27
Superefficiency

A. W. van der Vaart[1]

ABSTRACT We review the history and several proofs of the famous result of Le Cam that a sequence of estimators can be superefficient on at most a Lebesgue null set.

27.1 Introduction

The method of maximum likelihood as a general method of estimation in statistics was introduced and developed by Fisher (1912, 1922, 1925, 1934). It gained popularity as it appeared that the method automatically produces efficient estimators if the number of observations is large. The concept of asymptotic efficiency was invented by Fisher as early as 1922 roughly in the form as we use it for regular models today: a sequence of statistics is efficient if it tends to a normal distribution with the least possible standard deviation. In the 1930s and 1940s there were many steps in the direction of a rigorous foundation of Fisher's remarkable insights. These consisted both of proofs of the asymptotic normality of maximum likelihood estimators and of obtaining lower bounds for the variance of estimators.

Chapters 32 and 33 of Cramér (1946) give a summary of the state of affairs in the mid 1940s, even though some work carried out in the early war years, notably Wald's, had been unavailable to him. Chapter 32 gives a rigorous proof of what we now know as the Cramér-Rao inequality and next goes on to define the asymptotic efficiency of an estimator as the quotient of the inverse Fisher information and the asymptotic variance. Next Chapter 33 gives a rigorous proof of asymptotic normality of the maximum likelihood estimator, based on work by Dugué (1937).

Cramér defines an estimator sequence to be *asymptotically efficient* if its asymptotic efficiency (the quotient mentioned previously) equals one. Thus combination of the results of the two chapters leads to the correct conclusion that the method of maximum likelihood produces asymptotically efficient estimators, under some regularity conditions on the underlying densities. Apparently the conceptual hole in the definition was not fully recognized until 1951, even though the difficulty must have been clear to several authors who had worked on establishing efficiency within restricted classes of estimators.

[1] Vrije Universiteit

In 1951 Hodges produced the first example of a *superefficient* estimator sequence: an estimator sequence with efficiency at least one for all θ and more than one for some θ. An abstraction of Hodges' example is the following. Let T_n be a sequence of estimators of a real parameter θ such that the sequence $\sqrt{n}(T_n - \theta)$ converges to some limit distribution if θ is the true parameter, under every θ. If $S_n = T_n 1\{|T_n| > n^{-1/4}\}$, then the probability of the sequence of events $\{T_n = S_n\}$ converges to one under every $\theta \neq 0$, while under $\theta = 0$ the probability of the event $\{S_n = 0\}$ converges to one. In particular, if the first sequence of estimators T_n is asymptotically efficient in the sense of Cramér, then the sequence S_n is superefficient at $\theta = 0$.

Hodges' example revealed a difficulty with the definition of asymptotic efficiency and threw doubt on Fisher's assertion that the maximum likelihood estimator is asymptotically efficient. In this paper we review three lines of approach addressing the matter. They were all initiated by Le Cam. Already in 1952 Le Cam had announced in an abstract to the Annals of Mathematical Statistics that the set of superefficiency can never be larger than a Lebesgue null set. In the next section we review his proof, which appeared in Le Cam (1953). Le Cam's second approach is present in Le Cam (1973) and is based on automatic invariance. We discuss it in Section 3. The third approach combines elements of both papers and is given in Section 4. Particularly in this last section we do not strive for the utmost generality. We hope that simple proofs may help these beautiful results finally find their way into text books and lecture notes.

In the following, *superefficiency* of a sequence of estimators in the locally asymptotically normal case will be understood in the sense that

$$\limsup_{n\to\infty} \mathrm{E}_\theta \ell\big(\sqrt{n}(T_n - \theta)\big) \leq \int \ell \, dN_{0, I_\theta^{-1}},$$

for every θ, with strict inequality for some θ. Here I_θ is the Fisher information matrix and ℓ a given loss function.

Convergence in distribution is denoted $\rightsquigarrow$ and convergence in distribution under a law given by a parameter θ by $\overset{\theta}{\rightsquigarrow}$.

27.2 The 1953 proof

Le Cam (1953) started his paper with examples of superefficient estimator sequences. These include Hodges' example, but also estimators that are superefficient on a dense set of parameters. Next he went on to prove that superefficiency can occur only on Lebesgue null sets. The main idea is that the sequence of maximum likelihood estimators is asymptotically Bayes with respect to Lebesgue absolutely continuous priors on the parameter set. Specifically, let $\hat{\theta}_n$ be the maximum likelihood estimator based on a sample of size n from a

density p_θ. Under smoothness conditions on the map $\theta \mapsto p_\theta$ Le Cam showed that

$$\limsup_{n\to\infty} \int \mathrm{E}_\theta \ell\big(\sqrt{n}(\hat\theta_n - \theta)\big)\,\pi(\theta)\,d\theta \le \liminf_{n\to\infty} \int \mathrm{E}_\theta \ell\big(\sqrt{n}(T_n - \theta)\big)\,\pi(\theta)\,d\theta, \tag{1}$$

for every sequence of estimators T_n, most prior densities $\pi(\theta)$ and most symmetric, bounded, continuous loss functions ℓ. Since the standardized sequence of maximum likelihood estimators $\sqrt{n}(\hat\theta_n - \theta)$ is asymptotically normal $N(0, I_\theta^{-1})$, the first limsup exists as an ordinary limit. Application of Fatou's lemma immediately yields that

$$\int \Big(\int \ell\, dN_{0,I_\theta^{-1}} - \limsup_{n\to\infty} \mathrm{E}_\theta \ell\big(\sqrt{n}(T_n - \theta)\big)\Big)\,\pi(\theta)\,d\theta \le 0.$$

Superefficiency of the sequence T_n would imply that the integrand is nonnegative. Since it integrates nonpositive it must be zero almost surely under π.

Rigorous proofs of the asymptotic normality of the sequence of maximum likelihood estimators were available, for instance from Cramér (1946). The essential part of the preceding argument was therefore the proof of (1). Le Cam based his proof on a version of the Bernstein-von Mises theorem. Let Θ be a random variable with Lebesgue density π on the parameter set and consider $\prod_{i=1}^n p_\theta(x_i)$ as the conditional density of $(X_1, \ldots, X_n)$ given $\Theta = \theta$. Le Cam (1953) proved (under regularity conditions) that, for every θ, with $\|\cdot\|$ denoting the total variation norm,

$$\Big\| \mathcal{L}\big(\sqrt{n}(\Theta - \hat\theta_n) \mid X_1, \ldots, X_n\big) - N(0, I_\theta^{-1}) \Big\| \to 0, \qquad \text{a.s. } [P_\theta].$$

This strengthened earlier results by Bernstein (1934) and von Mises (1931) to the point where application towards proving (1) is possible. In the present notation the Bayes risk of T_n can be written

$$\begin{aligned} &\int \mathrm{E}_\theta \ell\big(\sqrt{n}(T_n - \theta)\big)\,\pi(\theta)\,d\theta \\ &\qquad = \mathrm{E}\,\mathrm{E}\Big(\ell\big(\sqrt{n}(T_n - \hat\theta_n) - \sqrt{n}(\Theta - \hat\theta_n)\big) \;\Big|\; X_1, \ldots, X_n\Big). \end{aligned}$$

According to the Bernstein-von Mises theorem the conditional expectation in this expression satisfies, setting $\mu_n = \sqrt{n}(T_n - \hat\theta_n)$,

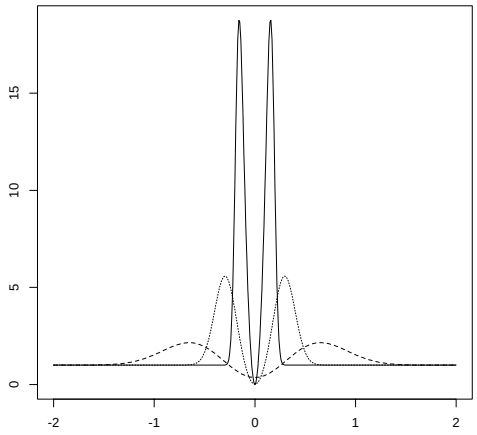
$$\mathrm{E}\Big(\ell\big(\mu_n - \sqrt{n}(\Theta - \hat\theta_n)\big) \;\Big|\; X_1, \ldots, X_n\Big) - \int \ell\, dN_{-\mu_n, I_{\hat\theta_n}^{-1}} \to 0,$$

almost surely under every θ, hence also under the mixtures $\int P_\theta^\infty \pi(\theta)\,d\theta$. It is assumed at this point that the loss function ℓ is bounded; otherwise a stronger version of the Bernstein-von Mises theorem would be necessary. For the usual symmetric loss functions the normal expectation in the preceding display decreases if μ_n is replaced by zero. This readily leads to (1).

From today's perspective the references to asymptotic normality and the maximum likelihood estimator are striking. As Le Cam was later to point out neither of the two are essential for the principle of superefficiency. The

special role of maximum likelihood estimators was removed in Le Cam (1956), where they were replaced by one-step estimators. Next Le Cam (1960, 1964) abstracted the structure of asymptotically normal problems into the 'local asymptotic normality' condition and finally removed even the latter in Le Cam (1972, 1973).

The use of Bayes estimators is in tune with the statistical paradigm of the 1940s and 1950s. Wald (1950)'s book firmly established statistical decision theory as the basis of statistical reasoning. A main result was that Bayes estimators (or rather their limit points) form a complete class. Wolfowitz (1953) exploited this to explain the impossibility of superefficiency in an informal manner. The preceding argument shows that the risk functions of Bayes estimators are asymptotically equivalent to the risk function of the maximum likelihood estimator. Thus asymptotically the maximum likelihood estimator is the only Bayes estimator. This would establish its asymptotic admissibility, and also its optimality. Only a more precise argument would properly explain the role of sets of Lebesgue measure zero.



Quadratic risk function of the Hodges estimator based on a sample of size 10 (dashed), 100 (dotted) and 1000 (solid) observations from the $N(\theta, 1)$-distribution.

In the final section of his paper Le Cam (1953) also showed that in the case of one-dimensional parameters superefficient estimators necessarily have undesirable properties. For the Hodges' estimator $T = \bar{X}1\{|\bar{X}| > n^{-1/4}\}$ based on a sample of size n from the $N(\theta, 1)$-distribution this is illustrated in the Figure, which shows the risk function $\theta \mapsto n\mathrm{E}_\theta(T - \theta)^2$ for three different values of n. Le Cam shows that this behaviour is typical: superefficiency at the point θ for a loss function ℓ implies the existence of a sequence $\theta_n \to \theta$ such that $\liminf \mathrm{E}_{\theta_n} \ell\big(\sqrt{n}(T_n - \theta_n)\big)$ is strictly larger than $\int \ell \, dN_{0,1/I_\theta}$. For the extreme case where the asymptotic risk at θ is zero, the liminf is even infinite for a sequence $\theta_n \to \theta$.

This result may be considered a forerunner of the local asymptotic minimax theorem of Hájek (1972), which states that the maximum risk over a shrinking neighbourhood of θ is asymptotically bounded below by $\int \ell \, dN_{0,1/I_\theta}$. A much earlier result of this type was obtained by Chernoff (1956), who essentially showed that

$$\lim_{c\to\infty} \liminf_{n\to\infty} \sup_{-c<h<c} \mathrm{E}_{\theta+h/\sqrt{n}}\Big(\sqrt{n}\big|T_n - \theta - h/\sqrt{n}\big| \wedge c\Big)^2 \geq \frac{1}{I_\theta}.$$

Chernoff's proof is based on a version of the Cramér-Rao inequality, which he attributed to Stein and Rubin. The theorem may have looked somewhat too complicated to gain popularity. Nevertheless Hájek's result, for general locally asymptotically normal models and general loss functions, is now considered the final result in this direction. Hájek wrote:

> The proof that local asymptotic minimax implies local asymptotic admissibility was first given by LeCam (1953, Theorem 14). $\cdots$. Apparently not many people have studied Le Cam's paper so far as to read this very last theorem, and the present author is indebted to Professor LeCam for giving him the reference.

Not reading to the end of Le Cam's papers became not uncommon in later years. His ideas have been regularly rediscovered.

Le Cam's Theorem 14 about the bad properties of superefficient estimators only applies to one-dimensional estimators. Le Cam commented:

> In the case of an r-dimensional parameter the problem becomes more complicated. The difficulties involved are conceptual as well as mathematical.

This is very true: we now know that a similar result is false in dimensions three and up. Only three years after Le Cam's paper, Stein (1956) published his famous paper on estimating a multivariate normal mean. The James-Stein estimator

$$T_n = \bar{X} - (d-2)\frac{\bar{X}}{n\|\bar{X}\|^2}$$

is superefficient at $\theta = 0$ for the loss function $\ell(x) = \|x\|^2$, and it does not behave badly in a neighbourhood of this point (for this loss function).

27.3 Automatic invariance

The second approach to proving the impossibility of superefficiency is based on the remarkable fact that the usual rescaling of the parameter (for instance considering $h = \sqrt{n}(\theta - \theta_0)$ as the parameter, rather than θ) automatically leads to asymptotic equivariance at almost every θ_0. This idea is put forward in Le Cam (1973) and leads to a completely different proof than the proof in

Le Cam (1953). On comparing the two papers the difference in style is also apparent. The main result of Le Cam (1973) asserts shift invariance of limit experiments and is stated within the abstract framework of L- and M-spaces. The important application is only indicated in the second last paragraph:

> Toutefois, et pour conclure, mentionnons que la démonstration du résultat de convolution de Hájek (1970) s'étend à tous les cas considérés ici, pourvu qu'elle soit faite par la méthode décrite dans Le Cam (1972).

The application must have been obvious to Le Cam. This would explain that Section 8.4 of Le Cam (1986), which is concerned with the same subject, seems to end without a conclusion regarding superefficiency as well. In this section we present a simplified version of Le Cam's (1973) result, suited to superefficiency.

Asymptotic equivariance subsumes the Hájek regularity property, which was the key requirement for Hájek (1970)'s convolution theorem. He defined an estimator sequence T_n based on n observations from a smooth parametric model to be *regular* at the parameter θ if

$$\sqrt{n}(T_n - \theta - h/\sqrt{n}) \overset{\theta+h/\sqrt{n}}{\rightsquigarrow} L_\theta, \qquad \text{every } h, \tag{2}$$

for some fixed probability distribution L_θ. The independence of L_θ of h is the crucial feature of regularity. Hájek's (1970) convolution theorem states that in this situation the limiting distribution L_θ is a convolution of the type

$$L_\theta = N(0, I_\theta^{-1}) * M_\theta.$$

This certainly implies that the covariance matrix of L_θ is bounded below by the inverse I_θ^{-1} of the Fisher information matrix.

The 'local uniformity' in the weak convergence required by Hájek regularity looks not too unnatural, though on closer inspection not all interesting estimators turn out to be regular. Shrinkage estimators are not regular at the shrinkage point; estimators that are truncated to a parameter set (such as $\bar{X} \vee 0$ if a mean is known to be positive) are not regular at the boundary of the parameter set. In these examples the set of points of irregularity is very small. It turns out that this is necessarily the case. Below we shall show that the regularity holds automatically at almost all parameter points for any estimator sequence such that $\sqrt{n}(T_n - \theta)$ has a limiting distribution under every θ.

This 'automatic regularity' is the key connection to superefficiency, for it follows that the limit distributions L_θ are convolutions for almost every θ. In particular the asymptotic covariance matrix is bounded below by the inverse Fisher information matrix for almost every θ. This constitutes a modern proof of the fact that superefficiency can occur only on Lebesgue null sets.

Hájek's proof of the convolution theorem is based on delicate calculations using the special character of local asymptotic normality. Le Cam (1972)'s theory of limiting experiments puts the result in a very general framework. Not only does it offer much insight in the Gaussian situation, it also allows

superefficiency statements in many other situations. We shall now carry out the preceding steps in more detail and in much greater generality.

From the more general point of view regularity is better described as (local) asymptotic equivariance. My favourite version (Van der Vaart (1991)) of Le Cam's result is as follows. A sequence of experiments (or statistical models) is said to *converge* to a limit if the marginals of the likelihood ratio processes converge in distribution to the corresponding marginals of the likelihood ratio processes in the limit experiment. The precise definition of convergence is not important for this paper: convergence of experiments only enters as a condition of the following theorem and through statements regarding concrete examples, which are not proven here.

(3) **Proposition** *Let the experiments $(\mathcal{X}_n, \mathcal{A}_n, P_{n,h}: h \in H)$ converge to a dominated experiment $(P_h: h \in H)$. Let $\kappa_n: H \mapsto \mathbb{D}$ be maps with values in a Banach space $\mathbb{D}$ such that*

$$r_n\big(\kappa_n(h) - \kappa_n(h_0)\big) \to \dot\kappa h - \dot\kappa h_0,$$

for some map $\dot\kappa: H \mapsto \mathbb{D}$ and linear maps $r_n: \mathbb{D} \mapsto \mathbb{D}$. Let $T_n: \mathcal{X}_n \mapsto \mathbb{D}$ be arbitrary maps with values in $\mathbb{D}$ such that the sequence $r_n\big(T_n - \kappa_n(h)\big)$ converges in distribution under h, for every h, to a probability distribution that is supported on a fixed separable, Borel measurable subset of $\mathbb{D}$. Then there exists a randomized estimator T in the limit experiment such that $r_n\big(T_n - \kappa_n(h)\big)$ converges under h to $T - \dot\kappa h$, for every h.

In this proposition a randomized estimator is a measurable map $T: \mathcal{X} \times [0,1] \mapsto \mathbb{D}$ whose law is to be calculated under the product of P_h and the uniform law. Thus $T = T(X, U)$ is based on an observation X in the limit experiment and an independent uniform variable U.

We could call the estimator sequence T_n in the preceding proposition *regular* if the limiting distribution under h of the sequence $r_n\big(T_n - \kappa_n(h)\big)$ is the same for every h. Then the matching randomized estimator in the limit experiment satisfies

$$\mathcal{L}_h\big(T - \dot\kappa h\big) = \mathcal{L}_0(T), \qquad \text{every } h.$$

This may be expressed as: T is *equivariant-in-law* for estimating the parameter $\dot\kappa h$.

Within the context of the proposition 'regularity' has lost its interpretation as a local uniformity requirement. This is recovered when the proposition is applied to 'localized' experiments. For instance, local asymptotic normality of the sequence of experiments $(\mathcal{X}_n, \mathcal{A}_n, P_\theta^n: \theta \in \Theta)$ at θ entails convergence of the sequence of local experiments to a Gaussian experiment:

$$(\mathcal{X}_n, \mathcal{A}_n, P^n_{\theta + h/\sqrt{n}}: h \in \mathbb{R}^d) \to \big(N_d(h, I_\theta^{-1}): h \in \mathbb{R}^d\big).$$

Regularity of a given estimator sequence for the functionals $\kappa_n(h) = \theta + h/\sqrt{n}$ in this sequence of localized experiments means exactly Hájek regularity as in (2). The proposition shows that the limit distribution L_θ is the distribution under $h = 0$ of an equivariant-in-law estimator for h in the Gaussian experiment.

Hájek's convolution theorem is reproved once it has been shown that every such equivariant-in-law estimator can be decomposed as a sum $X + W$ of two independent variables, where X is $N_d(0, I_\theta^{-1})$-distributed. In any case the 'best' equivariant-in-law estimator is X itself, so that the covariance of L_θ is not smaller than I_θ^{-1}.

The following theorem shows that estimator sequences in rescaled experiments are automatically (almost) regular, at almost every parameter. The proof of the theorem is based on an extension of a lemma by Bahadur (1964), who used his lemma to rederive Le Cam (1953)'s result for one-dimensional parameters. Denote the d-dimensional Lebesgue measure by λ^d.

(4) **Theorem** *Let $(\mathcal{X}_n, \mathcal{A}_n, P_{n,\theta}: \theta \in \Theta)$ be experiments indexed by a measurable subset Θ of $\mathbb{R}^d$. Let $T_n: \mathcal{X}_n \mapsto \mathbb{D}$ and $\kappa_n: \Theta \mapsto \mathbb{D}$ be maps into a complete metric space such that the maps $\theta \mapsto \mathrm{E}_\theta^* f\big(r_{n,\theta}(T_n - \kappa_n(\theta))\big)$ are measurable for every Lipschitz function $f: \mathbb{D} \mapsto [0, 1]$. Suppose that*

$$r_{n,\theta}\big(T_n - \kappa_n(\theta)\big) \overset{\theta}{\rightsquigarrow} L_\theta, \qquad \lambda^d - \text{a.e. } \theta,$$

for probability distributions L_θ supported on a fixed separable, Borel measurable subset of $\mathbb{D}$. Then for any matrices $\Gamma_n \to 0$ there exists a subsequence of $\{n\}$ such that

$$r_{n,\theta+\Gamma_n h}\big(T_n - \kappa_n(\theta + \Gamma_n h)\big) \overset{\theta+\Gamma_n h}{\rightsquigarrow} L_\theta, \qquad \lambda^{2d} - \text{a.e. } (\theta, h),$$

along the subsequence.

(5) **Lemma** *For $n \in \mathbf{N}$ let $g_n, g: \mathbb{R}^d \mapsto [0, 1]$ be arbitrary measurable real functions such that*

$$g_n \to g, \qquad \lambda^d - \text{a.e.}.$$

Then given any sequences of vectors $\gamma_n \to 0$ and matrices $\Gamma_n \to 0$ there exists a subsequence of $\{n\}$ such that

$$\begin{array}{lll} \textit{(i)} & g_n(\theta + \gamma_n) \to g(\theta), & \lambda^d - \text{a.e. } \theta, \\ \textit{(ii)} & g_n(\theta + \Gamma_n h) \to g(\theta), & \lambda^{2d} - \text{a.e. } (\theta, h), \end{array}$$

along the subsequence. If $g_n(\theta + \Gamma_n h_n) - g_n(\theta + \Gamma_n h) \to 0$ for every sequence $h_n \to h$, then (ii) is true for every $h \in \mathbb{R}^d$, for almost every θ.

Proof. We prove statement (ii), the proof of (i) being slightly simpler. We may assume without loss of generality that the function g is integrable; otherwise we first multiply each g_n and g with a suitable, fixed, positive, continuous function. Write p for the standard normal density on $\mathbb{R}^d$. Then

$$\iint \big|g(\theta + \Gamma_n h) - g(\theta)\big| \, p(\theta)\, d\theta \; p(h)\, dh \to 0.$$

This follows since the inner integral converges to zero for every h by the L_1-continuity theorem (e.g. Theorem 8.19 of Wheeden and Zygmund) and next the outer integral converges to zero by the dominated convergence theorem.

If p_n is the density of the $N_d(0, I + \Gamma_n'\Gamma_n)$-distribution, then

$$\iint \left|g_n(\theta + \Gamma_n h) - g(\theta + \Gamma_n h)\right| p(\theta) p(h)\, d\theta dh = \int \left|g_n(u) - g(u)\right| p_n(u)\, du.$$

The sequence p_n converges in L_1 to the standard normal density. Thus the integral on the right converges to zero by the dominated convergence theorem. Combination with the preceding display shows that the sequence of functions $(\theta, h) \mapsto g_n(\theta+\Gamma_n h)-g(\theta)$ converges to zero in mean, and hence in probability, under the standard normal measure. There exists a subsequence along which it converges to zero almost surely.

In the proof of the theorem abbreviate $r_{n,\theta}\big(T_n - \kappa_n(\theta)\big)$ to $T_{n,\theta}$. Assume without loss of generality that $\Theta = \mathbb{R}^d$; otherwise fix θ_0 such that $T_{n,\theta_0} \overset{\theta_0}{\rightsquigarrow} L_{\theta_0}$ and let $P_{n,\theta} = P_{n,\theta_0}$ for every θ not in Θ. Let $\mathbb{D}_0$ be the separable Borel subset of $\mathbb{D}$ on which the limit distributions L_θ concentrate. There exists a countable collection $\mathcal{F}$ of Lipschitz functions $f: \mathbb{D} \mapsto [0, 1]$, depending only on $\mathbb{D}_0$, such that weak convergence of a sequence of maps $T_n: \mathcal{X}_n \mapsto \mathbb{D}$ to a Borel measure L on $\mathbb{D}_0$ is equivalent to $\mathrm{E}^* f(T_n) \to \int f\, dL$ for every $f \in \mathcal{F}$. Consider the functions

$$g_n(\theta;\, f) = \mathrm{E}_\theta^* f(T_{n,\theta}); \qquad g(\theta;\, f) = \int f\, dL_\theta.$$

For every fixed f these functions are measurable by assumption and $g_n \to g$ pointwise. By the lemma there exists a subsequence of $\{n\}$ along which

$$\mathrm{E}^*_{\theta+\Gamma_n h} f(T_{n,\theta+\Gamma_n h}) \to \int f\, dL_\theta, \qquad \lambda^{2d} - \text{a.e.}.$$

The subsequence depends on f, but by a diagonalization scheme we can construct a subsequence for which this is valid for every f in the countable set $\mathcal{F}$. □

(6) **Example** Suppose that for numbers $r_n \to \infty$ the sequence of variables $r_n(T_n - \theta)$ converges under θ to a limit distribution L_θ for almost every θ. The preceding theorem asserts that for almost every θ there exists a set H_θ with $\lambda^d(H_\theta^c) = 0$ such that

$$r_n(T_n - \theta - h/r_n) \overset{\theta+h/r_n}{\rightsquigarrow} L_\theta, \qquad \text{every } h \in H_\theta.$$

Thus the sequence T_n is almost Hájek regular, at almost every θ. The first 'almost' refers to the set H_θ which is almost all of $\mathbb{R}^d$. In most situations this 'almost' can be removed from the statement. It is often the case that

$$\| P_{n,\theta+h_n/r_n} - P_{n,\theta+h/r_n} \| \to 0, \qquad \text{every } h_n \to h.$$

Then the total variation distance between the laws of $T_n - \theta - h/r_n$ under $\theta + h_n/r_n$ and $\theta + h/r_n$ converges to zero as well and in view of the last assertion of Lemma 5 the set H_θ can be taken equal to $\mathbb{R}^d$. □

The combination of Theorem 4 and the proposition gives interesting applications far beyond the Gaussian case. The key is that for almost all θ the limit distribution L_θ of an estimator sequence is not 'better' than the null distribution of the best equivariant estimator in the limit experiment. Also when the latter cannot be characterized as a convolution, the equivariance implies a lower bound on the risk. The locally asymptotically mixed normal case is well-documented in Jeganathan (1982, 1983). We give two other examples.

(7) **Example** Suppose the problem is to estimate θ based on a sample of size n from the uniform distribution P_θ on the interval $[\theta, \theta+1]$. The sequence of experiments $(P^n_{\theta+h/n}: h \in \mathbb{R})$ converges for each θ to the experiment consisting of observing a pair with the same distribution as $(V+h, h-W)$ for independent standard exponential variables V and W. If the sequence $n(T_n-\theta)$ converges in distribution to a limit for every θ, then for almost every θ the limit distribution L_θ is the distribution of an equivariant-in-law estimator T based on $(V+h, h-W)$. The best such estimator in terms of bowl-shaped loss functions is $\frac{1}{2}\big((V+h)+(h-W)\big) = \frac{1}{2}(V-W)+h$. Its invariant standardized law is the Laplace distribution, so that we may conclude

$$\int \ell \, dL_\theta \geq \int \ell(x)\, e^{-2|x|}\, dx, \qquad \lambda^d - \text{a.e. } \theta.$$

In this problem a characterization as a convolution is impossible. This can be seen from the fact that $V+h$ and $\frac{1}{2}(V-W)$ are both equivariant estimators, but their laws have no convolution factor in common. □

(8) **Example** Suppose the problem is to estimate θ based on a sample of size n from the distribution P_θ with density $p(\cdot-\theta)$ on the real line, where $p(x)$ is differentiable at every $x \notin \{a_1, \ldots, a_m\}$ with $\int |p'(x)|\, dx < \infty$ and has discontinuities at each a_i with $p(a_i-) = 0 < p(a_i+)$. Then the sequence $(P^n_{\theta-h/n}: h \in \mathbb{R})$ converges for each θ to the experiment consisting of observing a single random variable with the same distribution as $V+h$ for a standard exponential variable V with mean $1/\sum p(a_i+)$. Since the limit experiment is a full shift experiment, it admits a convolution theorem. If the sequence $n(T_n-\theta)$ converges in distribution to a limit for every θ, then for almost every θ the limit distribution L_θ contains the distribution of V as a convolution factor. □

27.4 Superefficiency and loss functions

The combined results of the preceding section give a deep characterization of the limiting distributions of a sequence of estimators, valid at almost every θ. Apart from measurability the only assumption is the mere existence of limiting distributions.

The latter is a fair assumption for this type of result, but what can be said without it? Equation (1) and asymptotic normality of the maximum likelihood estimator show that for any estimator sequence T_n

$$\liminf_{n\to\infty} \int \mathrm{E}_\theta \ell\big(\sqrt{n}(T_n-\theta)\big)\,\pi(\theta)\,d\theta \geq \iint \ell\,dN_{0,I_\theta^{-1}}\,\pi(\theta)\,d\theta, \tag{9}$$

for most prior densities π. In view of Fatou's lemma this readily gives

$$\limsup_{n\to\infty} \mathrm{E}_\theta \ell\big(\sqrt{n}(T_n-\theta)\big) \geq \int \ell\,dN_{0,I_\theta^{-1}}, \qquad \lambda^d - \text{a.e.}.$$

This cannot be strengthened by replacing the limsup by a liminf, basically because the sequence $\{n\}$ has too many subsequences.

(10) **Example** For the parameter set equal to the unit interval and $k \in \mathbf{N}$ define estimators $T_{2^k+i} = i2^{-k}$ for $i = 1, \ldots, 2^k$. Given a parameter θ define a subsequence of $\{n\}$ by $n_k = 2^k + i_k$, for i_k the integer such that $(i_k-1)2^{-k} < \theta \leq i_k 2^{-k}$. Then $\sqrt{n_k}|T_{n_k}-\theta| \leq \sqrt{2}2^{-k/2}$, whence $\liminf \mathrm{E}_\theta \ell\big(\sqrt{n}(T_n-\theta)\big) = \ell(0)$ for every symmetric loss function which is continuous at zero, and every θ. □

Le Cam (1953) established (9) under regularity conditions somewhat better than those given in Cramér's book. His result was improved in his later papers and also in Strasser (1978). The integral $\int \ell\,dN_{0,I_\theta^{-1}}$ is the minimax risk for estimating h based on a single observation from the $N(h, I_\theta^{-1})$-distribution, the limit of the local experiments around θ. The following theorem establishes a relationship between the limiting pointwise risk and the minimax risk in the local limit experiments in great generality, under very mild conditions. It is assumed that the sequence of experiments $(P_{n,\theta+h/r_n}: h \in \mathbb{R}^d)$ converges for almost every θ to a dominated experiment $\mathcal{E}_\theta$ in which the minimax theorem is valid in the form

$$\sup_P \inf_T \int \mathrm{E}_{\theta,h}\ell(T-h)\,dP(h) = \inf_T \sup_h \mathrm{E}_{\theta,h}\ell(T-h).$$

Here the first supremum is taken over all probability measures with compact support and the infimum over all randomized estimators in $\mathcal{E}_\theta$. This is generally the case, perhaps under some regularity condition on the loss function. Le Cam has broadened the definition of estimators to ensure that the minimax theorem is always true, but we wish to keep the statement simple. According to Le Cam (1973) the local limit experiments are for almost all θ shift-invariant. In Euclidean shift experiments the minimax risk is typically obtained as the limit of Bayes risks for a sequence of uniform priors that approach the improper Lebesgue prior.

(11) **Theorem** *Let $(P_{n,\theta}: \theta \in \Theta)$ be measurable experiments indexed by an open subset $\Theta \subset \mathbb{R}^d$. Suppose that for almost every θ the local experiments $(P_{n,\theta+h/r_n}: h \in \mathbb{R}^d)$ converge to dominated experiments $\mathcal{E}_\theta$ in which the minimax*

theorem holds as mentioned for a given bounded subcompact loss function ℓ. Then for every estimator sequence T_n

$$\limsup_{n\to\infty} \mathrm{E}_\theta \ell\big(r_n(T_n-\theta)\big) \geq \inf_T \sup_h \mathrm{E}_{\theta,h}\ell(T-h), \qquad \lambda^d - \text{a.e. } \theta,$$

where the infimum is taken over all randomized estimators in $\mathcal{E}_\theta$.

Proof. Let π be a probability density on a subset of Θ that is bounded away from the boundary of Θ. Let P be a probability measure with compact support and set $\gamma_n = r_n^{-1}$. Abbreviate $R_n(\theta) = \mathrm{E}_\theta \ell\big(r_n(T_n-\theta)\big)$. By a change of variables

$$\begin{aligned}\Big|\int R_n(\theta)\,\pi(\theta)\,d\theta - \iint R_n(\theta+\gamma_n h)\,dP(h)\,\pi(\theta)\,d\theta\Big| \\ \leq \|\ell\|_\infty \iint \big|\pi(\theta)-\pi(\theta-\gamma_n h)\big|\,d\theta\,dP(h) \to 0,\end{aligned}$$

by the L_1-continuity theorem and the dominated convergence theorem. Essentially as a consequence of Proposition 3, applied to a compactification of $\mathbb{R}^d$,

$$\liminf_{n\to\infty} \int \mathrm{E}_{\theta+\gamma_n h}\ell\big(r_n(T_n-\theta-\gamma_n h)\big)\,dP(h) \geq \inf_T \int \mathrm{E}_{\theta,h}\ell(T-h)\,dP(h),$$

where the infimum is taken over all randomized estimators T for the parameter $h \in \mathbb{R}^d$ in $\mathcal{E}_\theta$. This is valid for almost every θ. Combination of the preceding displays and Fatou's lemma gives

$$\liminf_{n\to\infty} \int \mathrm{E}_\theta \ell\big(r_n(T_n-\theta)\big)\,\pi(\theta)\,d\theta \geq \int \inf_T \int \mathrm{E}_{\theta,h}\ell(T-h)\,dP(h)\,\pi(\theta)\,d\theta.$$

By assumption there exists for each θ and m a probability measure $P_{\theta,m}$ such that the inner integral is within distance $1/m$ of the minimax risk in $\mathcal{E}_\theta$. The desired conclusion follows by the monotone convergence theorem as $m \to \infty$. □

27.5 References

Bahadur, R. R. (1964), 'On Fisher's bound for asymptotic variances', *Annals of Mathematical Statistics* **35**, 1545–1552.

Bernstein, S. (1934), *Theory of Probability*, GTTI, Moscow. (Russian).

Chernoff, H. (1956), 'Large sample theory: parametric case', *Annals of Mathematical Statistics* **27**, 1–22.

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton University Press.

Fisher, R. A. (1912), 'On an absolute criterion for fitting frequency curves', *Messenger of Mathematics* **41**, 155–160.

Fisher, R. A. (1922), 'On the mathematical foundations of theoretical statistics', *Philosophical Transactions of the Royal Society of London, Series A* **222**, 309–368.

Fisher, R. A. (1925), 'Theory of statistical estimation', *Proceedings of the Cambridge Philosophical Society* **22**, 700–725.

Fisher, R. A. (1934), 'Two new properties of mathematical likelihood', *Proceedings of the Royal Society of London, Series A* **144**, 285–307.

Hájek, J. (1970), 'A characterization of limiting distributions of regular estimators', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **14**, 323–330.

Hájek, J. (1972), Local asymptotic minimax and admissibility in estimation, *in* L. Le Cam, J. Neyman & E. L. Scott, eds, 'Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability', Vol. I, University of California Press, Berkeley, pp. 175–194.

Jeganathan, P. (1982), 'On the asymptotic theory of estimation when the limit of the log-likelihood ratios is mixed normal', *Sankhyā: The Indian Journal of Statistics, Series A* **44**, 173–212.

Jeganathan, P. (1983), 'Some asymptotic properties of risk functions when the limit of the experiment is mixed normal', *Sankhyā: The Indian Journal of Statistics, Series A* **45**, 66–87.

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

Le Cam, L. (1956), On the asymptotic theory of estimation and testing hypotheses, *in* J. Neyman, ed., 'Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability', Vol. I, University of California Press, Berkeley, pp. 129–156.

Le Cam, L. (1960), 'Locally asymptotically normal families of distributions', *University of California Publications in Statistics* **3**, 37–98.

Le Cam, L. (1972), Limits of experiments, *in* L. Le Cam, J. Neyman & E. L. Scott, eds, 'Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability', Vol. I, University of California Press, Berkeley, pp. 245–261.

Le Cam, L. (1973), 'Sur les contraintes imposeés par les passages à la limite usuels en statistique', *Proceedings 39th Session of the International Statistical Insitute* **XLV**, 169–177.

Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.

Stein, C. (1956), Inadmissibility of the usual estimator for the mean of a normal distribution, *in* J. Neyman, ed., 'Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability', Vol. I, University of California Press, Berkeley, pp. 197–206.

Strasser, H. (1978), 'Global asymptotic properties of risk functions in estimation', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **45**, 35–48.

van der Vaart, A. (1991), 'An asymptotic representation theorem', *International Statistical Review* **59**, 97–121.

von Mises, R. (1931), *Wahrscheinlichkeitsrechnung*, Springer-Verlag, Berlin.

Wald, A. (1950), *Statistical Decision Functions*, Wiley, New York.
Wolfowitz, J. (1953), 'The method of maximum likelihood and the Wald theory of decision functions', *Indagationes Mathematicae* **15**, 114–119.