KYBERNETIKA — VOLUME 31 (1995), NUMBER 3, PAGES 221 – 237

THE ROLE OF HÁJEK'S CONVOLUTION THEOREM IN STATISTICAL THEORY

RUDOLF BERAN

Hájek's [17] convolution theorem was a major advance in understanding the classical information inequality. This re-examination of the convolution theorem discusses historical background to asymptotic estimation theory; the role of superefficiency in current estimation practice; the link between convergence of bootstrap distributions and convolution structure; and a dimensional asymptotics view of superefficiency.

1. INTRODUCTION

In 1970, Hájek established sharp, general criteria for asymptotically efficient estimation in locally asymptotically normal parametric models. His *Zeitschrift* paper that year characterized, through the convolution theorem, the structure of possible limiting distributions for regular estimators of a parameter. His talk that summer at the Sixth Berkeley Symposium formulated the local asymptotic minimax bound for all estimators of a parameter. Both of Hájek's papers built on a pre-history; and both stimulated important further work by LeCam, by the Russian school, and by others. A quarter of a century later, the convolution theorem sheds light on topics such as bootstrap consistency, model selection, and signal recovery. This paper describes how.

Fisher's Program. As a model for the sample $X_n = (X_{n,1}, \ldots, X_{n,n})$, suppose that the $\{X_{n,i}\}$ are iid with distribution P_θ. The value of θ is unknown but lies in Θ, an open subset of R^k. The distribution P_θ has a density p_θ with respect to a dominating σ-finite measure that does not depend on θ. Suppose that the gradient ∇p_θ of p_θ with respect to θ exists. In this classical setting, the information matrix is defined to be

$$I(\theta) = \mathrm{Cov}_\theta[p_\theta^{-1}(X_n)\nabla p_\theta(X_n)], \tag{1}$$

provided the covariances on the right side exist. Savage [47 p. 456] gave historical background for this information concept.

Consider the problem of estimating a differentiable parametric function $\tau(\theta)$. For simplicity of exposition, suppose that $\tau(\theta) = \theta$. Let T_n be an unbiased estimator of

θ and let $|\cdot|$ denote Euclidean norm in R^k. The information inequality, formulated by Fréchet [16], Darmois [12], Rao [44], and Cramér [11], implies that

$$nE_\theta|T_n - E_\theta T_n|^2 \geq \operatorname{tr}[I^{-1}(\theta)] \qquad \forall\, \theta \in \Theta. \tag{2}$$

This variance inequality assumes that the information matrix is nonsingular and that the model meets certain other regularity conditions (cf. Lehmann [36, p. 128]).

Two other strong notions existed in estimation theory earlier this century. First was the idea that a good real-valued estimator has a bias that is much smaller than its variance when the sample size is large. Second was the belief that the maximum likelihood estimator $T_{n,ML}$ of θ is asymptotically normal and that the limiting distribution of $n^{1/2}(T_{n,ML} - \theta)$ is $N(0, I^{-1}(\theta))$. Combining these ideas with inequality (1.2) leads to two conjectures:

A. For any estimator T_n,

$$\liminf_{n \longrightarrow \infty} nE_\theta|T_n - \theta|^2 \geq \operatorname{tr}[I^{-1}(\theta)] \qquad \forall\, \theta \in \Theta. \tag{3}$$

B. The maximum likelihood estimator $T_{n,ML}$ satisfies

$$\lim_{n \to \infty} nE_\theta|T_{n,ML} - \theta|^2 = \operatorname{tr}[I^{-1}(\theta)] \qquad \forall\, \theta \in \Theta. \tag{4}$$

This pair of statements is sometimes called *Fisher's program*, in recognition of Fisher's [15] influential paper on estimation in parametric models. The program implies the conjecture that maximum likelihood estimators are asymptotically efficient, in the sense of attaining the asymptotic lower bound in A at every θ. Pratt [43] drew attention to related previous work by Edgeworth and others. Pfanzagl [39, pp. 207–208] summarized the history of early work on maximum likelihood estimation, from Laplace and Gauss onwards.

Superefficiency and Other Surprises. As stated, both conjectures A and B are false. Possible difficulties with uniform integrability in B may be resolved by a continuous, monotone, bounded transformation of the loss function. Deeper, however, is the possibly bad behavior of maximum likelihood estimators in regular parametric models.

Example 1. Suppose that the distribution of $\log(X_i - \gamma)$ is $N(\mu, \sigma^2)$. This is a model for the time at which disease symptoms are first observed in a patient who was exposed to infection at time γ. Here the unknown parameter θ is the triple (μ, σ^2, γ). The distribution of X_i is called the three-parameter lognormal distribution. The information matrix for this model is finite and is continuous in θ. However, it was not noticed for many years that the likelihood function climbs a ridge to infinity as γ tends to the smallest observation (Hill [20]). While maximum likelihood estimation thus fails, the model has the LAN property to be discussed in Section 2. Consequently, LeCam's [29, pp. 138–139] one-step estimator achieves the asymptotic efficiency that eludes maximum likelihood in this example.

Deeper still is the possible failure of inequality (1.3) at some values of θ, called points of superefficiency. LeCam [28] and Bahadur [1] showed that superefficiency points necessarily constitute a Lebesgue null set in R^k. While "null set" sounds innocuous, it is not, as we will see in the next two examples.

Example 2. Suppose that the $\{X_{n,i}\}$ are iid random variables, each distributed according to $N(\theta, 1)$. Here the parameter dimension $k = 1$. Let $\bar{X}_n$ to be the sample mean and let $T_{n,H}$ be the Hodges estimator of θ, given by

$$T_{n,H} = \begin{cases} b\bar{X}_n & \text{if } |\bar{X}_n| \leq n^{-1/4} \\ \bar{X}_n & \text{otherwise,} \end{cases} \tag{5}$$

where $b^2 < 1$ (LeCam [28]). Note that, when b is zero, $T_{n,H}$ is a model selection estimator that chooses between fitting the $N(0,1)$ and $N(\theta, 1)$ models on the basis of the data.

The limiting distribution of $n^{1/2}(T_{n,H} - \theta)$ is $N(0,1)$ when $\theta \neq 0$ but is $N(0, b^2)$ when $\theta = 0$. Moreover,

$$\lim_{n\to\infty} nE_\theta(T_{n,H} - \theta)^2 = \begin{cases} b^2 & \text{if } \theta = 0 \\ 1 & \text{if } \theta \neq 0 \end{cases} \tag{6}$$

while the Fisher information bound is 1. Thus, the origin is a point of superefficiency. For fixed n, the risk of $T_{n,H}$ is less than 1 in a neighborhood of the origin, then rises steeply above one, and subsequently drops slowly towards 1 as $|\theta|$ tends to infinity (cf. Lehmann [36, Chapter 6]). The neighborhood of improved risk narrows as n increases, so that the asymptotic picture is (1.6). At finite n, the Hodges estimator has larger risk than the sample mean for most values of θ. Such poor risk near points of superefficiency is characteristic of one-dimensional estimators (LeCam [28], Hájek [18]).

From this example, one might form the impression that model selection estimators are to be avoided. This impression is wrong. Consider model selection estimators for $\theta \in R^k$ that, as in Pötscher [42], select among submodels indexed by proper subspaces of R^k. Under asymptotics where parameter dimension k is fixed while n increases, the points of superefficiency are the union of proper subspaces of Θ. Though uncountable, these superefficiency points form a Lebesgue null set. However, under asymptotics where k increases while n is fixed, model selection can improve risk over the entire parameter space (Beran [5]). The next example makes the role of dimension clearer.

Example 3. Let I_k denote the $k \times k$ identity matrix. Suppose that the $\{X_{n,i}\}$ are iid random k-vectors, each distributed according to $N_k(\theta, I_k)$, where $\theta \in R^k$. This is a simple model for n repeated observations on a discrete time series measured at k time points. The goal is to estimate the unknown signal θ. Let $\bar{X}_n$ be the sample mean and suppose that k is at least 3. The James and Stein [23] estimator is

$$T_{n,S} = \left[1 - (k-2)/|n\bar{X}_n|^2\right] \bar{X}_n. \tag{7}$$

This celebrated estimator is a sharp early example of what are now called regularization methods for signal recovery (cf. Titterington [51]). For $k \geq 3$, $\lim_{n\to\infty} nE_\theta|T_{n,S} - \theta|^2$ equals the information bound k when $\theta \neq 0$ but is strictly smaller when $\theta = 0$. Thus, the origin is a point of superefficiency under quadratic loss.

Unlike Example 2, the risk $nE_\theta|T_{n,S} - \theta|^2$ is uniformly smaller than k for all values of θ and n (James and Stein [23]). Moreover, as $k \to \infty$ and $k^{-1}n|\theta|^2 \to c$, the normalized quadratic risk $k^{-1}nE_\theta|T_{n,S} - \theta|^2$ converges to $c/(1+c)$. For details, see Casella and Hwang [9]. Consequently, when k is large and n is fixed, the estimator $T_{n,S}$ improves substantially upon the sample mean $\bar{X}_n$ over compact balls about $\theta = 0$. The superefficiency at $\theta = 0$ that is detected by asymptotics in n alone is a ghost of what actually happens for finite k and n.

As an extension of this example, consider the Stein estimator that shrinks each component of $\bar{X}_n = (\bar{X}_{n,1}, \ldots, \bar{X}_{n,n})$ towards the average component $\hat{m}_n = n^{-1}\sum_{i=1}^n \bar{X}_{n,i}$. Let e denote the vector in R^k whose components each equal 1. Define

$$T_{n,SM} = \hat{m}_n e + [1 - (k-3)/(n|\bar{X}_n - \hat{m}_n e|^2)](\bar{X}_n - \hat{m}_n e). \tag{8}$$

This estimator is superefficient at every $\theta \in R^k$ whose components are equal, an uncountable Lebesgue null set. Moreover, the estimator $T_{n,SM}$ dominates $\bar{X}_n$ over the entire parameter space, substantially so when k is much larger than n (cf. Lehmann [36, p. 305]).

As these examples suggest, the possibility of superefficiency is at heart of modern estimation theory. Current signal estimators or model selection estimators tacitly create points of superefficiency, though usually without articulating this strategy; and they do so because superefficiency points can reduce risk over the entire parameter space when k and n are finite. That points of superefficiency form a Lebesgue null set does not make them unimportant.

2. HÁJEK'S CONVOLUTION THEOREM

For what class of parametric models is Fisher's program pertinent? LeCam [30] made a fundamental advance on this question by formulating the concept of a locally asymptotically normal model. For simplicity, we will assume that the parameter space Θ is an open subset of R^k and that the rate of convergence is $n^{1/2}$, as in classical iid models.

Definition. For $\theta_n = \theta_0 + n^{-1/2}h$, where $h \in R^k$, let $P^c_{\theta_n,n}$ denote the absolutely continuous part of $P_{\theta_n,n}$ with respect to $P_{\theta_0,n}$. Let $L_n(h,\theta_0)$ denote the log-likelihood ratio of $P^c_{\theta_n,n}$ with respect to $P_{\theta_0,n}$. Suppose there exist random vectors $Y_n(\theta_0)$, depending on the sample as well as on θ_0, and a nonsingular, nonrandom matrix $I(\theta_0)$ such that

$$L_n(h,\theta_0) = h'Y_n(\theta_0) - 2^{-1}h'I(\theta_0)h + o_p(1), \tag{9}$$

the remainder term tending to zero in $P_{\theta_0,n}$-probability. Suppose in addition that

$$\mathcal{L}[Y_n(\theta_0)|P_{\theta_0,n}] \Longrightarrow N(0, I(\theta_0)). \tag{10}$$

Then the model $\{P_{\theta,n}\colon \theta \in \Theta\}$ is said to be *locally asymptotically normal* (LAN) at θ_0.

The LAN property, which redefines the information matrix, is possessed by classical models such as smooth exponential families. Hájek and Šidák [19] and Hájek [18] included convenient sufficient conditions for LAN that developed LeCam's earlier work. For an LAN model, the log-likelihood ratio behaves asymptotically like the log-likelihood ratio of $N(h, I^{-1}(\theta_0))$ with respect to $N(0, I^{-1}(\theta_0))$. This suggests that good statistical procedures in the normal limit experiment may have counterparts that are approximately good, for large n, in the model $P_{\theta,n}$. Hájek's papers on the convolution theorem [17] and on the local asymptotic minimax bound [18] gave substance to this idea.

The summary paragraph at the start of Hájek [17] states: "Under certain very general conditions we prove that the limiting distribution of the estimates, if properly normed, is a convolution of a certain normal distribution, which depends only of the underlying distributions, and of a further distribution, which depends on the choice of the estimate. As corollaries we obtain inequalities for asymptotic variance and for asymptotic probabilities of certain sets, generalizing to some results of J. Wolfowitz [54], S. Kaufman [26], L. Schmetterer [48] and G. G. Roussas [45]." This describes the content and historical setting with admirable succinctness.

Let us consider now what Hájek did in this paper, specializing for convenience to the case where the rate of convergence is $n^{1/2}$. For any estimator T_n of θ, let

$$H_n(\theta) = \mathcal{L}[n^{1/2}(T_n - \theta)|P_{\theta,n}]. \tag{11}$$

Definition. Let $\theta_n = \theta_0 + n^{-1/2}h$, where $h \in R^k$. A sequence of estimators $\{T_n\colon n \geq 1\}$ is *regular* at θ_0 if

$$H_n(\theta_n) \Longrightarrow H(\theta_0) \qquad \forall\, h \in R^k \tag{12}$$

for some limit distribution $H(\theta_0)$ that does not depend on h.

Suppose that d is a metric for weak convergence of distributions on R^k. A little stronger than regularity is the property

$$\lim_{n\to\infty} \sup_{n^{1/2}|\theta-\theta_0|\leq c} d[H_n(\theta), H(\theta_0)] = 0 \tag{13}$$

for every finite positive c. Prior to Hájek [17], papers on asymptotic estimation typically imposed a requirement such as uniform weak convergence of the distributions $\{H_n(\theta)\}$ to a weakly continuous limit over some fixed neighborhood of θ_0 — an assumption considerably stronger than (2.5) or (2.4). Such was also the case for Inagaki's [22] independent discovery of the convolution theorem. We will see, in the course of this paper, how Hájek's weaker regularity assumption is important for statistical theory.

Theorem 2.1. (Convolution theorem) Suppose that the model $P_{\theta,n}$ is LAN at θ_0. If $\{T_n : n \geq 1\}$ is a sequence of estimators regular at θ_0, then

$$H(\theta_0) = N(0, I^{-1}(\theta_0)) * \nu(\theta_0) \tag{14}$$

for some distribution $\nu(\theta_0)$ on R^k. Moreover $H(\theta_0) = N(0, I^{-1}(\theta_0))$ if and only if

$$n^{1/2}(T_n - \theta_0) = I^{-1}(\theta_0)Y_n(\theta_0) + o_p(1), \tag{15}$$

the remainder term tending to zero in $P_{\theta_0,n}$-probability.

Not long after Hájek's announcement of the convolution theorem, P. J. Bickel sent him a letter that sketched a shorter characteristic function proof for the result. This writer saw a copy of the letter a few months later. Even though the characteristic function argument is now well-known, we give a version here because the argument leads to further insights.

Proof. Fix $h \in R^k$. The LAN assumption implies that the sequences $\{P_{\theta_n,n}\}$ and $\{P_{\theta_0,n}\}$ are contiguous (LeCam [30]). In particular, the total variation norm of the singular component of $P_{\theta_n,n}$ relative to $P_{\theta_0,n}$ tends to zero as n increases. Let $\phi_n(u,\theta)$ and $\phi(u,\theta)$ denote, respectively, the characteristic functions of $H_n(\theta)$ and $H(\theta)$. Then

$$\phi_n(u,\theta_n) = E_{\theta_0}[iu'n^{1/2}(T_n - \theta_0) - iu'h + L_n(h,\theta_0)] + o(1). \tag{16}$$

Because of (2.4) and (2.2), by going to a subsequence we can assume, without loss of generality, that

$$(n^{1/2}(T_n - \theta_0), Y_n(\theta_0)) \Longrightarrow (S, I^{1/2}(\theta_0)Z) \tag{17}$$

under $P_{\theta_0,n}$. Here S has marginal distribution $H(\theta_0)$ while Z has a standard normal distribution on R^k. Let $n \to \infty$ in (2.8). Then (2.9) and a uniform integrability argument establish

$$\phi(u,\theta_0) = E\exp[iu'S - iu'h]\exp[h'I^{1/2}(\theta_0)Z - 2^{-1}h'I(\theta_0)h]. \tag{18}$$

Equation (2.10) holds for every $h \in R^k$. Since the right side of (2.10) is analytic in h while the left side does not depend on h, equation (2.10) continues to hold for all $h \in C^k$. Setting $h = -iI^{-1}(\theta_0)u$ in (2.10) yields

$$\phi(u,\theta_0) = E\exp[iu'(S - I^{-1/2}(\theta_0)Z)]\exp[-2^{-1}u'I^{-1}(\theta_0)u], \tag{19}$$

which is equivalent to assertion (2.6).

The if and only if part: Suppose that (2.7) holds. By the LAN property and contiguity reasoning,

$$\mathcal{L}[Y_n(\theta_0)|P_{\theta_n,n}] \Longrightarrow N(I(\theta_0)h, I(\theta_0)). \tag{20}$$

This and (2.7) imply that $H_n(\theta_n) \Longrightarrow N(0, I^{-1}(\theta_0))$, as asserted by the theorem.

Conversely, suppose that $H(\theta_0) = N(0, I^{-1}(\theta_0))$ but the approximation (2.7) does not hold. By going to a subsequence, assume without loss of generality that

$$P_{\theta_0,n}[|n^{1/2}(T_n - \theta_0) - I^{-1}(\theta_0)Y_n(\theta_0)| \geq \epsilon] > \delta \tag{21}$$

for every n and some positive ϵ and δ. By going to a further subsequence, as in the first part of the proof, assume without loss of generality that (2.9) holds. From this and (2.13),

$$P[|S - I^{-1/2}(\theta_0)Z| \geq \epsilon] > \delta. \tag{22}$$

At the same time, (2.11) also holds and implies that

$$S = I^{-1/2}(\theta_0)Z \qquad \text{w.p.1.} \tag{23}$$

The contradiction between (2.14) and (2.15) establishes (2.7). □

The convolution theorem supports a portion of Fisher's program. Suppose that w is a symmetric, subconvex, and continuous loss function on R^k. If $\{T_n\}$ is any sequence of estimators whose limiting distribution at θ_0 has the convolution structure (2.6), then

$$\liminf_{n\to\infty} E_{\theta_0} w[n^{1/2}(T_n - \theta_0)] \geq Ew[I^{-1/2}(\theta_0)Z] \tag{24}$$

by Fatou's lemma and Anderson's lemma (cf. Ibragimov and Has'minskii [21], p. 157). The particular choice $w(x) = |x|^2$ establishes (1.3) for values of θ at which $\{T_n\}$ is regular. If the loss function w is also bounded, then any estimator sequence $\{T_n\}$ that satisfies (2.7) attains the bound (2.16) in the sense that

$$\lim_{n\to\infty} E_{\theta_0} w[n^{1/2}(T_n - \theta_0)] = Ew[I^{-1/2}(\theta_0)Z]. \tag{25}$$

The assumption of regularity in Theorem 2.1 can be weakened technically without changing the conclusions. This observation, recorded in Corollary 2.2 below, was made by Droste and Wefelmeyer [13]. Deeper is the result that an almost everywhere variant of the convolution theorem holds without the regularity assumption (Theorem 2.3 below). We will see that both extensions of Theorem 2.1 have implications for Fisher's program and for the convergence of bootstrap distributions.

A subset $D \subset R^k$ is called a *uniqueness set* if any analytic function defined on an open, connected set that contains D is uniquely determined by its values in D. For example, D could be R^k, as in the proof of Theorem 2.1, or a k-dimensional box in R^k with edges parallel to the coordinate axes, or a dense subset of these.

Definition. A sequence of estimators $\{T_n : n \geq 1\}$ is *essentially regular* at θ_0 if the following holds: There exists a uniqueness set $D \subset R^k$ and, for every $h \in D$, a sequence $\{h_n \in R^k\}$ converging to h such that

$$H_n(\theta_0 + n^{1/2}h_n) \Longrightarrow H(\theta_0) \tag{26}$$

for some limit distribution $H(\theta_0)$ that does not depend on h.

Corollary 2.2. The assumption of regularity in Theorem 2.1 may be replaced by essential regularity without changing the conclusions.

Proof. By reasoning like that in the previous proof, equation (2.10) holds for every $h \in D$. Since the right side of (2.10) is analytic in h and D is a uniqueness set, (2.10) again holds for every $h \in C^k$. The rest of the proof is unchanged. □

Theorem 2.3. Suppose that $P_{\theta,n}$ is LAN at every $\theta \in \Theta$. Let $\{T_n : n \geq 1\}$ be a sequence of estimators such that

$$H_n(\theta) \Longrightarrow H(\theta) \qquad \forall\, \theta \in \Theta. \tag{27}$$

Then there exists a distribution $\nu(\theta)$ and a Lebesgue null set $N \subset \Theta$ such that

$$H(\theta) = N(0, I^{-1}(\theta)) * \nu(\theta) \qquad \forall\, \theta \in \Theta - N. \tag{28}$$

Suppose in addition that both $H(\theta)$ and $I(\theta)$ are continuous at θ_0, the former in a metric for weak convergence. Then the convolution structure (2.20) holds at θ_0.

Remarks on pages 169 and 176 of LeCam [32] imply Theorem 2.3. Jeganathan [24] and Droste and Wefelmeyer [13] gave characteristic function proofs of the theorem. The proof below combines Theorem 2.1 with the following lemma, which is due to Droste and Wefelmeyer [13, pp. 140–141] and extends Lemma 4 in Bahadur [1]. See also Pfanzagl [39, pp. 285–286].

Lemma 2.4. Let $\{f_n : n \geq 1\}$ and f be Lebesgue measurable real-valued functions on R^k such that

$$\lim_{n\to\infty} f_n(x) = f(x) \qquad \text{a. e. Lebesgue.} \tag{29}$$

Then, for every sequence $\{y_n \in R^k\}$ converging to zero, there exists a subsequence M such that

$$\lim_{n\in M} f_n(x + y_n) = f(x) \quad \text{a. e. Lebesgue.} \tag{30}$$

Proof of Theorem 2.3. Let $\phi_n(u, \theta)$ and $\phi(u, \theta)$ again denote the characteristic functions of $H_n(\theta)$ and $H(\theta)$ respectively. Let D be a countable dense subset of R^k. Fix $(h, u) \in D^2$. By the hypothesis (2.19) and Lemma 2.4 with $y_n = n^{-1/2}h$, there is a subsequence $M(h, u)$ and a Lebesgue null set $N(h, u)$ such that

$$\lim_{n\in M(u,h)} \phi_n(u, \theta + n^{1/2}h) = \phi(u, \theta) \qquad \forall\, \theta \in \Theta - N(h, u). \tag{31}$$

Let

$$N = \bigcup_{(h,u)\in D^2} N(h, u). \tag{32}$$

By reasoning like that for Theorem 2.1, equation (2.10) holds for every $(h, u) \in D^2$ and for every $\theta \in \Theta - N$. Because D is a uniqueness class, this implies that equation (2.11) holds for every $u \in D$ and for every $\theta \in \Theta - N$. Conclusion (2.20) follows because a characteristic function is determined by its values on the dense subset $D \subset R^k$.

The last assertion of the theorem already holds if $\theta_0 \in \Theta - N$. Suppose that $\theta_0 \in N$ and that $\{\theta_n \in \Theta - N\}$ converges to θ_0. Let $\psi(u, \theta)$ denote the characteristic function of $\nu(\theta)$, defined for $\theta \in \Theta - N$. Then, from (2.20),

$$\psi(u, \theta_n) = \phi(u, \theta_n) \exp[2^{-1} u' I(\theta_n) u]. \tag{33}$$

Continuity of $I(\theta)$ and weak continuity of $H(\theta)$ now imply that

$$\lim_{n \to \infty} \psi(u, \theta_n) = \phi(u, \theta_0) \exp[2^{-1} u' I(\theta_0) u]. \tag{34}$$

Since the right side of (2.26) is continuous in u, it must be a characteristic function. Consequently, (2.20) holds at θ_0. □

What does the theory of this section entail for Fisher's program? Let w be a symmetric, subconvex, and continuous loss function of R^k. We say that $\{T_n\}$ is *superefficient* at θ_0 for loss function w if

$$\limsup_{n \to \infty} E_{\theta_0} w[n^{1/2}(T_n - \theta_0)] < Ew[I^{-1/2}(\theta_0) Z], \tag{35}$$

where Z has a standard normal distribution on R^k.

Inequality (2.16) shows that superefficiency cannot occur when the limit distribution $H(\theta_0)$ has the convolution structure (2.6). Consequently, lack of regularity at θ_0 is a necessary condition for superefficiency there (Theorem 2.1). Discontinuity of $H(\theta)$ or $I(\theta)$ at θ_0 is also a necessary condition for superefficiency there (Theorem 2.3). The Hodges and the Stein estimators illustrate both necessary conditions.

Example 2 (continued). By (1.6), the Hodges estimator $T_{n,H}$ is superefficient at $\theta_0 = 0$ for $w(x) = x^2$. Let $\theta_n = \theta_0 + n^{-1/2} h$. Then

$$H_n(\theta_n) \Longrightarrow \begin{cases} N((b-1)h, b^2) & \text{if } \theta_0 = 0 \\ N(0,1) & \text{if } \theta_0 \neq 0. \end{cases} \tag{36}$$

This shows directly that $\{T_{n,H}\}$ is not regular at $\theta_0 = 0$. Moreover, the pointwise limit distribution is

$$H_0(\theta_0) = \begin{cases} N(0, b^2) & \text{if } \theta_0 = 0 \\ N(0,1) & \text{if } \theta_0 \neq 0, \end{cases} \tag{37}$$

which has the expected discontinuity at $\theta_0 = 0$ in the topology of weak convergence.

Example 3 (continued). As discussed in the Introduction, the Stein estimator $T_{n,S}$ is superefficient at $\theta_0 = 0$ for $w(x) = |x|^2$. In this instance,

$$H_n(\theta_n) \Longrightarrow \begin{cases} \mathcal{L}[Z - (k-2)(Z+h)/|Z+h|^2] & \text{if } \theta_0 = 0 \\ N(0, I_k) & \text{if } \theta_0 \neq 0 \end{cases} \tag{38}$$

so that $\{T_{n,S}\}$ is not regular at the origin. Let Z be a random vector with standard normal distribution on R^k. The pointwise limit distribution in this example is

$$H_0(\theta_0) = \begin{cases} \mathcal{L}[Z - (k-2)Z/|Z|^2] & \text{if } \theta_0 = 0 \\ N(0, I_k) & \text{if } \theta_0 \neq 0. \end{cases} \tag{39}$$

It has the foreseen discontinuity at $\theta_0 = 0$.

3. BOOTSTRAP CONVERGENCE AND CONVOLUTION

Matters such as estimating the risk of T_n or constructing confidence sets for θ around T_n lead to the question of estimating $H_n(\theta)$, the distribution of $n^{1/2}(T_n - \theta)$. Suppose that $\hat{\theta}_n$ is an estimator of θ, possibly T_n itself. The implied plug-in estimator of $H_n(\theta)$ is then $H_n(\hat{\theta}_n)$.

The random probability measure $H_n(\hat{\theta}_n)$ can also be interpreted as a conditional distribution. Let X_n^* be an artificial sample of size n whose conditional distribution, given the observed sample X_n, is the fitted model $P_{n,\hat{\theta}_n}$. Let $T_n^* = T_n(X_n^*)$ denote the recalculation of T_n from X_n^*. Then

$$H_n(\hat{\theta}_n) = \mathcal{L}[n^{1/2}(T_n^* - \hat{\theta}_n)|X_n]. \tag{40}$$

Efron [14] called $H_n(\hat{\theta}_n)$ the *parametric bootstrap* estimator of $H_n(\theta)$, gave the interpretation as conditional distribution, and drew attention to Monte Carlo approximations for this conditional distribution.

Suppose that $H_n(\theta)$ converges weakly to a limit distribution $H(\theta)$ as n increases. When does the bootstrap distribution $H_n(\hat{\theta}_n)$ converge in probability to the correct limit $H(\theta)$? A substantial literature has grown around this question; two early papers are Bickel and Freedman [6] and Beran [4]. The next theorems link bootstrap convergence with convolution structure. As in the previous section, d is any metric for weak convergence on R^k.

Theorem 3.1. Suppose that $\{T_n : n \geq 1\}$ is a sequence of estimators for which (2.5) holds. Suppose that $\{\hat{\theta}_n : n \geq 1\}$ is a sequence of estimators such that

$$\mathcal{L}[n^{1/2}(\hat{\theta}_n - \theta_0)|P_{\theta_0,n}] \Longrightarrow J(\theta_0) \tag{41}$$

for some limit distribution $J(\theta_0)$. Then

$$d[H_n(\hat{\theta}_n), H(\theta_0)] \to \infty \tag{42}$$

in $P_{\theta_0,n}$-probability.

Proof. Let $V_n = n^{1/2}(\hat{\theta}_n - \theta_0)$. By a Skorokhod construction, there exist random vectors $\{V_n^*(\omega)\}$ and $V^*(\omega)$, defined on a common probability space, such that $\mathcal{L}(V_n^*) = \mathcal{L}(V_n)$, $\mathcal{L}(V^*) = J(\theta_0)$ and $\lim_{n\to\infty} V_n^*(\omega) = V^*(\omega)$ for every elementary event ω. Assumption (2.5) implies that

$$\lim_{n\to\infty} H_n(\theta_0 + n^{-1/2}V_n^*) = H(\theta_0) \qquad \text{w.p.1.} \tag{43}$$

Conclusion (3.3) now follows. □

If the conditions for Theorem 3.1 hold and the model $P_{\theta,n}$ is LAN at θ_0, then the limit distribution $H(\theta_0)$ must have the convolution structure (2.6), because of Theorem 2.1. The next theorem shows that bootstrap convergence itself can imply convolution structure in the limit distribution.

Theorem 3.2. Suppose that the model $P_{\theta,n}$ is LAN at θ_0, that (3.2) and (3.3) hold, and that the support of $J(\theta_0)$ contains a uniqueness set. Then

$$H(\theta_0) = N(0, I^{-1}(\theta_0)) * \nu(\theta_0) \tag{44}$$

for some distribution $\nu(\theta_0)$ on R^k.

Proof. In the notation of the preceding proof, (3.3) implies that

$$d[H_n(\theta_0 + n^{-1/2}V_n^*), H(\theta_0)] \to 0 \tag{45}$$

in probability. Hence, there exists a subsequence M such that

$$\lim_{n\in M} d[H_n(\theta_0 + n^{-1/2}V_n^*), H(\theta_0)] = 0 \qquad \lim_{n\in M} V_n^* = V^* \qquad \text{w.p.1.} \tag{46}$$

Since the possible values of V^* are dense in the support of $J(\theta_0)$, they also form a uniqueness set. Corollary 2.2 thus implies (3.5). □

Corollary 3.3. Suppose that $P_{\theta,n}$ is LAN at every $\theta \in \Theta$. Let $\{\hat{\theta}_n : n \geq 1\}$ be a sequence of estimators such that

$$\mathcal{L}[n^{1/2}(\hat{\theta}_n - \theta) | P_{\theta,n}] \Longrightarrow J(\theta) \qquad \forall\, \theta \in \Theta, \tag{47}$$

for some limit distribution $J(\theta)$. Let $\{T_n : n \geq 1\}$ be a sequence of estimators such that

$$d[H_n(\hat{\theta}_n), H(\theta_0)] \to 0 \tag{48}$$

in $P_{\theta_0,n}$-probability. If both $J(\theta)$ and $I(\theta)$ are continuous at θ_0 or if $\{\hat{\theta}_n : n \geq 1\}$ is regular at θ_0, then (3.5) holds.

Proof. Applying, respectively, Theorem 2.3 or Theorem 2.1 to the estimators $\{\hat{\theta}_n\}$ shows that the support of $J(\theta_0)$ is R^k. Theorem 3.2 then completes the proof. □

Implications. In the setting of Theorem 3.2, superefficiency of $\{T_n\}$ at θ_0 ensures that, under $P_{\theta_0,n}$, the bootstrap distribution $H_n(\hat{\theta}_n)$ cannot converge in probability to the correct limit distribution $H(\theta_0)$. Whether the superefficiency is beneficial to the risk of T_n at $\theta \neq \theta_0$, as in Example 3, or detrimental, as in Example 2, has no effect on the question of bootstrap convergence. The necessary conditions for superefficiency mentioned at the end of Section 2 — lack of regularity at θ_0 or a discontinuity in $H(\theta)$ or $I(\theta)$ at θ_0 — both signal possible bootstrap failure at θ_0.

Example 2 (continued). Let $\hat{\theta}_n$ be the sample mean $\bar{X}_n$. Condition (3.2) is satisfied and $J(\theta_0)$ has full support for every θ_0. By simple extension of (2.28), the Hodges estimator $T_{n,H}$ satisfies (2.5) for every $\theta_0 \neq 0$. Thus, by Theorem 3.1, the bootstrap distribution $H_n(\bar{X}_n)$ for $T_{n,H}$ converges correctly, in probability, whenever $\theta_0 \neq 0$.

By Theorem 3.2, $H_n(\bar{X}_n)$ cannot converge properly at $\theta_0 = 0$, because the origin is a superefficiency point for the Hodges estimator. In this exceptional case, reasoning related to (2.28) shows that $H_n(\bar{X}_n)$ converges *in distribution*, as a random element of the space of all probability measures on R^k metrized by weak convergence, to the random probability measure $N((b-1)Z, b^2)$. Here Z has a standard $N(0,1)$ distribution.

Example 3 (continued). By extension of (2.30), the Stein estimator $T_{n,S}$ satisfies (2.5) for every $\theta_0 \neq 0$. By Theorem 3.1, the bootstrap distribution $H_n(\bar{X}_n)$ for $T_{n,S}$ converges correctly whenever $\theta_0 \neq 0$. At the superefficiency point $\theta_0 = 0$, the failure of the bootstrap distribution to converge correctly (Theorem 3.2) can be clarified as follows. Let Z be a random vector with standard normal distribution on R^k and let

$$\pi(h) = \mathcal{L}[Z - (k-2)(Z+h)/|Z+h|^2]. \tag{49}$$

An argument akin to (2.30) shows that $H_n(\bar{X}_n)$ converges in distribution, as a random probability measure, to the random probability measure $\pi(Z)$.

The behavior at the origin of the Hodges and Stein estimators, as well as other examples, motivates the following abstraction.

Definition. A sequence of estimators $\{T_n\}$ is *locally uniformly weakly convergent* at θ_0 if there exists a family of distributions $\{\pi(\theta_0, h)\colon h \in R^k\}$ such that

$$H_n(\theta_0 + n^{-1/2}h_n) \Longrightarrow \pi(\theta_0, h) \tag{50}$$

for every $h \in R^k$ and every sequence $\{h_n \in R^k\}$ converging to h.

Theorem 3.4. Suppose that the estimators $\{T_n\}$ are locally uniformly weakly convergent at θ_0. Let $\{\hat{\theta}_n\}$ be a sequence of estimators that satisfies (3.2). Let V be a random vector whose distribution is $J(\theta_0)$. Under $P_{\theta_0,n}$, $H_n(\hat{\theta}_n)$ converges in distribution, as a random probability measure, to the random probability measure $\pi(\theta_0, V)$. In general, this limit differs from $H(\theta_0) = \pi(\theta_0, 0)$.

Proof. Let V_n, V_n^* and V^* be as in the proof of Theorem 3.1. Condition (3.11) implies that

$$H_n(\theta_0 + n^{-1/2}V_n^*) \Longrightarrow \pi(\theta_0, V^*) \qquad \text{w.p.1.} \tag{51}$$

The result follows. □

This theorem on possible bootstrap failure also suggests a remedy: use a bootstrap sample size m_n much smaller than n. The idea of the cure is due to Bretagnolle [8].

Corollary 3.5. Suppose that the conditions for Theorem 3.4 hold. Let $\{m_n : n \geq 1\}$ be a sequence of integers such that

$$\lim_{n\to\infty} m_n/n = 0 \qquad \lim_{n\to\infty} m_n = 0. \tag{52}$$

Then

$$d[H_{m_n}(\hat{\theta}_n), H(\theta_0)] \to 0 \tag{53}$$

in $P_{\theta_0,n}$ probability.

Proof. Observe, in the notation of the previous proof, that $H_{m_n}(\hat{\theta}_n)$ has the same distribution as $H_{m_n}(\theta_0 + m_n^{-1/2} W_n^*)$, where $W_n^* = (m_n/n)^{1/2} V_n^*$. Since $\lim_{n\to\infty} W_n^* = 0$ and $\lim_{n\to\infty} m_n = \infty$,

$$H_{m_n}(\theta_0 + m_n^{-1/2} W_n^*) \to \pi(\theta_0, 0) = H(\theta_0) \qquad \text{w.p.1.} \tag{54}$$

as in (3.12). The corollary follows. □

The best choice of m_n in this *subsample* bootstrap is the subject of current research. One difficulty is that $H_{m_n}(\hat{\theta}_n)$ can be highly inefficient as an estimator of $H_n(\theta_0)$ when (3.13) holds and θ_0 is a regularity point (cf. Beran [3]).

4. LAM VIEWS OF SUPEREFFICIENCY

For estimation in the normal model of Example 2, Chernoff [10] stated a local asymptotic minimax (LAM) bound that he attributed to unpublished work by C. Stein and by H. Rubin. This bound, like Theorem 14 in LeCam [28], brings out what is wrong with the Hodges estimator in small neighborhoods of its superefficiency point. Only years later, in Hájek [18], was the LAM approach formulated for general LAN parametric models. A convenient version of Hájek's result is the following (c.f. Ibragimov and Has'minskii [21, p. 162]:

Theorem 4.1. Suppose that the model $P_{\theta,n}$ is LAN at θ_0. Let w be a symmetric, subconvex, and continuous loss function. Let Z be a random vector with standard normal distribution on R^k. Then

$$\lim_{c\to\infty} \liminf_{n\to\infty} \inf_{T_n} \sup_{n^{1/2}|\theta-\theta_0|\leq c} E_\theta w[n^{1/2}(T_n - \theta)] \geq Ew[I^{-1/2}(\theta_0)Z], \tag{55}$$

the infimum being taken over all estimators of θ.

Unlike the convolution theorem, the statement of Theorem 4.1 applies to all estimators at every $\theta \in \Theta$. Suppose that a sequence of estimators $\{T_n\}$ satisfies (2.7) and that the loss function w is also bounded. Then, the lower bound (4.1) is attained asymptotically in the sense that

$$\lim_{c\to\infty} \lim_{n\to\infty} \sup_{n^{1/2}|\theta-\theta_0|\leq c} E_\theta w[n^{1/2}(T_n - \theta)] = Ew[I^{-1/2}(\theta_0)Z]. \tag{56}$$

Hájek [18] proved a partial converse to this last statement. When $k = 1$ and w is nonconstant, then (4.2) implies that $\{T_n\}$ must satisfy (2.7). This local asymptotic admissibility result is restricted to estimation of low-dimensional parameters. It typically breaks down for $k \geq 3$, as noted by van der Vaart [53, pp. 1491–1492].

Example 2 (continued). Suppose that $k = 1$ and $w(x) = x^2$. At every $\theta_0 \neq 0$, the Hodges estimator is LAM in the sense that (4.2) holds. However, at $\theta_0 = 0$,

$$\lim_{c\to\infty} \lim_{n\to\infty} \sup_{n^{1/2}|\theta|\leq c} nE_\theta(T_{n,H} - \theta)^2 = \infty. \tag{57}$$

The infinite limit in (4.3) indicates dramatically the poor performance of the Hodges estimator near its point of superefficiency. By contrast, the limiting maximum risk of the sample mean $\bar{X}_n$ is 1.

Example 3 (continued). Suppose that $k \geq 3$ and $w(x) = |x|^2$. At every $\theta_0 \in R^k$, including the superefficiency point $\theta_0 = 0$, the Stein estimator $T_{n,S}$ is LAM in the sense that (4.2) holds. The same is true of the sample mean vector $\bar{X}_n$. Thus, the LAM criterion of Theorem 4.1 fails to detect the improved performance of $T_{n,S}$ around the point of superefficiency $\theta_0 = 0$.

This insensitivity in the LAM criterion can be overcome by suitably linking the value of c to the dimension k and then letting the latter increase. In place of T_n or θ, we will write $T_{n,k}$ or θ_k to emphasize that the dimension of these quantities is varying. For every finite positive b,

$$\liminf_{k\to\infty} \liminf_{n\to\infty} \inf_{T_{n,k}} \sup_{n^{1/2}|\theta_k|\leq k^{1/2}b} k^{-1}nE_{\theta_k}|T_{n,k} - \theta_k|^2 \geq b^2/(1+b^2), \tag{58}$$

the infimum being taken over all estimators $T_{n,k}$. To prove this, apply Pinsker's [41] minimax bound for estimation in a Gaussian process to the limit experiment here, which is a multivariate normal location model with identity covariance matrix. Stein [50] pioneered dimensional asymptotics for this normal model.

The Stein estimator $T_{n,k,S}$ in k dimensions achieves the asymptotic lower bound (4.4) because

$$\lim_{k\to\infty} \lim_{n\to\infty} \sup_{n^{1/2}|\theta_k|\leq k^{1/2}b} k^{-1}nE_{\theta_k}|T_{n,k,S} - \theta_k|^2 = b^2/(1+b^2). \tag{59}$$

For the sample mean vector, the right side of (4.5) must be replaced by 1. Unlike the LAM bound of Theorem 4.1, version (4.4) detects the improvement achieved by the Stein estimator around the point of superefficiency when dimension k is large.

5. EPILOG

Hájek's [17] and [18] papers inspired work by an international array of authors. LeCam [31] gave a deep generalization of the convolution theorem and of the LAM bound to models whose limit experiment need not be normal. LeCam [33] presented an abstract version of Hájek's asymptotic admissibility result for one-dimensional estimators that extends beyond estimation and the LAN setup. Nonparametric forms of the convolution theorem were found by Beran [2], Millar [38], and others. Nonparametric versions of the LAM bound were obtained by Levit [37], Koshevnik and Levit [27], and others. The ideas of Stein [49] played an important role in these extensions. Jeganathan [25] initiated detailed study of the case where the limit experiment is mixed normal. More recently, van der Vaart [53] treated quadratic mean differentiable models whose tangent sets are not necessarily linear spaces. Much of what we have learned since Hájek's two papers is covered in monographs by Roussas [46], Ibragimov and Has'minskii [21], Pfanzagl and Wefelmeyer [40], LeCam [34], van der Vaart [52], LeCam and Yang [35], Bickel, Klaassen, Ritov, and Wellner [7], Pfanzagl [39], and in other books cited by these authors.

ACKNOWLEDGEMENT

This research was supported in part by National Science Foundation grant DMS 9224868.

(Received October 19, 1994.)

REFERENCES

[1] R. R. Bahadur: A note on Fisher's bound for asymptotic variance. Ann. Math. Statist. *35* (1964), 1545–1552.
[2] R. Beran: Estimating a distribution function. Ann. Statist. *5* (1977), 400–404.
[3] R. Beran: Estimated sampling distributions: the bootstrap and competitors. Ann. Statist. *10* (1982), 212–225.
[4] R. Beran: Bootstrap methods in statistics. Jahresber. Deutsch. Math.-Verein. *86* (1984), 212–225.
[5] R. Beran: Confidence sets centered at C_p-estimators. Unpublished preprint (1994).
[6] P. J. Bickel and D. F. Freedman: Some asymptotic theory for the bootstrap. Ann. Statist. *9* (1981), 1196–1217.
[7] P. J. Bickel, C. A. Klaassen, Y. Ritov and J. A. Wellner: Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins Univ. Press, Baltimore 1993.
[8] J. Bretagnolle: Lois limites du bootstrap de certaines fonctionnelles. Ann. Inst. Henri Poincaré *X9* (1983), 281–296.
[9] G. Casella and T. J. Hwang: Limit expressions for the risk of James–Stein estimators. Canad. J. Statist. *10* (1982), 305–309.
[10] H. Chernoff: Large sample theory: parametric case. Ann. Math. Statist. *27* (1956), 1–22.
[11] H. Cramér: A contribution to the theory of statistical estimation. Skand. Akt. Tidskr. *29* (1946), 85–94.
[12] G. Darmois: Sur les lois limites de la dispersion de certaines estimations. Rev. Inst. Internat. Statist. *13* (1945), 9–15.
[13] W. Droste and W. Wefelmeyer: On Hájek's convolution theorem. Statist. Decisions *2* (1984), 131–144.
[14] B. Efron: Bootstrap methods: Another look at the jackknife. Ann. Statist. *7* (1979), 1–26
[15] R. A. Fisher: Theory of statistical estimation. Proc. Cambridge Philos. Soc. *22* (1925), 700–725.

[16] M. Fréchet: Sur l'extension de certaines evaluations statistiques de petit echantillons. Rev. Inst. Internat. Statist. *11* (1943), 182–205.
[17] J. Hájek: A characterization of limiting distributions of regular estimates. Z. Wahrsch. verw. Gebiete *14* (1970), 323–330.
[18] J. Hájek: Local asymptotic minimax and admissibility in estimation. In: Proc. Sixth Berkeley Symp. Math. Statist. Prob. (L. M. LeCam, J. Neyman and E. M. Scott, eds.), *1*, Univ. California Press, Berkeley 1972, pp. 175–194.
[19] J. Hájek and Z. Šidák: Theory of Rank Tests. Academic Press, New York 1967.
[20] B. M. Hill: The three-parameter lognormal distribution and Bayesian analysis of a point source epidemic. J. Amer. Statist. Assoc. *58* (1963), 72–84.
[21] I. A. Ibragimov and R. Z. Has'minskii: Statistical Estimation: Asymptotic Theory. Springer, New York 1981.
[22] N. Inagaki: On the limiting distribution of a sequence of estimators with uniformity property. Ann. Inst. Statist. Math. *22* (1970), 1–13.
[23] W. James and C. Stein: Estimation with quadratic loss. In: Proc. Fourth Berkeley Symp. Math. Statist. Prob. (J. Neyman, ed.) *1*, Univ. California Press, Berkeley 1961, pp. 361–380.
[24] P. Jeganathan: On a decomposition of the limit distribution of a sequence of estimators. Sankhyā (Ser. A) *43* (1981), part 1, 26–36.
[25] P. Jeganathan: On the asymptotic theory of estimation when the limit of the log-likelihood ratios is mixed normal. Sankhyā (Ser. A) *44* (1982), part 2, 173–212.
[26] S. Kaufman: Asymptotic efficiency of the maximum likelihood estimator. Ann. Inst. Statist. Math. *18* (1966), 155–178.
[27] Yu. A. Koshevnik and B. Ya. Levit: On a nonparametric analog of the information matrix. Theory Probab. Appl. *21* (1975), 738–753.
[28] L. LeCam: On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. Univ. California Publ. Statist. *1* (1953), 277–330.
[29] L. LeCam: On the asymptotic theory of estimation and testing hypotheses. In: Proc. Third Berkeley Symp. Math. Statist. Prob. (J. Neyman, ed.) *1*, Univ. California Press, Berkeley 1956, pp. 129–156.
[30] L. LeCam: Locally asymptotically normal families of distributions. Univ. California Publ. Statist. *3* (1960), 27–98.
[31] L. LeCam: Limits of experiments. In: Proc. Sixth Berkeley Symp. Math. Statist. Prob. (L. M. LeCam, J. Neyman, and E. M. Scott, eds.), *1*, Univ. California Press, Berkeley 1972, pp. 245–261.
[32] L. LeCam: Sur les contraintes imposées par les passages à limite usuels en statististique. Bull. Inst. Internat. Statist. *45* (1972), 169–180.
[33] L. LeCam: On a theorem of J. Hájek. In: Contributions to Statistics: J. Hájek Memorial Volume (J. Jurečková, ed.), Academia, Prague 1979, pp. 119–137.
[34] L. LeCam: Asymptotic Methods in Statistical Decision Theory. Springer, New York 1986.
[35] L. LeCam and G. L. Yang: Asymptotics in Statistics. Springer, New York 1990.
[36] E. L. Lehmann: Theory of Point Estimation. Wiley, New York 1983.
[37] B. Ya. Levit: On the efficiency of a class of nonparametric estimates. Theory Probab. Appl. *20* (1975), 723–740.
[38] P. W. Millar: Non-parametric applications of an infinite dimensional convolution theorem. Z. Wahrsch. verw. Gebiete *68* (1985), 545–556.
[39] J. Pfanzagl: Parametric Statistical Theory. Walter de Gruyter, Berlin 1994.
[40] J. Pfanzagl and W. Wefelmeyer: Contributions to a General Asymptotic Statistical Theory. (Lecture Notes in Statistics 13.) Springer, New York 1982.
[41] M. S. Pinsker: Optimal filtration of square-integrable signals in Gaussian white noise. Problems Inform. Transmission *16* (1980), 120–133.

[42] B. Pötscher: Comment on "The effect of model selection on confidence regions and prediction regions" by P. Kabaila. Econometric Theory, to appear.
[43] J. W. Pratt: F. Y. Edgeworth and R. A. Fisher on the efficiency of maximum likelihood estimation. Ann. Statist. *4* (1976), 501–514.
[44] C. R. Rao: Information and accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. *37* (1945), 81–91.
[45] G. G. Roussas: Some applications of the asymptotic distribution of likelihood functions to the asymptotic efficiency of estimates. Z. Wahrsch. verw. Gebiete *10* (1968), 252–260.
[46] G. R. Roussas: Contiguous Probability Measures: Some Applications in Statistics. Cambridge Univ. Press, Cambridge 1972.
[47] L. J. Savage: On rereading R. A. Fisher. Ann. Statist. *4* (1976), 441-500.
[48] L. Schmetterer: On asymptotic efficiency of estimates. In: Research Papers in Statistics: Festschrift for J. Neyman (F. N. David, ed.), Wiley, New York 1966, pp. 310–317.
[49] C. Stein: Efficient nonparametric testing and estimation. In: Proc. Third Berkeley Symp. Math. Statist. Prob. (J. Neyman, ed.), *1*, Univ. California Press, Berkeley 1956, pp. 187–195.
[50] C. Stein: Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Proc. Third Berkeley Symp. Math. Statist. Prob. (J. Neyman, ed.), *1*, Univ. California Press, Berkeley 1956, pp. 197–206.
[51] D. M. Titterington: Choosing the regularization parameter in image restoration. In: Spatial Statistics and Imaging (A. Possolo, ed.), IMS Lecture Notes – Monograph Series *20* (1991), 392–402.
[52] A. van der Vaart: Statistical Estimation in Large Parameter Spaces. CWI Tracts *44*, Centrum voor Wiskunde en Informatica, Amsterdam 1988.
[53] A. van der Vaart: On the asymptotic information bound. Ann. Statist. *17* (1989), 1487–1500.
[54] J. Wolfowitz: Asymptotic efficiency of the maximum likelihood estimator. Theory Probab. Appl. *10* (1965), 247–260.

Prof. Dr. Rudolf Beran, Department of Statistics, University of California, Berkeley, CA 94720–3860. U. S. A.