_\theta: \theta \in \Theta\}$ and the actual distribution of the data. The trivial choice, $\hat{c}_n = n^{-1/4+\delta}$ for some $\delta \in (0, 1/4)$, leads to an unsatisfactory asymptotically minimax estimate T_n which is not robust globally. The problem arises because this $\hat{c}_n$ confuses sample size with the amount of contamination present. A more effective $\hat{c}_n$ would be a goodness-of-fit test statistic which is sensitive to fairly arbitrary departures from the parametric model.

For instance, let $\hat{F}_n$ be the empirical c.d.f. of the sample $(x_1, x_2, \dots, x_n)$ and let F_θ be the c.d.f. of P_θ. For some $\alpha > 0$, $\delta \in (0, 1/4)$, let

$$c_n^* = \alpha [\sup_x |\hat{F}_n(x) - F_{\hat\theta_n}(x)|]^{1/2-2\delta}, \tag{3.12}$$

where $\hat\theta_n$ is the discretized initial estimate discussed in Subsect. 3.1.

It is readily verified, in view of (2.4) and the properties of $\{\hat\theta_n\}$, that $\{n^{1/2} \sup_x |\hat{F}_n(x) - F_{\hat\theta_n}(x)|\}$ is tight under $\{Q_n^n\}$; hence $\{n^{1/4-\delta} c_n^*\}$ is also tight under $\{Q_n^n\}$ in this case.

3.4 The Estimates

Let ξ_n be a weighted and re-centered version of the pseudo-derivative $\psi_{n,\hat{\theta}_n}$, defined as follows:

$$\xi_n(x)=[\psi_{n,\hat{\theta}_n}(x)-[\int w_n(t)\,dP_{\hat{\theta}_n}]^{-1}\int \psi_{n,\hat{\theta}_n}(t)\,w_n(t)\,dP_{\hat{\theta}_n}]\,w_n(x). \tag{3.13}$$

Key properties of ξ_n are

$$\int \xi_n(x)\,dP_{\hat{\theta}_n}=0 \tag{3.14}$$

and the two convergences described in the next lemma.

Lemma 4. *Suppose the assumptions of Lemma 1 are satisfied. Let $\{Q_n; n\geqq 1\}$ be any sequence of probabilities on $(\mathscr{X},\mathscr{B})$ such that $\{n^{1/2}[(dQ_n)^{1/2}-(dP_\theta)^{1/2}]; n\geqq 1\}$ converges weakly to an element in H. Then*

$$\|\xi_n(dP_\theta)^{1/2}-\eta_\theta(dP_\theta)^{1/2}\|\xrightarrow{Q_n^n}0 \tag{3.15}$$

and

$$n^{-1/4}\sup_x|\xi_n(x)|\xrightarrow{Q_n^n}0. \tag{3.16}$$

Proof. Because of Lemma 2 and (2.4), $\int \psi_{n,\hat{\theta}_n}\,dP_{\hat{\theta}_n}\xrightarrow{Q_n^n}\int \eta_\theta\,dP_\theta=0$. Using this and (3.8) of Lemma 3 yields

$$|\int \psi_{n,\hat{\theta}_n}w_n\,dP_{\hat{\theta}_n}|\leqq|\int \psi_{n,\hat{\theta}_n}\,dP_{\hat{\theta}_n}|+\|\psi_{n,\hat{\theta}_n}(dP_{\hat{\theta}_n})^{1/2}\|\cdot\|(w_n-1)(dP_{\hat{\theta}_n})^{1/2}\|\xrightarrow{Q_n^n}0. \tag{3.17}$$

Similarly, $\int w_n\,dP_{\hat{\theta}_n}\xrightarrow{Q_n^n}1$. Hence

$$\sup_x|\xi_n(x)-\psi_{n,\hat{\theta}_n}(x)\,w_n(x)|\xrightarrow{Q_n^n}0 \tag{3.18}$$

which, together with (3.9), implies (3.14).

On the other hand,

$$\begin{aligned}\|\xi_n(dP_\theta)^{1/2}-\eta_\theta(dP_\theta)^{1/2}\|\leqq&\|\xi_n(dP_\theta)^{1/2}-\psi_{n,\hat{\theta}_n}w_n(dP_\theta)^{1/2}\|\\&+\|\psi_{n,\hat{\theta}_n}w_n(dP_\theta)^{1/2}-\psi_{n,\hat{\theta}_n}w_n(dP_{\hat{\theta}_n})^{1/2}\|\\&+\|\psi_{n,\hat{\theta}_n}w_n(dP_{\hat{\theta}_n})^{1/2}-\eta_\theta w_n(dP_\theta)^{1/2}\|\\&+\|\eta_\theta w_n(dP_\theta)^{1/2}-\eta_\theta(dP_\theta)^{1/2}\|.\end{aligned} \tag{3.19}$$

Every term on the right side of (3.17) converges in Q_n^n-probability to zero. For the first term, use (3.18). The second term is bounded from above by $n^{-1/2}\sup_x|\psi_{n,\hat{\theta}_n}(x)\,w_n(x)|\cdot[n^{1/2}\,\|(dP_\theta)^{1/2}-(dP_{\hat{\theta}_n})^{1/2}\|]$, the first factor converging in probability to zero by (3.9) while the second factor is tight because of (2.4) and Lemma 1. For the third term, use (3.2) of Lemma 2. Asymptotic negligibility of the final term follows from the bounds, for any $c>0$,

$$\begin{aligned}&\int_{|\eta_\theta|^2>c}[w_n-1]^2|\eta_\theta|^2\,dP_\theta\leqq\int_{|\eta_\theta|^2>c}|\eta_\theta|^2\,dP_\theta\\&\int_{|\eta_\theta|^2\leqq c}[w_n-1]^2|\eta_\theta|^2\,dP_\theta\leqq c\int(w_n-1)^2\,dP_\theta\end{aligned} \tag{3.20}$$

and (3.8), Lemma 1, and (2.4). This completes the proof.

Let $I_n = 4\int \xi_n(x)\,\xi_n'(x)\,dP_{\hat\theta_n}$ and let T_n be the adaptively modified, one step maximum likelihood estimate of θ defined by

$$T_n = \hat\theta_n + 2n^{-1} I_n^{-1} \sum_{i=1}^{n} \xi_n(x_i), \tag{3.21}$$

where $(x_1, x_2, \ldots, x_n)$ is the sample. The next result shows that I_n converges in probability to a nonsingular matrix and that the estimates $\{T_n; n \geqq 1\}$ are asymptotically minimax in that they satisfy (2.14). This proves Theorem 2. In essence, (3.21) is a robust version of the one-step MLE studied by Le Cam (1956), (1969).

Proposition 1. *Suppose* $\{P_\theta: \theta \in \Theta\}$ *satisfies* (2.4), (2.5) *and the mapping* $\theta \rightarrow P_\theta$ *is one-to-one. Let* $\{Q_n; n \geqq 1\}$ *be any sequence of probabilities in* $S_n(\theta, c)$ *such that* $\{n^{1/2}[(dQ_n)^{1/2} - (dP_\theta)^{1/2}]; n \geqq 1\}$ *converges weakly to an element* $\zeta(dR)^{1/2}$ *in* H. *Then, the limiting distribution of* $\{n^{1/2}(T_n - T(\theta, Q_n));\ n \geqq 1\}$ *under* $\{Q_n^n\}$ *is* $N(0, I^{-1}(\theta))$. *Consequently, the* $\{T_n\}$ *satisfy* (2.14) *for every* $c \in (0, 2)$.

Proof. Fix $\theta \in \Theta$ and c. The random vector $n^{-1/2} \sum_{i=1}^{n} \xi_n(x_i)$ can be expressed as the sum of three integrals,

$$\begin{aligned} A_{1n} &= n^{1/2} \int \xi_n(x)\, d(\hat P_n - Q_n) \\ A_{2n} &= n^{1/2} \int \xi_n(x)\, d(Q_n - P_\theta) \\ A_{3n} &= n^{1/2} \int \xi_n(x)\, d(P_\theta - P_{\hat\theta_n}). \end{aligned} \tag{3.22}$$

$\hat P_n$ being the empirical distribution of the sample and the last integral drawing on (3.14).

Let $\theta_n = \theta + n^{-1/2} h$, $c_n = n^{-1/4+\delta} d$ for $h \in R^k$, $d \in R$ and define $\bar\xi_n$ by replacing $(\hat\theta_n, \hat c_n)$ with (θ_n, c_n) in the definition (3.13) of ξ_n. Let $\{a_n; n \geqq 1\}$ be any sequence of positive constants such that $\lim_{n\to\infty} a_n = \infty$, $\lim_{n\to\infty} n^{-1/2} a_n = 0$. Let $D_n = \{x: |\eta_\theta(x)|^2 \leqq a_n\}$. The limiting distribution of $\{n^{1/2} \int_{D_n} \eta_\theta(x)\, d(\hat P_n - Q_n)\}$ under $\{Q_n^n\}$ is evidently $N(0, 4^{-1} I(\theta))$. Both $n^{1/2} \int_{D_n} (\bar\xi_n(x) - \eta_\theta(x))\, d(\hat P_n - Q_n)$ and $n^{1/2} \int_{\bar D_n} \xi_n(x)\, d(\hat P_n - Q_n)$ converge to zero in Q_n^n-probability, by Cebyšev's inequality. Indeed

$$\begin{aligned} \int_{D_n} |\bar\xi_n(x) - \eta_\theta(x)|^2\, dQ_n \leqq & \int_{D_n} |\bar\xi_n(x) - \eta_\theta(x)|^2\, dP_\theta \\ & + [\sup_x |\bar\xi_n(x)|^2 + a_n] \cdot O(n^{-1/2}) \rightarrow 0 \end{aligned} \tag{3.23}$$

as in Lemma 4. A similar argument shows that $\int_{\bar D_n} |\bar\xi_n(x)|^2\, dQ_n \rightarrow 0$. Thus,

$$n^{1/2} \int \bar\xi_n(x)\, d(\hat P_n - Q_n) - n^{1/2} \int_{D_n} \eta_\theta(x)\, d(\hat P_n - Q_n) \xrightarrow{Q_n^n} 0. \tag{3.24}$$

Because of the tightness properties and discreteness of $(\hat\theta_n, \hat c_n)$, (3.24) remains valid when $\bar\xi_n$ is replaced by ξ_n.

We conclude that the limiting distribution of $\{A_{1n}\}$ under $\{Q_n^n\}$ is $N(0, 4^{-1} I(\theta))$.

Also, because of Lemma 4,

$$\begin{aligned}|A_{2n} - 2n^{1/2}\langle \xi_n (dP_\theta)^{1/2}, (dQ_n)^{1/2} - (dP_\theta)^{1/2}\rangle| \\ = n^{1/2} |\langle \xi_n (dQ_n)^{1/2} - \xi_n (dP_\theta)^{1/2}, (dQ_n)^{1/2} - (dP_\theta)^{1/2}\rangle| \\ \leqq n^{-1/2} \sup_x |\xi_n(x)| \cdot [-2 \log(c/2)]^{1/2} \xrightarrow{Q_n^n} 0\end{aligned} \tag{3.25}$$

and, therefore, $A_{2n} \xrightarrow{Q_n^n} 2\langle \eta_\theta (dP_\theta)^{1/2}, \zeta (dR)^{1/2}\rangle$. Similarly,

$$\begin{aligned}|A_{3n} + 2n^{1/2}\langle \xi_n (dP_\theta)^{1/2}, (dP_{\hat\theta_n})^{1/2} - (dP_\theta)^{1/2}\rangle| \\ \leqq n^{-1/2} \sup_x |\xi_n(x)| \cdot [n \,\|(dP_{\hat\theta_n})^{1/2} - (dP_\theta)^{1/2}\|^2] \xrightarrow{Q_n^n} 0\end{aligned} \tag{3.26}$$

by (2.4) and Lemmas 1, 4. Since

$$\begin{aligned}n^{1/2}\langle \xi_n (dP_\theta)^{1/2}, (dP_{\hat\theta_n})^{1/2} - (dP_\theta)^{1/2}\rangle \\ = [\int \xi_n(x)\, \eta'_\theta(x)\, dP_\theta] \cdot n^{1/2}(\hat\theta_n - \theta) + o_p(1) \\ = 4^{-1} I(\theta) \cdot n^{1/2}(\hat\theta_n - \theta) + o_p(1),\end{aligned} \tag{3.27}$$

it follows that $|A_{3n} + 2^{-1} I(\theta) \cdot n^{1/2}(\hat\theta_n - \theta)| \xrightarrow{Q_n^n} 0$.

On the other hand, $I_n \xrightarrow{Q_n^n} I(\theta)$ because, for every $a \in R^k$,

$$\begin{aligned}\|a' \xi_n (dP_{\hat\theta_n})^{1/2} - a' \eta_\theta (dP_\theta)^{1/2}\| \leqq \sup_x |a' \xi_n(x)| \cdot \|(dP_{\hat\theta_n})^{1/2} - (dP_\theta)^{1/2}\| \\ + \|a' \xi_n (dP_\theta)^{1/2} - a' \eta_\theta (dP_\theta)^{1/2}\| \xrightarrow{Q_n^n} 0\end{aligned} \tag{3.28}$$

by (2.4) and Lemmas 1, 4. Substituting the asymptotic approximations for A_{2n}, A_{3n}, and I_n into the definition (3.21) of T_n and recalling the definition (2.8) of $T(\theta, Q)$ yields

$$n^{1/2}(T_n - T(\theta, Q_n)) = 2I^{-1}(\theta)\, A_{1n} + o_p(1). \tag{3.29}$$

Thus, the distributions of $\{n^{1/2}(T_n - T(\theta, Q_n))\}$ converge weakly, under $\{Q_n^n\}$, to a $N(0, I^{-1}(\theta))$ distribution.

To verify that the $\{T_n\}$ satisfy (2.14), we will prove the technically stronger result,

$$\lim_{n\to\infty} \sup_{Q \in S_n(\theta, c)} |R_n(T_n, Q) - R_0(u)| = 0. \tag{3.30}$$

Suppose (3.30) is false. Then, there exists a sequence of probabilities $\{Q_n \in S_n(\theta, c);\ n \geqq 1\}$ such that $R_n(T_n, Q_n)$ does not converge to $R_0(u)$. Note that $Q_n \in S_n(\theta, c)$ entails $n \,\|(dQ_n)^{1/2} - (dP_\theta)^{1/2}\|^2 \leqq -2 \log(c/2)$. By considering subsequences, we may assume without loss of generality that $\{n^{1/2}[(dQ_n)^{1/2} - (dP_\theta)^{1/2}]\}$ converges weakly in H while $R_n(T_n, Q_n)$ does not converge to $R_0(u)$. But u is continuous a.e. (Lebesgue) because it is bounded and monotone increasing. Thus, the first part of Proposition 1 implies convergence of $\{R_n(T_n, Q_n)\}$ to $R_0(u)$. The contradiction proves (3.30).

Remarks. The matrix I_n appearing in the definition (3.21) of T_n can be replaced with $J_n = 4n^{-1} \sum_{i=1}^{n} \xi_n(x_i)\, \xi_n'(x_i)$. Under the assumptions of Proposition 1, $J_n - 4\int \xi_n(x)\, \xi_n'(x)\, dQ_n \xrightarrow{Q_n^n} 0$ because of Lemma 4 and discretization; also $4\int \xi_n(x)\, \xi_n'(x)\, dQ_n \xrightarrow{Q_n^n} I(\theta)$ by an argument similar to (3.28).

Some models $\{P_\theta : \theta \in \Theta\}$ satisfying (2.4) and (2.5) have the additional property that

$$\lim_{t \to 0} \|\eta_{\theta+t}(dP_{\theta+t})^{1/2} - \eta_\theta (dP_\theta)^{1/2}\| = 0. \tag{3.31}$$

In this event, $\psi_{n,\theta}$ may be replaced with η_θ in Lemmas 2, 3 and hence in the definitions of $\xi_n(x)$ and T_n. If, moreover, η_θ is bounded, then the choice $w_n(x) \equiv 1$ suffices to make T_n asymptotically minimax.

When $k \geqq 2$, a different random window may be used to define each component of ξ_n.

Example 2. Suppose Θ is the real line and P_θ is $N(\theta, 1)$. Since $\eta_\theta(x) = 2^{-1}(x-\theta)$, (3.31) holds, and P_θ is symmetric, we may take $\xi_n(x) = 2^{-1}(x - \hat{\theta}_n)\, w_n(x)$. Any discretized M-estimate of θ having the tightness property described in Lemma 1 can be selected as initial estimate $\hat{\theta}_n$. Possible choices of the function m which determines the shape of the random window w_n include: $m(x) = \min\{1, x^{-1}\}$; $m(x) = \max\{(1-x^2)^2, 0\}$; $m(x) = x^{-1} \sin(x)$. The corresponding asymptotically minimax estimates T_n are adaptively scaled versions of the one-step M-estimates associated with the names Huber, Tukey, and Andrews respectively (cf. Andrews et al. (1972)).

Looking at the formula (3.21) for T_n, we can describe the role of $\hat{c}_n$ in this example as follows. If $\hat{c}_n$ is small (i.e. the sample appears normal), $w_n(x)$ remains close to 1 for all moderate x and T_n is, approximately, the average of those observations not too far from the initial estimate $\hat{\theta}_n$; the more distant observations are downweighted. If $\hat{c}_n$ is large (i.e. the sample contains a few extreme outliers, or just appears non-normal), $w_n(x)$ drops to zero quickly as x moves from the origin and T_n disregards all but the observations nearest $\hat{\theta}_n$. Thus, the estimate T_n is very conservatively robust in the presence of detectable contamination and remains cautious even when the sample appears to be normally distributed.

Example 3. Suppose Θ is the real line and $dP_\theta/dx = f(x-\theta)$, where f is absolutely continuous with derivative f' such that $\int [f'(x)]^2/f(x)\, dx$ is finite, nonzero. Then $\{P_\theta : \theta \in \Theta\}$ satisfies (2.4) and (2.5) with $\eta_\theta(x) = -2^{-1} f^{-1}(x-\theta) f'(x-\theta)$ (Hájek (1972)) and the mapping $\theta \to P_\theta$ is one-to-one. Thus, this location model is covered by Theorem 2 and the construction of T_n in Subsect. 3.5.

Example 4. Consider the special case of Example 3 when f is the Cauchy density. Since, for this model, $\eta_\theta(x) = [1 + (x-\theta)^2]^{-1}(x-\theta)$ is bounded and satisfies (3.31), we may take $\xi_n = \eta_{\hat{\theta}_n}$ and $I_n = 2^{-1}$ in defining T_n. The one-step MLE of θ based upon a robust initial estimate $\hat{\theta}_n$ is already asymptotically minimax in the sense of Theorem 2.

Example 5. Let μ be a σ-finite measure on $(\mathscr{X}, \mathscr{B})$ and let $h(x) = (h_1(x), h_2(x), \dots, h_k(x))'$ be a Borel measurable function mapping $\mathscr{X}$ into R^k. Let $\psi(x) = (h_1(x), h_2(x), \dots, h_k(x), 1)$ and suppose the $\{h_j\}$ are such that $\int (a'\psi(x))^2 \, d\mu > 0$ for every nonzero column vector $a \in R^k$. The canonical exponential family $\{P_\theta: \theta \in \Theta\}$ generated by μ and h has densities

$$\frac{dP_\theta}{d\mu}(x) = \exp[\theta' h(x) - b(\theta)], \tag{3.32}$$

where θ is a column vector in R^k, $b(\theta)$ is the normalizing constant, and $\Theta = \operatorname{int}\{\theta \in R^k: \exp[b(\theta)] < \infty\}$. This parametric model satisfies (2.4) and (2.5) with derivative $\eta_\theta(x) = 2^{-1}[h(x) - E(h(x) \mid P_\theta)]$ and information matrix $I(\theta) = \operatorname{Cov}(h(x) \mid P_\theta)$. The mapping $\theta \to P_\theta$ is one-to-one (cf. Berk (1972)). Thus, the assumptions of Theorem 2 are fulfilled.

4. Modified MLE's in Exponential Families

Let $\{P_\theta: \theta \in \Theta\}$ be the canonical exponential family described in the preceding example. It is well-known that the maximum likelihood estimate of θ exists with P_θ^∞-probability one in sufficiently large samples, is unique and asymptotically efficient (cf. Berk (1972)), but may not be robust in the sense of Theorem 2 (Example 1 of Sect. 2). This section describes robust modifications of the MLE which possess some or all of the good properties just cited. In particular, we obtain modified MLE's which are asymptotically minimax in the sense of Theorem 2.

Let $\{v_n(x); n \geqq 1\}$ be a sequence of Borel measurable functions mapping $\mathscr{X}$ into $(0, 1]$, which has the following properties for every $\theta \in \Theta$:

$$n^{-1/4} \sup_x |h(x) \, v_n(x)| \to 0 \tag{4.1}$$

and there exists a Borel measurable function $v: \mathscr{X} \to (0, 1]$ such that

$$\int [v_n(x) - v(x)]^2 \, dP_\theta \to 0. \tag{4.2}$$

Particular choices of $\{v_n\}$ and v will be discussed later in this section.

Let $b_{v(n)}(\theta)$ be the function on Θ determined by the equation

$$\exp[b_{v(n)}(\theta)] = \int \exp(\theta' h(x)) \, v_n(x) \, d\mu.$$

The parametric family of probabilities $\{P_{\theta, v(n)}: \theta \in \Theta\}$ on $(\mathscr{X}, \mathscr{B})$ having densities $dP_{\theta, v(n)}/d\mu = \exp[\theta' h(x) - b_{v(n)}(\theta)] \, v_n(x)$ is canonical exponential. Thus, $b_{v(n)}(\theta)$ is strictly convex with

$$\begin{gathered} \nabla b_{v(n)}(\theta) = E[h(x) \mid P_{\theta, v(n)}] = [\int v_n(x) \, dP_\theta]^{-1} \int h(x) \, v_n(x) \, dP_\theta \\ \nabla^2 b_{v(n)}(\theta) = \operatorname{Cov}(h(x) \mid P_{\theta, v(n)}), \end{gathered} \tag{4.3}$$

the covariance matrix being positive definite and continuous in θ. Replacing v_n

by v in these formulae yields another exponential family $\{P_{\theta,v}: \theta\in\Theta\}$ and a function $b_v(\theta)$ with parallel properties.

Given a sample $(x_1, x_2, \ldots, x_n)$, define the estimate T_n^* as the unique value of $s\in\Theta$ which maximizes the strictly concave function

$$M_n(s)=n\left[\sum_{i=1}^{n} v_n(x_i)\right]^{-1} s' \sum_{i=1}^{n} h(x_i)\, v_n(x_i)-nb_{v(n)}(s) \tag{4.4}$$

if such a maximizing value exists; otherwise let $T_n^*=\theta_0$, an arbitrary element of Θ. Equivalently, T_n^* is the unique solution to the equation

$$[\int v_n(x)\,dP_s]^{-1} \int h(x)\, v_n(x)\,dP_s=\left[\sum_{i=1}^{n} v_n(x_i)\right]^{-1} \sum_{i=1}^{n} h(x_i)\, v_n(x_i) \tag{4.5}$$

if such exists and is θ_0 otherwise. When $v_n(x)\equiv 1$, T_n^* reduces to the maximum likelihood estimate of θ.

Let $D_v(\theta)$ be the $k\times(k+1)$ partitioned matrix

$$[\int v(x)\,dP_\theta]^{-1}(I_k\colon -[\int v(x)\,dP_\theta]^{-1} \int h(x)\, v(x)\,dP_\theta),$$

where I_k is the $k\times k$ identity matrix. Let $A_v(\theta)=[\nabla^2 b_v(\theta)]^{-1} D_v(\theta)$.

Proposition 2. *Let $\{Q_n; n\geqq 1\}$ be any sequence of probabilities in $S_n(\theta, c)$ such that $\{n^{1/2}[(dQ_n)^{1/2}-(dP_\theta)^{1/2}]; n\geqq 1\}$ converges weakly to an element $\zeta(dR)^{1/2}$ in H. Suppose the windows $\{v_n; n\geqq 1\}$, v satisfy* (4.1) *and* (4.2). *Then, the limiting distribution of $\{n^{1/2}(T_n^*-\theta); n\geqq 1\}$ under Q_n^n is $N(\mu_v, \Sigma_v)$, with*

$$\begin{aligned}\mu_v&=2A_v(\theta)\langle\psi v(dP_\theta)^{1/2}, \zeta(dR)^{1/2}\rangle\\ \Sigma_v&=A_v(\theta)\operatorname{Cov}(\psi(x)\, v(x)\,|\,P_\theta)\, A_v'(\theta).\end{aligned} \tag{4.6}$$

For the canonical exponential family under discussion, $\eta_\theta(x)=2^{-1}[h(x)-E(h(x)\,|\,P_\theta)]$ and $I(\theta)=\operatorname{Cov}(h(x)\,|\,P_\theta)$. The weak convergence of $\{n^{1/2}[(dQ_n)^{1/2}-(dP_\theta)^{1/2}]\}$ to $\zeta(dR)^{1/2}$ entails orthogonality in H of $\zeta(dR)^{1/2}$ and $(dP_\theta)^{1/2}$. When $v(x)\equiv 1$, Eq. (4.3) and this orthogonality permit the simplifications

$$\begin{aligned}\mu_v&=2I^{-1}(\theta)\langle h(dP_\theta)^{1/2}, \zeta(dR)^{1/2}\rangle\\ &=4I^{-1}(\theta)\langle\eta_\theta(dP_\theta)^{1/2}, \zeta(dR)^{1/2}\rangle\\ \Sigma_v&=I^{-1}(\theta).\end{aligned} \tag{4.7}$$

Consequently, by the argument used in Proposition 1, the estimates T_n^* satisfy (2.14) for every $c\in(0, 2)$.

Proof. Fix $\theta\in\Theta$. For $t\in R^k$, let

$$L_n(t)=\begin{cases}M_n(\theta+n^{-1/2}t)-M_n(\theta) & \text{if } \theta+n^{1/2}t\in\Theta\\ -\infty & \text{otherwise.}\end{cases} \tag{4.8}$$

Let $t_{M,n}$ denote the unique value of t which maximizes $L_n(t)$, if it exists. The estimate T_n^* equals $\theta+n^{-1/2}t_{M,n}$ in that event. Let $\hat{P}_n$ be the empirical distribution of the sample. For $\theta+n^{-1/2}t\in\Theta$, it follows from (4.4) and (4.3) that

$$\begin{aligned} L_n(t) &= n^{1/2}[\int v_n(x)\, d\hat{P}_n]^{-1}\, t' \int h(x)\, v_n(x)\, d\hat{P}_n \\ &\quad - n[b_{v(n)}(\theta+n^{-1/2}t) - b_{v(n)}(\theta)] \\ &= t' Z_n - n[b_{v(n)}(\theta+n^{-1/2}t) - b_{v(n)}(\theta) - n^{-1/2}\, t'\, \nabla b_{v(n)}(\theta)] \end{aligned} \tag{4.9}$$

where $Z_n = n^{1/2} D_n \int \psi(x)\, v_n(x)\, d(\hat{P}_n - P_\theta)$ and D_n is the $k\times(k+1)$ partitioned matrix $[\int v_n(x)\, dP_\theta]^{-1}(I_k\colon -[\int v_n(x)\, d\hat{P}_n]^{-1} \int h(x)\, v_n(x)\, d\hat{P}_n)$.

The first stage of the proof is to show that the limiting distribution of $\{Z_n\}$ under $\{Q_n^n\}$ is normal with mean $2\, D_v(\theta)\langle \psi v (dP_\theta)^{1/2}, \zeta (dR)^{1/2}\rangle$ and covariance matrix $D_v(\theta)\, \mathrm{Cov}(\psi(x)\, v(x) \,|\, P_\theta)\, D_v'(\theta)$. Indeed, this is a consequence of three facts:

$$n^{1/2} \int \psi(x)\, v_n(x)\, d(\hat{P}_n - Q_n) \xrightarrow{Q_n^n} N(0, \mathrm{Cov}(\psi(x)\, v(x) \,|\, P_\theta)) \tag{4.10}$$

and

$$n^{1/2} \int \psi(x)\, v_n(x)\, d(Q_n - P_\theta) \to 2\langle \psi v (dP_\theta)^{1/2}, \zeta (dR)^{1/2}\rangle \tag{4.11}$$

and $D_n \xrightarrow{Q_n^n} D_v(\theta)$.

To verify (4.10), let $\{a_n;\, n \geqq 1\}$ be any sequence of positive constants such that $\lim_{n\to\infty} a_n = \infty$, $\lim_{n\to\infty} n^{-1/2} a_n = 0$ and let $B_n = \{x\colon |\psi(x)|^2 \leqq a_n\}$. Then

$$\begin{aligned} n^{1/2} \int_{B_n} \psi(x)\, v_n(x)\, d(\hat{P}_n - Q_n) &= n^{1/2} \int_{B_n} \psi(x)\, v(x)\, d(\hat{P}_n - Q_n) \\ &\quad + n^{1/2} \int_{B_n} \psi(x)[v_n(x) - v(x)]\, d(\hat{P}_n - Q_n). \end{aligned} \tag{4.12}$$

The limiting distribution of the first integral on the right side of (4.12) is $N(0, \mathrm{Cov}(\psi(x)\, v(x) \,|\, P_\theta))$, because $\lim_{n\to\infty} \mathrm{Cov}[n^{1/2} \int_{B_n} \psi(x)\, v(x)\, d(\hat{P}_n - Q_n)] = \mathrm{Cov}(\psi(x)\, v(x) \,|\, P_\theta)$ by the choice of $\{a_n\}$. On the other hand, the second integral on the right converges in Q_n^n-probability to zero because

$$\begin{aligned} \int_{B_n} |\psi(x)|^2 [v_n(x) - v(x)]^2\, dQ_n &\leqq a_n \int [v_n(x) - v(x)]^2\, dP_\theta \\ &\quad + O(n^{-1/2}) \cdot [\sup_x |\psi(x)\, v_n(x)|^2 + a_n] \to 0 \end{aligned} \tag{4.13}$$

in view of (4.1) and (4.2). Moreover, $n^{1/2} \int_{\bar{B}_n} \psi(x)\, v_n(x)\, d(\hat{P}_n - Q_n) \xrightarrow{Q_n^n} 0$ because

$$\int_{\bar{B}_n} |\psi(x)\, v_n(x)|^2\, dQ_n \leqq \int_{\bar{B}_n} |\psi(x)|^2\, dP_\theta + O(n^{-1/2}) \sup_x |\psi(x)\, v_n(x)|^2 \to 0. \tag{4.14}$$

The calculations of this paragraph imply (4.10).

Justification of (4.11) rests on the inequalities

$$\begin{aligned} &|n^{1/2} \int \psi(x)\, v_n(x)\, d(Q_n - P_\theta) - 2\langle \psi v_n (dP_\theta)^{1/2}, n^{1/2}[(dQ_n)^{1/2} - (dP_\theta)^{1/2}]\rangle| \\ &\quad \leqq \sup_x |\psi(x)\, v_n(x)| \cdot n^{1/2}\, \|(dQ_n)^{1/2} - (dP_\theta)^{1/2}\|^2 \end{aligned} \tag{4.15}$$

and

$$\begin{aligned} &|\langle \psi(v_n - v)(dP_\theta)^{1/2}, n^{1/2}[(dQ_n)^{1/2} - (dP_\theta)^{1/2}]\rangle| \\ &\quad \leqq \|\psi(v_n - v)(dP_\theta)^{1/2}\| \cdot [-2\log(c/2)]^{1/2}, \end{aligned} \tag{4.16}$$

since $Q_n \in S_n(\theta, c)$. Both bounds converge to zero under the assumptions on $\{v_n\}$ and $\{Q_n\}$. Finally, combining (4.10), (4.11) with (4.2) yields $D_n \xrightarrow{Q_n^n} D_v(\theta)$.

The second stage of the proof considers the asymptotic behavior of $t_{M,n}$, the value of $t \in R^k$ which maximizes $L_n(t)$. By Skorohod's theorem, there exist versions of the random variables $\{Z_n\}$ and a random variable Z such that $Z_n \xrightarrow{\text{a.s.}} Z$ and Z has a normal distribution with mean $2D_v(\theta)\langle \psi v(dP_\theta)^{1/2}, \zeta(dR)^{1/2}\rangle$ and covariance matrix $D_v(\theta)\operatorname{Cov}(\psi(x)\,v(x)\,|\,P_\theta)\,D_v'(\theta)$. For $t \in R^k$, let

$$L(t) = t'Z - 2^{-1}\,t'\,\nabla^2 b_v(\theta)\,t. \tag{4.17}$$

We will show that for these random variables $\{Z_n\}$, Z and for every $a > 0$,

$$\sup_{|t| \leqq a} |L_n(t) - L(t)| \xrightarrow{\text{a.s.}} 0. \tag{4.18}$$

Since the value of t which maximizes $L(t)$ is $t_M = [\nabla^2 b_v(\theta)]^{-1} Z$ and since $L_n(t)$ is strictly concave or $-\infty$, it follows from (4.18) that $t_{M,n}$ exists a.s. if n is sufficiently large and $t_{M,n} \xrightarrow{\text{a.s.}} t_M$ as $n \to \infty$. Thus, $n^{1/2}(T_n - \theta) \xrightarrow{\text{a.s.}} t_M$, which implies the proposition.

The validity of (4.18) is clear from the definitions of $L_n(t)$ and $L(t)$, once it has been shown that

$$\sup_{|t| \leqq a} \nabla^2 b_{v(n)}(\theta + n^{-1/2}t) - \nabla^2 b_v(\theta)| \to 0 \tag{4.19}$$

for every $a > 0$. Since $\nabla^2 b_v(\theta)$ is continuous in θ, it suffices to check that

$$\sup_{|t| \leqq a} |\nabla^2 b_{v(n)}(\theta + n^{-1/2}t) - \nabla^2 b_v(\theta + n^{-1/2}t)| \to 0. \tag{4.20}$$

The last convergence can be verified by calculation, using (4.2), (2.4), (4.3) and the definitions of $P_{\theta,v}$ and $P_{\theta,v(n)}$.

Remarks. When $|h(x)|$ is bounded, the choices $v_n(x) = v(x) \equiv 1$ satisfy (4.1) and (4.2). Hence, the MLE of θ is asymptotically minimax under these circumstances.

For general $h(x)$, an initial robust estimate of θ can be obtained as follows. Let $v_n(x) = v(x) = q(|h(x) - \beta|)$, where β is a vector constant, q is a Borel measurable function mapping R^+ into $(0, 1]$, such that $q(0) = 1$ and $\sup_{x>0}[xq(x)] < \infty$. Ideally, β should approximate the unknown value of $E(h(x)\,|\,P_\theta)$. Let θ_n^* be the estimate of θ determined by the modified ML equations (4.5) corresponding to this choice of v_n. Since Proposition 2 is applicable, the estimates $\{\theta_n^*\}$ have the tightness property described in Lemma 1.

Once a suitable initial estimate θ_n^* has been found, an asymptotically minimax estimate can be constructed by the method of Sect. 3 and Example 5. Alternatively, we may solve the modified ML equations (4.5) with $v_n(x) \equiv w_n(x)$, the discretized random window defined in Subsect. 3.3. Proposition 2 can be extended to this random v_n, with $v(x) \equiv 1$. In view of (4.7), this modified ML estimate is asymptotically minimax, by the argument used in Proposition 1.

Example 6. For $r \geqq 2$, let $S_r = \{x \in R^r\colon |x| = 1\}$ and let μ be the rotation invariant measure on S_r, normalized so that $\mu(S_r) = 1$. Relative to μ, the Fisher-von Mises distribution has density proportional to $\exp(\kappa v'x)$, where $\kappa \geqq 0$ and $v \in S_r$. This model can be reparametrized into the form

$$\frac{dP_\theta}{d\mu}(x) = \exp[\theta' x - b(\theta)], \qquad x \in S_r, \tag{4.21}$$

with $\theta \in \Theta = R^r$. Since $|x| = 1$, the MLE of θ is already asymptotically minimax.

Example 7. Let $e = (1, 1, \ldots, 1)' \in R^r$, $r \geqq 2$. The submodel of $N(\mu, \Sigma)$ which corresponds to the one-way layout with random effects specifies that $\mu = \alpha e$ and $\sum = \sigma^2 [(1-\rho) I_r + \rho e e']$, with α real, $\sigma^2 > 0$, $0 < \rho < 1$. Reparametrization gives the equivalent form

$$\frac{dP_\theta}{dx} = \exp[\theta_1 |x|^2 + \theta_2 (e' x)^2 + \theta_3 (e' x) - b(\theta)] \tag{4.22}$$

for $x \in R^r$, $\theta_1 < 0$, $0 < \theta_2 < -r^{-1} \theta_1$, and θ_3 real. Since these constraints define an open subset of the natural parameter space, the estimation theory of this section is applicable.

Acknowledgement. Use of the simplified risk (2.9) was suggested by conversations with P.W. Millar. I am grateful for his interest and comments.

References

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., Tukey, J.W.: Robust Estimates of Location: Survey and Advances. Princeton: University Press 1972

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

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

Beran, R.J.: Asymptotic lower bounds for risk in robust estimation. Ann. Statist. **8**, 1252–1264 (1980)

Berk, R.H.: Consistency and asymptotic normality of MLE's for exponential models. Ann. Math. Statist. **43**, 193–204 (1972)

Chernoff, H.: Large sample theory: parametric case. Ann. Math. Statist. **27**, 1–22 (1956)

Dvoretzky, A., Kiefer, J., Wolfowitz, J.: Asymptotic minimax character of the sample distribution function and of the classical multinominal estimator. Ann. Math. Statist. **27**, 642–669 (1956)

Hâjek, J.: Local asymptotic minimax and admissibility in estimation. Proc. Sixth Berkeley Sympos. Math. Statist. Probab. **1**, 175–194. University of California Press (1972)

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

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

Le Cam, L.: On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. Univ. of California Publications in Statistics **1**, 277–330 (1953)

Le Cam, L.: On the asymptotic theory of estimation and testing hypotheses. Proc. Third Berkeley Sympos. Math. Statist. Probab. **1**, 129–156. University of California Press (1956)

Le Cam, L.: Théorie Asymptotique de la Décision Statistique. Les Presses de l'Université de Montréal 1969

Le Cam, L.: Limits of experiments. Proc. Sixth Berkeley Sympos. Math. Statist. Probab. **1**, 245–261. University of California Press (1972)

Millar, P.W.: Asymptotic minimax theorems for the sample distribution function. Z. Wahrscheinlichkeitstheorie und verw. Gebiete **48**, 233–252 (1979)

Moussatat, M.W.: On the asymptotic theory of statistical experiments and some of its applications. Ph.D dissertation, University of California, Berkeley 1976

Neveu, J.: Mathematical Foundations of the Calculus of Probability. San Francisco: Holden-Day 1969

Received October 1, 1979; in revised form July 2, 1980