8

Efficiency of Estimators

One purpose of asymptotic statistics is to compare the performance of estimators for large sample sizes. This chapter discusses asymptotic lower bounds for estimation in locally asymptotically normal models. These show, among others, in what sense maximum likelihood estimators are asymptotically efficient.

8.1 Asymptotic Concentration

Suppose the problem is to estimate $\psi(\theta)$ based on observations from a model governed by the parameter θ. What is the best asymptotic performance of an estimator sequence T_n for $\psi(\theta)$?

To simplify the situation, we shall in most of this chapter assume that the sequence $\sqrt{n}\big(T_n - \psi(\theta)\big)$ converges in distribution under every possible value of θ. Next we rephrase the question as: What are the best possible limit distributions? In analogy with the Cramér-Rao theorem a "best" limit distribution is referred to as an *asymptotic lower bound.* Under certain restrictions the normal distribution with mean zero and covariance the inverse Fisher information is an asymptotic lower bound for estimating $\psi(\theta) = \theta$ in a smooth parametric model. This is the main result of this chapter, but it needs to be qualified.

The notion of a "best" limit distribution is understood in terms of concentration. If the limit distribution is a priori assumed to be normal, then this is usually translated into asymptotic unbiasedness and minimum variance. The statement that $\sqrt{n}\big(T_n - \psi(\theta)\big)$ converges in distribution to a $N\big(\mu(\theta), \sigma^2(\theta)\big)$-distribution can be roughly understood in the sense that eventually T_n is approximately normally distributed with mean and variance given by

$$\psi(\theta) + \frac{\mu(\theta)}{\sqrt{n}} \quad \text{and} \quad \frac{\sigma^2(\theta)}{n}.$$

Because T_n is meant to estimate $\psi(\theta)$, optimal choices for the asymptotic mean and variance are $\mu(\theta) = 0$ and variance $\sigma^2(\theta)$ as small as possible. These choices ensure not only that the asymptotic mean square error is small but also that the limit distribution $N\big(\mu(\theta), \sigma^2(\theta)\big)$ is maximally concentrated near zero. For instance, the probability of the interval $(-a, a)$ is maximized by choosing $\mu(\theta) = 0$ and $\sigma^2(\theta)$ minimal.

We do not wish to assume a priori that the estimators are asymptotically normal. That normal limits are best will actually be an interesting conclusion. The concentration of a general limit distribution L_θ cannot be measured by mean and variance alone. Instead, we can employ a variety of concentration measures, such as

$$\int x^2\,dL_\theta(x); \qquad \int |x|\,dL_\theta(x); \qquad \int 1\{|x|>a\}\,dL_\theta(x); \qquad \int \bigl(|x|\wedge a\bigr)\,dL_\theta(x).$$

A limit distribution is "good" if quantities of this type are small. More generally, we focus on minimizing $\int \ell\,dL_\theta$ for a given nonnegative function ℓ. Such a function is called a *loss function* and its integral $\int \ell\,dL_\theta$ is the *asymptotic risk* of the estimator. The method of measuring concentration (or rather lack of concentration) by means of loss functions applies to one- and higher-dimensional parameters alike.

The following example shows that a definition of what constitutes asymptotic optimality is not as straightforward as it might seem.

8.1 ***Example (Hodges' estimator).*** Suppose that T_n is a sequence of estimators for a real parameter θ with standard asymptotic behavior in that, for each θ and certain limit distributions L_θ,

$$\sqrt{n}(T_n-\theta) \overset{\theta}{\rightsquigarrow} L_\theta.$$

As a specific example, let T_n be the mean of a sample of size n from the $N(\theta,1)$-distribution. Define a second estimator S_n through

$$S_n = \begin{cases} T_n & \text{if } |T_n| \geq n^{-1/4} \\ 0 & \text{if } |T_n| < n^{-1/4}. \end{cases}$$

If the estimator T_n is already close to zero, then it is changed to exactly zero; otherwise it is left unchanged. The truncation point $n^{-1/4}$ has been chosen in such a way that the limit behavior of S_n is the same as that of T_n for every $\theta \neq 0$, but for $\theta = 0$ there appears to be a great improvement. Indeed, for every r_n,

$$\begin{aligned} r_n S_n &\overset{0}{\rightsquigarrow} 0 \\ \sqrt{n}(S_n-\theta) &\overset{\theta}{\rightsquigarrow} L_\theta, \qquad \theta \neq 0. \end{aligned}$$

To see this, note first that the probability that T_n falls in the interval $(\theta - Mn^{-1/2}, \theta + Mn^{-1/2})$ converges to $L_\theta(-M, M)$ for most M and hence is arbitrarily close to 1 for M and n sufficiently large. For $\theta \neq 0$, the intervals $(\theta - Mn^{-1/2}, \theta + Mn^{-1/2})$ and $(-n^{-1/4}, n^{-1/4})$ are centered at different places and eventually disjoint. This implies that truncation will rarely occur: $P_\theta(T_n = S_n) \to 1$ if $\theta \neq 0$, whence the second assertion. On the other hand the interval $(-Mn^{-1/2}, Mn^{-1/2})$ is contained in the interval $(-n^{-1/4}, n^{-1/4})$ eventually. Hence under $\theta = 0$ we have truncation with probability tending to 1 and hence $P_0(S_n = 0) \to 1$; this is stronger than the first assertion.

At first sight, S_n is an improvement on T_n. For every $\theta \neq 0$ the estimators behave the same, while for $\theta = 0$ the sequence S_n has an "arbitrarily fast" rate of convergence. However, this reasoning is a bad use of asymptotics.

Consider the concrete situation that T_n is the mean of a sample of size n from the normal $N(\theta, 1)$-distribution. It is well known that $T_n = \overline{X}$ is optimal in many ways for every fixed n and hence it ought to be asymptotically optimal also. Figure 8.1 shows why $S_n = \overline{X}1\{|\overline{X}| \geq n^{-1/4}\}$ is no improvement. It shows the graph of the risk function $\theta \mapsto \mathrm{E}_\theta(S_n - \theta)^2$ for three different values of n. These functions are close to 1 on most

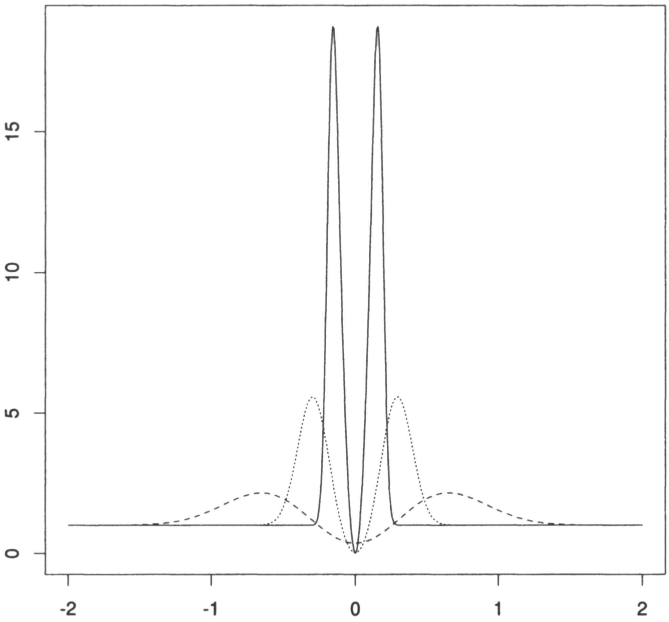

Figure 8.1. Quadratic risk functions of the Hodges estimator based on the means of samples of size 10 (dashed), 100 (dotted), and 1000 (solid) observations from the $N(\theta, 1)$-distribution.

of the domain but possess peaks close to zero. As $n \to \infty$, the locations and widths of the peaks converge to zero but their heights to infinity. The conclusion is that S_n "buys" its better asymptotic behavior at $\theta = 0$ at the expense of erratic behavior close to zero. Because the values of θ at which S_n is bad differ from n to n, the erratic behavior is not visible in the pointwise limit distributions under fixed θ. □

8.2 Relative Efficiency

In order to choose between two estimator sequences, we compare the concentration of their limit distributions. In the case of normal limit distributions and convergence rate $\sqrt{n}$, the quotient of the asymptotic variances is a good numerical measure of their relative efficiency. This number has an attractive interpretation in terms of the numbers of observations needed to attain the same goal with each of two sequences of estimators.

Let $\nu \to \infty$ be a "time" index, and suppose that it is required that, as $\nu \to \infty$, our estimator sequence attains mean zero and variance 1 (or $1/\nu$). Assume that an estimator T_n based on n observations has the property that, as $n \to \infty$,

$$\sqrt{n}\big(T_n - \psi(\theta)\big) \overset{\theta}{\rightsquigarrow} N\big(0, \sigma^2(\theta)\big).$$

Then the requirement is to use at time ν an appropriate number n_ν of observations such that, as $\nu \to \infty$,

$$\sqrt{\nu}\big(T_{n_\nu} - \psi(\theta)\big) \overset{\theta}{\rightsquigarrow} N(0, 1).$$

Given two available estimator sequences, let $n_{\nu,1}$ and $n_{\nu,2}$ be the numbers of observations needed to meet the requirement with each of the estimators. Then, if it exists, the limit

$$\lim_{\nu\to\infty}\frac{n_{\nu,2}}{n_{\nu,1}}$$

is called the *relative efficiency* of the estimators. (In general, it depends on the parameter θ.)

Because $\sqrt{\nu}\big(T_{n_\nu}-\psi(\theta)\big)$ can be written as $\sqrt{\nu/n_\nu}\sqrt{n_\nu}\big(T_{n_\nu}-\psi(\theta)\big)$, it follows that necessarily $n_\nu\to\infty$, and also that $n_\nu/\nu\to\sigma^2(\theta)$. Thus, the relative efficiency of two estimator sequences with asymptotic variances $\sigma_i^2(\theta)$ is just

$$\lim_{\nu\to\infty}\frac{n_{\nu,2}/\nu}{n_{\nu,1}/\nu}=\frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}.$$

If the value of this quotient is bigger than 1, then the second estimator sequence needs proportionally that many observations more than the first to achieve the same (asymptotic) precision.

8.3 Lower Bound for Experiments

It is certainly impossible to give a nontrivial lower bound on the limit distribution of a standardized estimator $\sqrt{n}\big(T_n-\psi(\theta)\big)$ for a single θ. Hodges' example shows that it is not even enough to consider the behavior under every θ, pointwise for all θ. Different values of the parameters must be taken into account simultaneously when taking the limit as $n\to\infty$. We shall do this by studying the performance of estimators under parameters in a "shrinking" neighborhood of a fixed θ.

We consider parameters $\theta+h/\sqrt{n}$ for θ fixed and h ranging over $\mathbb{R}^k$ and suppose that, for certain limit distributions $L_{\theta,h}$,

$$\sqrt{n}\left(T_n-\psi\left(\theta+\frac{h}{\sqrt{n}}\right)\right)\overset{\theta+h/\sqrt{n}}{\rightsquigarrow}L_{\theta,h},\qquad\text{every } h. \tag{8.2}$$

Then T_n can be considered a good estimator for $\psi(\theta)$ if the limit distributions $L_{\theta,h}$ are maximally concentrated near zero. If they are maximally concentrated for every h and some fixed θ, then T_n can be considered locally optimal at θ. Unless specified otherwise, we assume in the remainder of this chapter that the parameter set Θ is an open subset of $\mathbb{R}^k$, and that ψ maps Θ into $\mathbb{R}^m$. The derivative of $\theta\mapsto\psi(\theta)$ is denoted by $\dot\psi_\theta$.

Suppose that the observations are a sample of size n from a distribution P_θ. If P_θ depends smoothly on the parameter, then

$$\left(P^n_{\theta+h/\sqrt{n}}:h\in\mathbb{R}^k\right)\rightsquigarrow\left(N\big(h,I_\theta^{-1}\big):h\in\mathbb{R}^k\right),$$

as experiments, in the sense of Theorem 7.10. This theorem shows which limit distributions are possible and can be specialized to the estimation problem in the following way.

8.3 *Theorem.* *Assume that the experiment $(P_\theta:\theta\in\Theta)$ is differentiable in quadratic mean (7.1) at the point θ with nonsingular Fisher information matrix I_θ. Let ψ be differentiable at θ. Let T_n be estimators in the experiments $(P^n_{\theta+h/\sqrt{n}}:h\in\mathbb{R}^k)$ such that*

(8.2) *holds for every h. Then there exists a randomized statistic T in the experiment $\big(N(h, I_\theta^{-1}) : h \in \mathbb{R}^k\big)$ such that $T - \dot\psi_\theta h$ has distribution $L_{\theta,h}$ for every h.*

Proof. Apply Theorem 7.10 to $S_n = \sqrt{n}\big(T_n - \psi(\theta)\big)$. In view of the definition of $L_{\theta,h}$ and the differentiability of ψ, the sequence

$$S_n = \sqrt{n}\left(T_n - \psi\left(\theta + \frac{h}{\sqrt{n}}\right)\right) + \sqrt{n}\left(\psi\left(\theta + \frac{h}{\sqrt{n}}\right) - \psi(\theta)\right)$$

converges in distribution under h to $L_{\theta,h} * \delta_{\dot\psi_\theta h}$, where $*\delta_h$ denotes a translation by h. According to Theorem 7.10, there exists a randomized statistic T in the normal experiment such that T has distribution $L_{\theta,h} * \delta_{\dot\psi_\theta h}$ for every h. This satisfies the requirements. ■

This theorem shows that for most estimator sequences T_n there is a randomized estimator T such that the distribution of $\sqrt{n}\big(T_n - \psi(\theta + h/\sqrt{n})\big)$ under $\theta + h/\sqrt{n}$ is, for large n, approximately equal to the distribution of $T - \dot\psi_\theta h$ under h. Consequently the standardized distribution of the best possible estimator T_n for $\psi(\theta + h/\sqrt{n})$ is approximately equal to the standardized distribution of the best possible estimator T for $\dot\psi_\theta h$ in the limit experiment. If we know the best estimator T for $\dot\psi_\theta h$, then we know the "locally best" estimator sequence T_n for $\psi(\theta)$.

In this way, the asymptotic optimality problem is reduced to optimality in the experiment based on one observation X from a $N(h, I_\theta^{-1})$-distribution, in which θ is known and h ranges over $\mathbb{R}^k$. This experiment is simple and easy to analyze. The observation itself is the customary estimator for its expectation h, and the natural estimator for $\dot\psi_\theta h$ is $\dot\psi_\theta X$. This has several optimality properties: It is minimum variance unbiased, minimax, best equivariant, and Bayes with respect to the noninformative prior. Some of these properties are reviewed in the next section.

Let us agree, at least for the moment, that $\dot\psi_\theta X$ is a "best" estimator for $\dot\psi_\theta h$. The distribution of $\dot\psi_\theta X - \dot\psi_\theta h$ is normal with zero mean and covariance $\dot\psi_\theta I_\theta^{-1} \dot\psi_\theta{}^T$ for every h. The parameter $h = 0$ in the limit experiment corresponds to the parameter θ in the original problem. We conclude that the "best" limit distribution of $\sqrt{n}\big(T_n - \psi(\theta)\big)$ under θ is the $N(0, \dot\psi_\theta I_\theta^{-1} \dot\psi_\theta{}^T)$-distribution.

This is the main result of the chapter. The remaining sections discuss several ways of making this reasoning more rigorous. Because the expression $\dot\psi_\theta I_\theta^{-1} \dot\psi_\theta{}^T$ is precisely the Cramér-Rao lower bound for the covariance of unbiased estimators for $\psi(\theta)$, we can think of the results of this chapter as asymptotic Cramér-Rao bounds. This is helpful, even though it does not do justice to the depth of the present results. For instance, the Cramér-Rao bound in no way suggests that normal limiting distributions are best. Also, it is not completely true that an $N(h, I_\theta^{-1})$-distribution is "best" (see section 8.8). We shall see exactly to what extent the optimality statement is false.

8.4 Estimating Normal Means

According to the preceding section, the asymptotic optimality problem reduces to optimality in a normal location (or "Gaussian shift") experiment. This section has nothing to do with asymptotics but reviews some facts about Gaussian models.

Based on a single observation X from a $N(h, \Sigma)$-distribution, it is required to estimate Ah for a given matrix A. The covariance matrix Σ is assumed known and nonsingular. It is well known that AX is minimum variance unbiased. It will be shown that AX is also best-equivariant and minimax for many loss functions.

A randomized estimator T is called *equivariant-in-law* for estimating Ah if the distribution of $T - Ah$ under h does not depend on h. An example is the estimator AX, whose "invariant law" (the law of $AX - Ah$ under h) is the $N(0, A\Sigma A^T)$-distribution. The following proposition gives an interesting characterization of the law of general equivariant-in-law estimators: These are distributed as the sum of AX and an independent variable.

8.4 ***Proposition.*** *The null distribution L of any randomized equivariant-in-law estimator of Ah can be decomposed as $L = N(0, A\Sigma A^T) * M$ for some probability measure M. The only randomized equivariant-in-law estimator for which M is degenerate at* 0 *is AX.*

The measure M can be interpreted as the distribution of a noise factor that is added to the estimator AX. If no noise is best, then it follows that AX is best equivariant-in-law.

A more precise argument can be made in terms of loss functions. In general, convoluting a measure with another measure decreases its concentration. This is immediately clear in terms of variance: The variance of a sum of two independent variables is the sum of the variances, whence convolution increases variance. For normal measures this extends to all "bowl-shaped" symmetric loss functions. The name should convey the form of their graph. Formally, a function is defined to be *bowl-shaped* if the sublevel sets $\{x : \ell(x) \leq c\}$ are convex and symmetric about the origin; it is called *subconvex* if, moreover, these sets are closed. A *loss function* is any function with values in $[0, \infty)$. The following lemma quantifies the loss in concentration under convolution (for a proof, see, e.g., [80] or [114].)

8.5 ***Lemma (Anderson's lemma).*** *For any bowl-shaped loss function ℓ on $\mathbb{R}^k$, every probability measure M on $\mathbb{R}^k$, and every covariance matrix Σ*

$$\int \ell \, dN(0, \Sigma) \leq \int \ell \, d\big[N(0, \Sigma) * M\big].$$

Next consider the *minimax criterion*. According to this criterion the "best" estimator, relative to a given loss function, minimizes the maximum risk

$$\sup_h \mathrm{E}_h \ell(T - Ah),$$

over all (randomized) estimators T. For every bowl-shaped loss function ℓ, this leads again to the estimator AX.

8.6 ***Proposition.*** *For any bowl-shaped loss function ℓ, the maximum risk of any randomized estimator T of Ah is bounded below by $\mathrm{E}_0\ell(AX)$. Consequently, AX is a minimax estimator for Ah. If Ah is real and $\mathrm{E}_0(AX)^2\ell(AX) < \infty$, then AX is the only minimax estimator for Ah up to changes on sets of probability zero.*

Proofs. For a proof of the uniqueness of the minimax estimator, see [18] or [80]. We prove the other assertions for subconvex loss functions, using a Bayesian argument.

Let H be a random vector with a normal $N(0, \Lambda)$-distribution, and consider the original $N(h, \Sigma)$-distribution as the conditional distribution of X given $H = h$. The randomization variable U in $T(X, U)$ is constructed independently of the pair (X, H). In this notation, the distribution of the variable $T - AH$ is equal to the "average" of the distributions of $T - Ah$ under the different values of h in the original set-up, averaged over h using a $N(0, \Lambda)$-"prior distribution."

By a standard calculation, we find that the "a posteriori" distribution, the distribution of H given X, is the normal distribution with mean $(\Sigma^{-1} + \Lambda^{-1})^{-1}\Sigma^{-1}X$ and covariance matrix $(\Sigma^{-1} + \Lambda^{-1})^{-1}$. Define the random vectors

$$W_\Lambda = T - A(\Sigma^{-1} + \Lambda^{-1})^{-1}\Sigma^{-1}X, \quad G_\Lambda = -A\big(H - (\Sigma^{-1} + \Lambda^{-1})^{-1}\Sigma^{-1}X\big)$$

These vectors are independent, because W_Λ is a function of (X, U) only, and the conditional distribution of G_Λ given X is normal with mean 0 and covariance matrix $A(\Sigma^{-1} + \Lambda^{-1})^{-1}A^T$, independent of X. As $\Lambda = \lambda I$ for a scalar $\lambda \to \infty$, the sequence G_Λ converges in distribution to a $N(0, A\Sigma A^T)$-distributed vector G. The sum of the two vectors yields $T - AH$, for every Λ.

Because a supremum is larger than an average, we obtain, where on the left we take the expectation with respect to the original model,

$$\sup_h \mathrm{E}_h \ell(T - Ah) \geq \mathrm{E}\ell(T - AH) = \mathrm{E}\ell(G_\Lambda + W_\Lambda) \geq \mathrm{E}\ell(G_\Lambda),$$

by Anderson's lemma. This is true for every Λ. The lim inf of the right side as $\Lambda \to \infty$ is at least $\mathrm{E}\ell(G)$, by the portmanteau lemma. This concludes the proof that AX is minimax.

If T is equivariant-in-law with invariant law L, then the distribution of $G_\Lambda + W_\Lambda = T - AH$ is L, for every Λ. It follows that

$$\int e^{it^Tx}\, dL(x) = \mathrm{E}e^{it^TG_\Lambda}\mathrm{E}e^{it^TW_\Lambda}, \qquad \text{every } t.$$

As $\Lambda \to \infty$, the left side remains fixed; the first factor on the right side converges to the characteristic function of G, which is positive. Conclude that the characteristic functions of W_Λ converge to a continuous function, whence W_Λ converges in distribution to some vector W, by Lévy's continuity theorem. By the independence of G_Λ and W_Λ for every Λ, the sequence (G_Λ, W_Λ) converges in distribution to a pair (G, W) of independent vectors with marginal distributions as before. Next, by the continuous-mapping theorem, the distribution of $G_\Lambda + W_\Lambda$, which is fixed at L, "converges" to the distribution of $G + W$. This proves that L can be written as a convolution, as claimed in Proposition 8.4.

If T is an equivariant-in-law estimator and $\tilde{T}(X) = \mathrm{E}\big(T(X, U)|\, X\big)$, then

$$\mathrm{E}_h(\tilde{T} - AX) = \mathrm{E}_h(T - AX) = \mathrm{E}_h(T - Ah) - \mathrm{E}_h(AX - Ah)$$

is independent of h. By the completeness of the normal location family, we conclude that $\tilde{T} - AX$ is constant, almost surely. If T has the same law as AX, then the constant is zero. Furthermore, T must be equal to its projection $\tilde{T}$ almost surely, because otherwise it would have a bigger second moment than $\tilde{T} = AX$. Thus $T = AX$ almost surely. ■

8.5 Convolution Theorem

An estimator sequence T_n is called *regular* at θ for estimating a parameter $\psi(\theta)$ if, for every h,

$$\sqrt{n}\left(T_n - \psi\left(\theta + \frac{h}{\sqrt{n}}\right)\right) \overset{\theta+h/\sqrt{n}}{\rightsquigarrow} L_\theta.$$

The probability measure L_θ may be arbitrary but should be the same for every h.

A regular estimator sequence attains its limit distribution in a "locally uniform" manner. This type of regularity is common and is often considered desirable: A small change in the parameter should not change the distribution of the estimator too much; a disappearing small change should not change the (limit) distribution at all. However, some estimator sequences of interest, such as shrinkage estimators, are not regular.

In terms of the limit distributions $L_{\theta,h}$ in (8.2), regularity is exactly that all $L_{\theta,h}$ are equal, for the given θ. According to Theorem 8.3, every estimator sequence is matched by an estimator T in the limit experiment $\bigl(N(h, I_\theta^{-1}) : h \in \mathbb{R}^k\bigr)$. For a regular estimator sequence this matching estimator has the property

$$T - \dot\psi_\theta h \overset{h}{\sim} L_\theta, \qquad \text{every } h. \tag{8.7}$$

Thus a regular estimator sequence is matched by an equivariant-in-law estimator for $\dot\psi_\theta h$. A more informative name for "regular" is *asymptotically equivariant-in-law.*

It is now easy to determine a best estimator sequence from among the regular estimator sequences (a *best regular* sequence): It is the sequence T_n that corresponds to the best equivariant-in-law estimator T for $\dot\psi_\theta h$ in the limit experiment, which is $\dot\psi_\theta X$ by Proposition 8.4. The best possible limit distribution of a regular estimator sequence is the law of this estimator, a $N(0, \dot\psi_\theta I_\theta^{-1} \dot\psi_\theta{}^T)$-distribution.

The characterization as a convolution of the invariant laws of equivariant-in-law estimators carries over to the asymptotic situation.

8.8 ***Theorem (Convolution).*** *Assume that the experiment $(P_\theta : \theta \in \Theta)$ is differentiable in quadratic mean* (7.1) *at the point θ with nonsingular Fisher information matrix I_θ. Let ψ be differentiable at θ. Let T_n be an at θ regular estimator sequence in the experiments $(P_\theta^n : \theta \in \Theta)$ with limit distribution L_θ. Then there exists a probability measure M_θ such that*

$$L_\theta = N\bigl(0, \dot\psi_\theta I_\theta^{-1} \dot\psi_\theta{}^T\bigr) * M_\theta.$$

In particular, if L_θ has covariance matrix Σ_θ, then the matrix $\Sigma_\theta - \dot\psi_\theta I_\theta^{-1} \dot\psi_\theta{}^T$ is nonnegative-definite.

Proof. Apply Theorem 8.3 to conclude that L_θ is the distribution of an equivariant-in-law estimator T in the limit experiment, satisfying (8.7). Next apply Proposition 8.4. ■

8.6 Almost-Everywhere Convolution Theorem

Hodges' example shows that there is no hope for a nontrivial lower bound for the limit distribution of a standardized estimator sequence $\sqrt{n}\bigl(T_n - \psi(\theta)\bigr)$ for *every* θ. It is always possible to improve on a given estimator sequence for selected parameters. In this section it is shown that improvement over an $N(0, \dot\psi_\theta I_\theta^{-1} \dot\psi_\theta{}^T)$-distribution can be made on at most a Lebesgue null set of parameters. Thus the possibilities for improvement are very much restricted.

8.9 ***Theorem.*** *Assume that the experiment $(P_\theta : \theta \in \Theta)$ is differentiable in quadratic mean* (7.1) *at every θ with nonsingular Fisher information matrix I_θ. Let ψ be differentiable at every θ. Let T_n be an estimator sequence in the experiments $(P_\theta^n : \theta \in \Theta)$ such that $\sqrt{n}\big(T_n - \psi(\theta)\big)$ converges to a limit distribution L_θ under every θ. Then there exist probability distributions M_θ such that for Lebesgue almost every θ*

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

In particular, if L_θ has covariance matrix Σ_θ, then the matrix $\Sigma_\theta - \dot\psi_\theta I_\theta^{-1} \dot\psi_\theta{}^T$ is nonnegative definite for Lebesgue almost every θ.

The theorem follows from the convolution theorem in the preceding section combined with the following remarkable lemma. Any estimator sequence with limit distributions is *automatically* regular at almost every θ along a subsequence of $\{n\}$.

8.10 ***Lemma.*** *Let T_n be estimators in experiments $(P_{n,\theta} : \theta \in \Theta)$ indexed by a measurable subset Θ of $\mathbb{R}^k$. Assume that the map $\theta \mapsto P_{n,\theta}(A)$ is measurable for every measurable set A and every n, and that the map $\theta \mapsto \psi(\theta)$ is measurable. Suppose that there exist distributions L_θ such that for Lebesgue almost every θ*

$$r_n\big(T_n - \psi(\theta)\big) \overset{\theta}{\rightsquigarrow} L_\theta.$$

Then for every $\gamma_n \to 0$ there exists a subsequence of $\{n\}$ such that, for Lebesgue almost every (θ, h), along the subsequence,

$$r_n\big(T_n - \psi(\theta + \gamma_n h)\big) \overset{\theta+\gamma_n h}{\rightsquigarrow} L_\theta.$$

Proof. Assume without loss of generality that $\Theta = \mathbb{R}^k$; otherwise, fix some θ_0 and let $P_{n,\theta} = P_{n,\theta_0}$ for every θ not in Θ. Write $T_{n,\theta} = r_n\big(T_n - \psi(\theta)\big)$. There exists a countable collection $\mathcal{F}$ of uniformly bounded, left- or right-continuous functions f such that weak convergence of a sequence of maps T_n is equivalent to $\mathrm{E} f(T_n) \to \int f\, dL$ for every $f \in \mathcal{F}$.[†] Suppose that for every f there exists a subsequence of $\{n\}$ along which

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

Even in case the subsequence depends on f, we can, by a diagonalization scheme, construct a subsequence for which this is valid for every f in the countable set $\mathcal{F}$. Along this subsequence we have the desired convergence.

[†] For continuous distributions L we can use the indicator functions of cells $(-\infty, c]$ with c ranging over Q^k. For general L replace every such indicator by an approximating sequence of continuous functions. Alternatively, see, e.g., Theorem 1.12.2 in [146]. Also see Lemma 2.25.

Setting $g_n(\theta) = \mathrm{E}_\theta f(T_{n,\theta})$ and $g(\theta) = \int f\, dL_\theta$, we see that the lemma is proved once we have established the following assertion: Every sequence of bounded, measurable functions g_n that converges almost everywhere to a limit g, has a subsequence along which

$$g_n(\theta + \gamma_n h) \to g(\theta), \qquad \lambda^{2k} - \text{a.e. } (\theta, h).$$

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. It should also be verified that, under our conditions, the functions g_n are measurable.

Write p for the standard normal density on $\mathbb{R}^k$ and p_n for the density of the $N(0, I + \gamma_n^2 I)$-distribution. By Scheffé's lemma, the sequence p_n converges to p in L_1. Let Θ and H denote independent standard normal vectors. Then, by the triangle inequality and the dominated-convergence theorem,

$$\mathrm{E}\big|g_n(\Theta + \gamma_n H) - g(\Theta + \gamma_n H)\big| = \int \big|g_n(u) - g(u)\big| p_n(u)\, du \to 0.$$

Secondly for any fixed continuous and bounded function g_ε the sequence $\mathrm{E}\big|g_\varepsilon(\Theta + \gamma_n H) - g_\varepsilon(\Theta)\big|$ converges to zero as $n \to \infty$ by the dominated convergence theorem. Thus, by the triangle inequality, we obtain

$$\begin{aligned}\mathrm{E}\big|g(\Theta + \gamma_n H) - g(\Theta)\big| &\le \int |g - g_\varepsilon|(u)\,(p_n + p)(u)\, du + o(1)\\ &= 2\int |g - g_\varepsilon|(u)\, p(u)\, du + o(1).\end{aligned}$$

Because any measurable integrable function g can be approximated arbitrarily closely in L_1 by continuous functions, the first term on the far right side can be made arbitrarily small by choice of g_ε. Thus the left side converges to zero.

By combining this with the preceding display, we see that $\mathrm{E}\big|g_n(\Theta + \gamma_n H) - g(\Theta)\big| \to 0$. In other words, 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. ∎

*8.7 Local Asymptotic Minimax Theorem

The convolution theorems discussed in the preceding sections are not completely satisfying. The convolution theorem designates a best estimator sequence among the regular estimator sequences, and thus imposes an a priori restriction on the set of permitted estimator sequences. The almost-everywhere convolution theorem imposes no (serious) restriction but yields no information about some parameters, albeit a null set of parameters.

This section gives a third attempt to "prove" that the normal $N(0, \dot\psi_\theta I_\theta^{-1} \dot\psi_\theta{}^T)$-distribution is the best possible limit. It is based on the minimax criterion and gives a lower bound for the maximum risk over a small neighborhood of a parameter θ. In fact, it bounds the expression

$$\lim_{\delta\to 0} \liminf_{n\to\infty} \sup_{\|\theta'-\theta\|<\delta} \mathrm{E}_{\theta'} \ell\Big(\sqrt{n}\big(T_n - \psi(\theta')\big)\Big).$$

This is the asymptotic maximum risk over an arbitrarily small neighborhood of θ. The following theorem concerns an even more refined (and smaller) version of the local maximum risk.

8.11 ***Theorem.*** *Let the experiment $(P_\theta : \theta \in \Theta)$ be differentiable in quadratic mean (7.1) at θ with nonsingular Fisher information matrix I_θ. Let ψ be differentiable at θ. Let T_n be any estimator sequence in the experiments $(P_\theta^n : \theta \in \mathbb{R}^k)$. Then for any bowl-shaped loss function ℓ*

$$\sup_I \liminf_{n\to\infty} \sup_{h\in I} \mathrm{E}_{\theta+h/\sqrt{n}}\ell\left(\sqrt{n}\left(T_n - \psi\left(\theta + \frac{h}{\sqrt{n}}\right)\right)\right) \geq \int \ell\, dN\left(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta{}^T\right).$$

Here the first supremum is taken over all finite subsets I of $\mathbb{R}^k$.

Proof. We only give the proof under the further assumptions that the sequence $\sqrt{n}\big(T_n - \psi(\theta)\big)$ is uniformly tight under θ and that ℓ is (lower) semicontinuous.[†] Then Prohorov's theorem shows that every subsequence of $\{n\}$ has a further subsequence along which the vectors

$$\left(\sqrt{n}\big(T_n - \psi(\theta)\big), \frac{1}{\sqrt{n}} \sum \dot{\ell}_\theta(X_i)\right)$$

converge in distribution to a limit under θ. By Theorem 7.2 and Le Cam's third lemma, the sequence $\sqrt{n}\big(T_n - \psi(\theta)\big)$ converges in law also under every $\theta + h/\sqrt{n}$ along the subsequence. By differentiability of ψ, the same is true for the sequence $\sqrt{n}\big(T_n - \psi(\theta + h/\sqrt{n})\big)$, whence (8.2) is satisfied. By Theorem 8.3, the distributions $L_{\theta,h}$ are the distributions of $T - \dot{\psi}_\theta h$ under h for a randomized estimator T based on an $N(h, I_\theta^{-1})$-distributed observation. By Proposition 8.6,

$$\sup_{h\in\mathbb{R}^k} E_h \ell(T - \dot{\psi}_\theta h) \geq E_0 \ell(\dot{\psi}_\theta X) = \int \ell\, dN\left(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta{}^T\right).$$

It suffices to show that the left side of this display is a lower bound for the left side of the theorem.

The complicated construction that defines the asymptotic minimax risk (the lim inf sandwiched between two suprema) requires that we apply the preceding argument to a carefully chosen subsequence. Place the rational vectors in an arbitrary order, and let I_k consist of the first k vectors in this sequence. Then the left side of the theorem is larger than

$$R := \lim_{k\to\infty} \liminf_{n\to\infty} \sup_{h\in I_k} \mathrm{E}_{\theta+h/\sqrt{n}}\ell\left(\sqrt{n}\left(T_n - \psi\left(\theta + \frac{h}{\sqrt{n}}\right)\right)\right).$$

There exists a subsequence $\{n_k\}$ of $\{n\}$ such that this expression is equal to

$$\lim_{k\to\infty} \sup_{h\in I_k} E_{\theta+h/\sqrt{n_k}}\ell\left(\sqrt{n_k}\left(T_{n_k} - \psi\left(\theta + \frac{h}{\sqrt{n_k}}\right)\right)\right).$$

We apply the preceding argument to this subsequence and find a further subsequence along which T_n satisfies (8.2). For simplicity of notation write this as $\{n'\}$ rather than with a double subscript. Because ℓ is nonnegative and lower semicontinuous, the portmanteau lemma gives, for every h,

$$\liminf_{n'\to\infty} E_{\theta+h/\sqrt{n'}}\ell\left(\sqrt{n'}\left(T_{n'} - \psi\left(\theta + \frac{h}{\sqrt{n'}}\right)\right)\right) \geq \int \ell\, dL_{\theta,h}.$$

[†] See, for example, [146, Chapter 3.11] for the general result, which can be proved along the same lines, but using a compactification device to induce tightness.

Every rational vector h is contained in I_k for every sufficiently large k. Conclude that

$$R \geq \sup_{h \in \mathbb{Q}^k} \int \ell \, dL_{\theta,h} = \sup_{h \in \mathbb{Q}^k} E_h \ell(T - \dot{\psi}_\theta h).$$

The risk function in the supremum on the right is lower semicontinuous in h, by the continuity of the Gaussian location family and the lower semicontinuity of ℓ. Thus the expression on the right does not change if $\mathbb{Q}^k$ is replaced by $\mathbb{R}^k$. This concludes the proof. ■

*8.8 Shrinkage Estimators

The theorems of the preceding sections seem to prove in a variety of ways that the best possible limit distribution is the $N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta{}^T)$-distribution. At closer inspection, the situation is more complicated, and to a certain extent optimality remains a matter of taste, asymptotic optimality being no exception. The "optimal" normal limit is the distribution of the estimator $\dot{\psi}_\theta X$ in the normal limit experiment. Because this estimator has several optimality properties, many statisticians consider it best. Nevertheless, one might prefer a Bayes estimator or a shrinkage estimator. With a changed perception of what constitutes "best" in the limit experiment, the meaning of "asymptotically best" changes also. This becomes particularly clear in the example of shrinkage estimators.

8.12 ***Example (Shrinkage estimator).*** Let $X_1, \ldots, X_n$ be a sample from a multivariate normal distribution with mean θ and covariance the identity matrix. The dimension k of the observations is assumed to be at least 3. This is essential! Consider the estimator

$$T_n = \overline{X}_n - (k-2) \frac{\overline{X}_n}{n \|\overline{X}_n\|^2}.$$

Because $\overline{X}_n$ converges in probability to the mean θ, the second term in the definition of T_n is $O_P(n^{-1})$ if $\theta \neq 0$. In that case $\sqrt{n}(T_n - \overline{X}_n)$ converges in probability to zero, whence the estimator sequence T_n is regular at every $\theta \neq 0$. For $\theta = h/\sqrt{n}$, the variable $\sqrt{n}\overline{X}_n$ is distributed as a variable X with an $N(h, I)$-distribution, and for every n the standardized estimator $\sqrt{n}(T_n - h/\sqrt{n})$ is distributed as $T - h$ for

$$T(X) = X - (k-2) \frac{X}{\|X\|^2}.$$

This is the Stein *shrinkage estimator*. Because the distribution of $T - h$ depends on h, the sequence T_n is not regular at $\theta = 0$. The Stein estimator has the remarkable property that, for every h (see, e.g., [99, p. 300]),

$$E_h \|T - h\|^2 < \mathrm{E}_h \|X - h\|^2 = k.$$

It follows that, in terms of joint quadratic loss $\ell(x) = \|x\|^2$, the local limit distributions $L_{0,h}$ of the sequence $\sqrt{n}(T_n - h/\sqrt{n})$ under $\theta = h/\sqrt{n}$ are all better than the $N(0, I)$-limit distribution of the best regular estimator sequence $\overline{X}_n$. □

The example of shrinkage estimators shows that, depending on the optimality criterion, a normal $N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta{}^T)$-limit distribution need not be optimal. In this light, is it reasonable to uphold that maximum likelihood estimators are asymptotically optimal? Perhaps not. On the other hand, the possibility of improvement over the $N(0, \dot\psi_\theta I_\theta^{-1}\dot\psi_\theta{}^T)$-limit is restricted in two important ways.

First, improvement can be made only on a null set of parameters by Theorem 8.9. Second, improvement is possible only for special loss functions, and improvement for one loss function necessarily implies worse performance for other loss functions. This follows from the next lemma.

Suppose that we require the estimator sequence to be *locally asymptotically minimax* for a given loss function ℓ in the sense that

$$\sup_I \limsup_{n\to\infty} \sup_{h\in I} E_{\theta+h/\sqrt{n}}\ell\left(\sqrt{n}\left(T_n - \psi\left(\theta + \frac{h}{\sqrt{n}}\right)\right)\right) \le \int \ell\, dN\left(0, \dot\psi_\theta I_\theta^{-1}\dot\psi_\theta{}^T\right).$$

This is a reasonable requirement, and few statisticians would challenge it. The following lemma shows that for one-dimensional parameters $\psi(\theta)$ local asymptotic minimaxity for even a single loss function implies regularity. Thus, if it is required that all coordinates of a certain estimator sequence be locally asymptotically minimax for some loss function, then the best regular estimator sequence is optimal without competition.

8.13 Lemma. *Assume that the experiment* $(P_\theta : \theta \in \Theta)$ *is differentiable in quadratic mean* (7.1) *at θ with nonsingular Fisher information matrix I_θ. Let ψ be a real-valued map that is differentiable at θ. Then an estimator sequence in the experiments $(P_\theta^n : \theta \in \mathbb{R}^k)$ can be locally asymptotically minimax at θ for a bowl-shaped loss function ℓ such that $0 < \int x^2 \ell(x)\, dN(0, \dot\psi_\theta I_\theta^{-1}\dot\psi_\theta{}^T)(x) < \infty$ only if T_n is best regular at θ.*

Proof. We only give the proof under the further assumption that the sequence $\sqrt{n}\big(T_n - \psi(\theta)\big)$ is uniformly tight under θ. Then by the same arguments as in the proof of Theorem 8.11, every subsequence of $\{n\}$ has a further subsequence along which the sequence $\sqrt{n}\big(T_n - \psi(\theta + h/\sqrt{n})\big)$ converges in distribution under $\theta + h/\sqrt{n}$ to the distribution $L_{\theta,h}$ of $T - \dot\psi_\theta h$ under h, for a randomized estimator T based on an $N(h, I_\theta^{-1})$-distributed observation. Because T_n is locally asymptotically minimax, it follows that

$$\sup_{h\in\mathbb{R}^k} E_h \ell(T - \dot\psi_\theta h) = \sup_{h\in\mathbb{R}^k} \int \ell\, dL_{\theta,h} \le \int \ell\, dN\left(0, \dot\psi_\theta I_\theta^{-1}\dot\psi_\theta{}^T\right).$$

Thus T is a minimax estimator for $\dot\psi_\theta h$ in the limit experiment. By Proposition 8.6, $T = \dot\psi_\theta X$, whence $L_{\theta,h}$ is independent of h. ∎

*8.9 Achieving the Bound

If the convolution theorem is taken as the basis for asymptotic optimality, then an estimator sequence is best if it is asymptotically regular with a $N(0, \dot\psi_\theta I_\theta^{-1}\dot\psi_\theta{}^T)$-limit distribution. An estimator sequence has this property if and only if the estimator is asymptotically linear in the score function.

8.14 Lemma. *Assume that the experiment* $(P_\theta : \theta \in \Theta)$ *is differentiable in quadratic mean* (7.1) *at θ with nonsingular Fisher information matrix I_θ. Let ψ be differentiable at*

θ. Let T_n be an estimator sequence in the experiments $(P_\theta^n : \theta \in \mathbb{R}^k)$ such that

$$\sqrt{n}\big(T_n - \psi(\theta)\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \dot\psi_\theta I_\theta^{-1}\dot\ell_\theta(X_i) + o_{P_\theta}(1).$$

Then T_n is best regular estimator for $\psi(\theta)$ at θ. Conversely, every best regular estimator sequence satisfies this expansion.

Proof. The sequence $\Delta_{n,\theta} = n^{-1/2}\sum \dot\ell_\theta(X_i)$ converges in distribution to a vector Δ_θ with a $N(0, I_\theta)$-distribution. By Theorem 7.2 the sequence $\log dP^n_{\theta+h/\sqrt{n}}/dP^n_\theta$ is asymptotically equivalent to $h^T\Delta_{n,\theta} - \frac{1}{2}h^T I_\theta h$. If T_n is asymptotically linear, then $\sqrt{n}\big(T_n - \psi(\theta)\big)$ is asymptotically equivalent to the function $\dot\psi_\theta I_\theta^{-1}\Delta_{n,\theta}$. Apply Slutsky's lemma to find that

$$\left(\sqrt{n}\big(T_n - \psi(\theta)\big), \log\frac{dP_{\theta+h/\sqrt{n}}}{dP_\theta^n}\right) \overset{\theta}{\rightsquigarrow} \left(\dot\psi_\theta I_\theta^{-1}\Delta_\theta, h^T\Delta_\theta - \frac{1}{2}h^T I_\theta h\right)$$
$$\sim N\left(\begin{pmatrix} 0 \\ -\frac{1}{2}h^T I_\theta h\end{pmatrix}, \begin{pmatrix} \dot\psi_\theta I_\theta^{-1}\dot\psi_\theta{}^T & \dot\psi_\theta h \\ \dot\psi_\theta h^T & h^T I_\theta h\end{pmatrix}\right).$$

The limit distribution of the sequence $\sqrt{n}\big(T_n - \psi(\theta)\big)$ under $\theta + h/\sqrt{n}$ follows by Le Cam's third lemma, Example 6.7, and is normal with mean $\dot\psi_\theta h$ and covariance matrix $\dot\psi_\theta I_\theta^{-1}\dot\psi_\theta{}^T$. Combining this with the differentiability of ψ, we obtain that T_n is regular.

Next suppose that S_n and T_n are both best regular estimator sequences. By the same arguments as in the proof of Theorem 8.11 it can be shown that, at least along subsequences, the joint estimators (S_n, T_n) for $\big(\psi(\theta), \psi(\theta)\big)$ satisfy for every h

$$\left(\sqrt{n}\left(S_n - \psi\left(\theta + \frac{h}{\sqrt{n}}\right)\right), \sqrt{n}\left(T_n - \psi\left(\theta + \frac{h}{\sqrt{n}}\right)\right)\right) \overset{\theta+h/\sqrt{n}}{\rightsquigarrow} (S - \dot\psi_\theta h, T - \dot\psi_\theta h),$$

for a randomized estimator (S, T) in the normal-limit experiment. Because S_n and T_n are best regular, the estimators S and T are best equivariant-in-law. Thus $S = T = \dot\psi_\theta X$ almost surely by Proposition 8.6, whence $\sqrt{n}(S_n - T_n)$ converges in distribution to $S - T = 0$.

Thus every two best regular estimator sequences are asymptotically equivalent. The second assertion of the lemma follows on applying this to T_n and the estimators

$$S_n = \psi(\theta) + \frac{1}{\sqrt{n}}\dot\psi_\theta I_\theta^{-1}\Delta_{n,\theta}.$$

Because the parameter θ is known in the local experiments $(P^n_{\theta+h/\sqrt{n}} : h \in \mathbb{R}^k)$, this indeed defines an estimator sequence within the present context. It is best regular by the first part of the lemma. ■

Under regularity conditions, for instance those of Theorem 5.39, the maximum likelihood estimator $\hat\theta_n$ in a parametric model satisfies

$$\sqrt{n}(\hat\theta_n - \theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_\theta^{-1}\dot\ell_\theta(X_i) + o_{P_\theta}(1).$$

Then the maximum likelihood estimator is asymptotically optimal for estimating θ in terms of the convolution theorem. By the delta method, the estimator $\psi(\hat\theta_n)$ for $\psi(\theta)$ can be seen to be asymptotically linear as in the preceding theorem, so that it is asymptotically regular and optimal as well.

Actually, regular and asymptotically optimal estimators for θ exist in every parametric model $(P_\theta : \theta \in \Theta)$ that is differentiable in quadratic mean with nonsingular Fisher information throughout Θ, provided the parameter θ is identifiable. This can be shown using the discretized one-step method discussed in section 5.7 (see [93]).

*8.10 Large Deviations

Consistency of an estimator sequence T_n entails that the probability of the event $d\bigl(T_n, \psi(\theta)\bigr) > \varepsilon$ tends to zero under θ, for every $\varepsilon > 0$. This is a very weak requirement. One method to strengthen it is to make ε dependent on n and to require that the probabilities $\mathrm{P}_\theta\bigl(d\bigl(T_n, \psi(\theta)\bigr) > \varepsilon_n\bigr)$ converge to 0, or are bounded away from 1, for a given sequence $\varepsilon_n \to 0$. The results of the preceding sections address this question and give very precise lower bounds for these probabilities using an "optimal" rate $\varepsilon_n = r_n^{-1}$, typically $n^{-1/2}$.

Another method of strengthening the consistency is to study the speed at which the probabilities $\mathrm{P}_\theta\bigl(d\bigl(T_n, \psi(\theta)\bigr) > \varepsilon\bigr)$ converge to 0 for a fixed $\varepsilon > 0$. This method appears to be of less importance but is of some interest. Typically, the speed of convergence is exponential, and there is a precise lower bound for the exponential rate in terms of the Kullback-Leibler information.

We consider the situation that T_n is based on a random sample of size n from a distribution P_θ, indexed by a parameter θ ranging over an arbitrary set Θ. We wish to estimate the value of a function $\psi : \Theta \mapsto \mathbb{D}$ that takes its values in a metric space.

8.15 ***Theorem.*** *Suppose that the estimator sequence T_n is consistent for $\psi(\theta)$ under every θ. Then, for every $\varepsilon > 0$ and every θ_0,*

$$\limsup_{n\to\infty} -\frac{1}{n} \log P_{\theta_0}\Bigl(d\bigl(T_n, \psi(\theta_0)\bigr) > \varepsilon\Bigr) \leq \inf_{\theta : d(\psi(\theta), \psi(\theta_0)) > \varepsilon} -P_\theta \log \frac{p_{\theta_0}}{p_\theta}.$$

Proof. If the right side is infinite, then there is nothing to prove. The Kullback-Leibler information $-P_\theta \log p_{\theta_0}/p_\theta$ can be finite only if $P_\theta \ll P_{\theta_0}$. Hence, it suffices to prove that $-P_\theta \log p_{\theta_0}/p_\theta$ is an upper bound for the left side for every θ such that $P_\theta \ll P_{\theta_0}$ and $d\bigl(\psi(\theta), \psi(\theta_0)\bigr) > \varepsilon$. The variable $\Lambda_n = (n^{-1}) \sum_{i=1}^n \log(p_\theta/p_{\theta_0})(X_i)$ is well defined (possibly $-\infty$). For every constant M,

$$\begin{aligned} P_{\theta_0}\Bigl(d\bigl(T_n, \psi(\theta_0)\bigr) > \varepsilon\Bigr) &\geq P_{\theta_0}\Bigl(d\bigl(T_n, \psi(\theta_0)\bigr) > \varepsilon, \Lambda_n < M\Bigr) \\ &\geq E_\theta 1\Bigl\{d\bigl(T_n, \psi(\theta_0)\bigr) > \varepsilon, \Lambda_n < M\Bigr\} e^{-n\Lambda_n} \\ &\geq e^{-nM} P_\theta\Bigl(d\bigl(T_n, \psi(\theta_0)\bigr) > \varepsilon, \Lambda_n < M\Bigr). \end{aligned}$$

Take logarithms and multiply by $-(1/n)$ to conclude that

$$-\frac{1}{n} \log P_{\theta_0}\Bigl(d\bigl(T_n, \psi(\theta_0)\bigr) > \varepsilon\Bigr) \leq M - \frac{1}{n} \log P_\theta\Bigl(d\bigl(T_n, \psi(\theta_0)\bigr) > \varepsilon, \Lambda_n < M\Bigr).$$

For $M > P_\theta \log p_\theta/p_{\theta_0}$, we have that $P_\theta(\Lambda_n < M) \to 1$ by the law of large numbers. Furthermore, by the consistency of T_n for $\psi(\theta)$, the probability $P_\theta\bigl(d\bigl(T_n, \psi(\theta_0)\bigr) > \varepsilon\bigr)$

converges to 1 for every θ such that $d\big(\psi(\theta), \psi(\theta_0)\big) > \varepsilon$. Conclude that the probability in the right side of the preceding display converges to 1, whence the lim sup of the left side is bounded by M. ■

Notes

Chapter 32 of the famous book by Cramér [27] 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. Cramér defines an estimator as asymptotically efficient if its efficiency (the quotient mentioned previously) equals one. These definitions lead to the conclusion that the method of maximum likelihood produces asymptotically efficient estimators, as already conjectured by Fisher [48, 50] in the 1920s. That there is a conceptual hole in the definitions was clearly realized in 1951 when Hodges produced his example of a superefficient estimator. Not long after this, in 1953, Le Cam proved that superefficiency can occur only on a Lebesgue null set. Our present result, almost without regularity conditions, is based on later work by Le Cam (see [95].) The asymptotic convolution and minimax theorems were obtained in the present form by Hájek in [69] and [70] after initial work by many authors. Our present proofs follow the approach based on limit experiments, initiated by Le Cam in [95].

PROBLEMS

1. Calculate the asymptotic relative efficiency of the sample mean and the sample median for estimating θ, based on a sample of size n from the normal $N(\theta, 1)$ distribution.

2. As the previous problem, but now for the Laplace distribution (density $p(x) = \frac{1}{2}e^{-|x|}$).

3. Consider estimating the distribution function $\mathrm{P}(X \leq x)$ at a fixed point x based on a sample $X_1, \ldots, X_n$ from the distribution of X. The "nonparametric" estimator is $n^{-1}\#(X_i \leq x)$. If it is known that the true underlying distribution is normal $N(\theta, 1)$, another possible estimator is $\Phi(x - \overline{X})$. Calculate the relative efficiency of these estimators.

4. Calculate the relative efficiency of the empirical p-quantile and the estimator $\Phi^{-1}(p)S_n + \overline{X}_n$ for the estimating the p-th quantile of the distribution of a sample from the normal $N(\mu, \sigma^2)$-distribution.

5. Consider estimating the population variance by either the sample variance S^2 (which is unbiased) or else $n^{-1}\sum_{i=1}^n (X_i - \overline{X})^2 = (n-1)/n\, S^2$. Calculate the asymptotic relative efficiency.

6. Calculate the asymptotic relative efficiency of the sample standard deviation and the interquartile range (corrected for unbiasedness) for estimating the standard deviation based on a sample of size n from the normal $N(\mu, \sigma^2)$-distribution.

7. Given a sample of size n from the uniform distribution on $[0, \theta]$, the maximum $X_{(n)}$ of the observations is biased downwards. Because $\mathrm{E}_\theta(\theta - X_{(n)}) = \mathrm{E}_\theta X_{(1)}$, the bias can be removed by adding the minimum of the observations. Is $X_{(1)} + X_{(n)}$ a good estimator for θ from an asymptotic point of view?

8. Consider the Hodges estimator S_n based on the mean of a sample from the $N(\theta, 1)$-distribution.

(i) Show that $\sqrt{n}(S_n - \theta_n) \overset{\theta_n}{\rightsquigarrow} -\infty$, if $\theta_n \to 0$ in such a way that $n^{1/4}\theta_n \to 0$ and $n^{1/2}\theta_n \to \infty$.

(ii) Show that S_n is not regular at $\theta = 0$.

(iii) Show that $\sup_{-\delta<\theta<\delta} \mathrm{P}_\theta\bigl(\sqrt{n}|S_n - \theta| > k_n\bigr) \to 1$ for every k_n that converges to infinity sufficiently slowly.

9. Show that a loss function $\ell : \mathbb{R} \mapsto \mathbb{R}$ is bowl-shaped if and only if it has the form $\ell(x) = \ell_0\bigl(|x|\bigr)$ for a nondecreasing function ℓ_0.

10. Show that a function of the form $\ell(x) = \ell_0\bigl(\|x\|\bigr)$ for a nondecreasing function ℓ_0 is bowl-shaped.

11. Prove Anderson's lemma for the one-dimensional case, for instance by calculating the derivative of $\int \ell(x + h)\, dN(0, 1)(x)$. Does the proof generalize to higher dimensions?

12. What does Lemma 8.13 imply about the coordinates of the Stein estimator. Are they good estimators of the coordinates of the expectaction vector?

13. All results in this chapter extend in a straightforward manner to general locally asymptotically normal models. Formulate Theorem 8.9 and Lemma 8.14 for such models.