5

M- and Z-Estimators

This chapter gives an introduction to the consistency and asymptotic normality of M-estimators and Z-estimators. Maximum likelihood estimators are treated as a special case.

5.1 Introduction

Suppose that we are interested in a parameter (or "functional") θ attached to the distribution of observations $X_1, \ldots, X_n$. A popular method for finding an estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ is to maximize a criterion function of the type

$$\theta \mapsto M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i). \tag{5.1}$$

Here $m_\theta \colon \mathcal{X} \mapsto \overline{\mathbb{R}}$ are known functions. An estimator $\hat{\theta}_n$ maximizing $M_n(\theta)$ over Θ is called an *M-estimator*. In this chapter we investigate the asymptotic behavior of sequences of M-estimators.

Often the maximizing value is sought by setting a derivative (or the set of partial derivatives in the multidimensional case) equal to zero. Therefore, the name *M-estimator* is also used for estimators satisfying systems of equations of the type

$$\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \psi_\theta(X_i) = 0. \tag{5.2}$$

Here ψ_θ are known vector-valued maps. For instance, if θ is k-dimensional, then ψ_θ typically has k coordinate functions $\psi_\theta = (\psi_{\theta,1}, \ldots, \psi_{\theta,k})$, and (5.2) is shorthand for the system of equations

$$\sum_{i=1}^{n} \psi_{\theta,j}(X_i) = 0, \qquad j = 1, 2, \ldots, k.$$

Even though in many examples $\psi_{\theta,j}$ is the jth partial derivative of some function m_θ, this is irrelevant for the following. Equations, such as (5.2), defining an estimator are called *estimating equations* and need not correspond to a maximization problem. In the latter case it is probably better to call the corresponding estimators *Z-estimators* (for zero), but the use of the name *M-estimator* is widespread.

Sometimes the maximum of the criterion function M_n is not taken or the estimating equation does not have an exact solution. Then it is natural to use as estimator a value that almost maximizes the criterion function or is a near zero. This yields approximate M-estimators or Z-estimators. Estimators that are sufficiently close to being a point of maximum or a zero often have the same asymptotic behavior.

An operator notation for taking expectations simplifies the formulas in this chapter. We write P for the marginal law of the observations $X_1, \ldots, X_n$, which we assume to be identically distributed. Furthermore, we write Pf for the expectation $\mathrm{E} f(X) = \int f \, dP$ and abbreviate the average $n^{-1}\sum_{i=1}^n f(X_i)$ to $\mathbb{P}_n f$. Thus $\mathbb{P}_n$ is the *empirical distribution*: the (random) discrete distribution that puts mass $1/n$ at every of the observations $X_1, \ldots, X_n$. The criterion functions now take the forms

$$M_n(\theta) = \mathbb{P}_n m_\theta, \quad \text{and} \quad \Psi_n(\theta) = \mathbb{P}_n \psi_\theta.$$

We also abbreviate the centered sums $n^{-1/2}\sum_{i=1}^n \big(f(X_i) - Pf\big)$ to $\mathbb{G}_n f$, the *empirical process* at f.

5.3 ***Example (Maximum likelihood estimators).*** Suppose $X_1, \ldots, X_n$ have a common density p_θ. Then the *maximum likelihood estimator* maximizes the likelihood $\prod_{i=1}^n p_\theta(X_i)$, or equivalently the log likelihood

$$\theta \mapsto \sum_{i=1}^n \log p_\theta(X_i).$$

Thus, a maximum likelihood estimator is an M-estimator as in (5.1) with $m_\theta = \log p_\theta$. If the density is partially differentiable with respect to θ for each fixed x, then the maximum likelihood estimator also solves an equation of type (5.2), with ψ_θ equal to the vector of partial derivatives $\dot\ell_{\theta,j} = \partial/\partial\theta_j \log p_\theta$. The vector-valued function $\dot\ell_\theta$ is known as the *score function* of the model.

The definition (5.1) of an M-estimator may apply in cases where (5.2) does not. For instance, if $X_1, \ldots, X_n$ are i.i.d. according to the uniform distribution on $[0, \theta]$, then it makes sense to maximize the log likelihood

$$\theta \mapsto \sum_{i=1}^n \big(\log 1_{[0,\theta]}(X_i) - \log\theta\big).$$

(Define $\log 0 = -\infty$.) However, this function is not smooth in θ and there exists no natural version of (5.2). Thus, in this example the definition as the location of a maximum is more fundamental than the definition as a zero. □

5.4 ***Example (Location estimators).*** Let $X_1, \ldots, X_n$ be a random sample of real-valued observations and suppose we want to estimate the location of their distribution. "Location" is a vague term; it could be made precise by defining it as the mean or median, or the center of symmetry of the distribution if this happens to be symmetric. Two examples of location estimators are the sample mean and the sample median. Both are Z-estimators, because they solve the equations

$$\sum_{i=1}^n (X_i - \theta) = 0; \quad \text{and} \quad \sum_{i=1}^n \operatorname{sign}(X_i - \theta) = 0,$$

respectively.† Both estimating equations involve functions of the form $\psi(x-\theta)$ for a function ψ that is monotone and odd around zero. It seems reasonable to study estimators that solve a general equation of the type

$$\sum_{i=1}^{n}\psi(X_i-\theta)=0.$$

We can consider a Z-estimator defined by this equation a "location" estimator, because it has the desirable property of location equivariance. If the observations X_i are shifted by a fixed amount α, then so is the estimate: $\hat{\theta}+\alpha$ solves $\sum_{i=1}^{n}\psi(X_i+\alpha-\theta)=0$ if $\hat{\theta}$ solves the original equation.

Popular examples are the *Huber estimators* corresponding to the functions

$$\psi(x)=[x]_{-k}^{k}:=\begin{cases}-k & \text{if } x\le -k,\\ x & \text{if } |x|\le k,\\ k & \text{if } x\ge k.\end{cases}$$

The Huber estimators were motivated by studies in robust statistics concerning the influence of extreme data points on the estimate. The exact values of the largest and smallest observations have very little influence on the value of the median, but a proportional influence on the mean. Therefore, the sample mean is considered nonrobust against outliers. If the extreme observations are thought to be rather unreliable, it is certainly an advantage to limit their influence on the estimate, but the median may be too successful in this respect. Depending on the value of k, the Huber estimators behave more like the mean (large k) or more like the median (small k) and thus bridge the gap between the nonrobust mean and very robust median.

Another example are the *quantiles*. A pth sample quantile is roughly a point θ such that pn observations are less than θ and $(1-p)n$ observations are greater than θ. The precise definition has to take into account that the value pn may not be an integer. One possibility is to call a *pth sample quantile* any $\hat{\theta}$ that solves the inequalities

$$-1<\sum_{i=1}^{n}\bigl((1-p)1\{X_i<\theta\}-p1\{X_i>\theta\}\bigr)<1. \tag{5.5}$$

This is an approximate M-estimator for $\psi(x)=1-p,0,-p$ if $x<0$, $x=0$, or $x>0$, respectively. The "approximate" refers to the inequalities: It is required that the value of the estimating equation be inside the interval $(-1,1)$, rather than exactly zero. This may seem a rather wide tolerance interval for a zero. However, all solutions turn out to have the same asymptotic behavior. In any case, except for special combinations of p and n, there is no hope of finding an exact zero, because the criterion function is discontinuous with jumps at the observations. (See Figure 5.1.) If no observations are tied, then all jumps are of size one and at least one solution $\hat{\theta}$ to the inequalities exists. If tied observations are present, it may be necessary to increase the interval $(-1,1)$ to ensure the existence of solutions. Note that the present ψ function is monotone, as in the previous examples, but not symmetric about zero (for $p\neq 1/2$).

† The *sign-function* is defined as $\operatorname{sign}(x)=-1,0,1$ if $x<0$, $x=0$ or $x>0$, respectively. Also x^+ means $x\vee 0=\max(x,0)$. For the median we assume that there are no tied observations (in the middle).

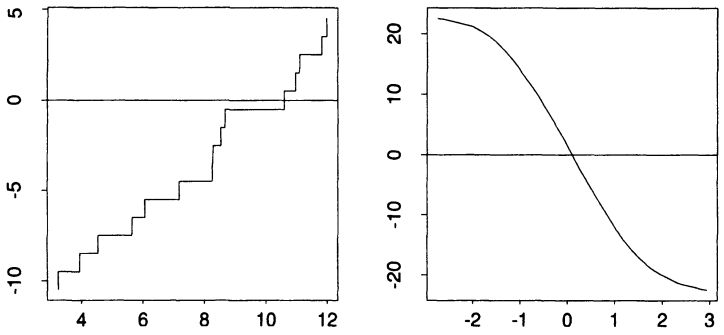

Figure 5.1. The functions $\theta \mapsto \Psi_n(\theta)$ for the 80% quantile and the Huber estimator for samples of size 15 from the gamma(8,1) and standard normal distribution, respectively.

All the estimators considered so far can also be defined as a solution of a maximization problem. Mean, median, Huber estimators, and quantiles minimize $\sum_{i=1}^n m(X_i - \theta)$ for m equal to x^2, $|x|$, $x^2 1_{|x|\le k} + \big(2k|x| - k^2\big)1_{|x|>k}$ and $(1-p)x^- + px^+$, respectively. □

5.2 Consistency

If the estimator $\hat\theta_n$ is used to estimate the parameter θ, then it is certainly desirable that the sequence $\hat\theta_n$ converges in probability to θ. If this is the case for every possible value of the parameter, then the sequence of estimators is called *asymptotically consistent*. For instance, the sample mean $\overline{X}_n$ is asymptotically consistent for the population mean EX (provided the population mean exists). This follows from the law of large numbers. Not surprisingly this extends to many other sample characteristics. For instance, the sample median is consistent for the population median, whenever this is well defined. What can be said about M-estimators in general? We shall assume that the set of possible parameters is a metric space, and write d for the metric. Then we wish to prove that $d(\hat\theta_n, \theta_0) \xrightarrow{P} 0$ for some value θ_0, which depends on the underlying distribution of the observations.

Suppose that the M-estimator $\hat\theta_n$ maximizes the random criterion function

$$\theta \mapsto M_n(\theta).$$

Clearly, the "asymptotic value" of $\hat\theta_n$ depends on the asymptotic behavior of the functions M_n. Under suitable normalization there typically exists a deterministic "asymptotic criterion function" $\theta \mapsto M(\theta)$ such that

$$M_n(\theta) \xrightarrow{P} M(\theta), \qquad \text{every } \theta. \tag{5.6}$$

For instance, if $M_n(\theta)$ is an average of the form $\mathbb{P}_n m_\theta$ as in (5.1), then the law of large numbers gives this result with $M(\theta) = Pm_\theta$, provided this expectation exists.

It seems reasonable to expect that the maximizer $\hat\theta_n$ of M_n converges to the maximizing value θ_0 of M. This is what we wish to prove in this section, and we say that $\hat\theta_n$ is (asymptotically) consistent for θ_0. However, the convergence (5.6) is too weak to ensure

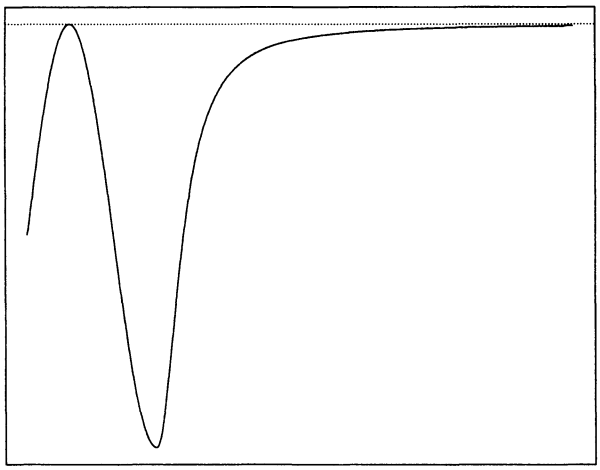

Figure 5.2. Example of a function whose point of maximum is not well separated.

the convergence of $\hat{\theta}_n$. Because the value $\hat{\theta}_n$ depends on the whole function $\theta \mapsto M_n(\theta)$, an appropriate form of "functional convergence" of M_n to M is needed, strengthening the pointwise convergence (5.6). There are several possibilities. In this section we first discuss an approach based on uniform convergence of the criterion functions. Admittedly, the assumption of uniform convergence is too strong for some applications and it is sometimes not easy to verify, but the approach illustrates the general idea.

Given an arbitrary random function $\theta \mapsto M_n(\theta)$, consider estimators $\hat{\theta}_n$ that *nearly maximize* M_n, that is,

$$M_n(\hat{\theta}_n) \geq \sup_{\theta} M_n(\theta) - o_P(1).$$

Then certainly $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$, which turns out to be enough to ensure consistency. It is assumed that the sequence M_n converges to a nonrandom map $M: \Theta \mapsto \overline{\mathbb{R}}$. Condition (5.8) of the following theorem requires that this map attains its maximum at a unique point θ_0, and only parameters close to θ_0 may yield a value of $M(\theta)$ close to the maximum value $M(\theta_0)$. Thus, θ_0 should be a *well-separated* point of maximum of M. Figure 5.2 shows a function that does not satisfy this requirement.

5.7 *Theorem.* *Let M_n be random functions and let M be a fixed function of θ such that for every $\varepsilon > 0$*[†]

$$\begin{aligned} &\sup_{\theta \in \Theta} \left| M_n(\theta) - M(\theta) \right| \xrightarrow{\mathrm{P}} 0, \\ &\sup_{\theta : d(\theta, \theta_0) \geq \varepsilon} M(\theta) < M(\theta_0). \end{aligned} \tag{5.8}$$

Then any sequence of estimators $\hat{\theta}_n$ with $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$ converges in probability to θ_0.

[†] Some of the expressions in this display may be nonmeasurable. Then the probability statements are understood in terms of outer measure.

Proof. By the property of $\hat{\theta}_n$, we have $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$. Because the uniform convergence of M_n to M implies the convergence of $M_n(\theta_0) \xrightarrow{P} M(\theta_0)$, the right side equals $M(\theta_0) - o_P(1)$. It follows that $M_n(\hat{\theta}_n) \geq M(\theta_0) - o_P(1)$, whence

$$\begin{aligned} M(\theta_0) - M(\hat{\theta}_n) &\leq M_n(\hat{\theta}_n) - M(\hat{\theta}_n) + o_P(1) \\ &\leq \sup_\theta |M_n - M|(\theta) + o_P(1) \xrightarrow{P} 0. \end{aligned}$$

by the first part of assumption (5.8). By the second part of assumption (5.8), there exists for every $\varepsilon > 0$ a number $\eta > 0$ such that $M(\theta) < M(\theta_0) - \eta$ for every θ with $d(\theta, \theta_0) \geq \varepsilon$. Thus, the event $\{d(\hat{\theta}_n, \theta_0) \geq \varepsilon\}$ is contained in the event $\{M(\hat{\theta}_n) < M(\theta_0) - \eta\}$. The probability of the latter event converges to 0, in view of the preceding display. ■

Instead of through maximization, an M-estimator may be defined as a zero of a criterion function $\theta \mapsto \Psi_n(\theta)$. It is again reasonable to assume that the sequence of criterion functions converges to a fixed limit:

$$\Psi_n(\theta) \xrightarrow{P} \Psi(\theta).$$

Then it may be expected that a sequence of (approximate) zeros of Ψ_n converges in probability to a zero of Ψ. This is true under similar restrictions as in the case of maximizing M estimators. In fact, this can be deduced from the preceding theorem by noting that a zero of Ψ_n maximizes the function $\theta \mapsto -\|\Psi_n(\theta)\|$.

5.9 *Theorem.* *Let Ψ_n be random vector-valued functions and let Ψ be a fixed vector-valued function of θ such that for every $\varepsilon > 0$*

$$\begin{aligned} \sup_{\theta \in \Theta} \|\Psi_n(\theta) - \Psi(\theta)\| &\xrightarrow{P} 0, \\ \inf_{\theta : d(\theta, \theta_0) \geq \varepsilon} \|\Psi(\theta)\| > 0 &= \|\Psi(\theta_0)\|. \end{aligned}$$

Then any sequence of estimators $\hat{\theta}_n$ such that $\Psi_n(\hat{\theta}_n) = o_P(1)$ converges in probability to θ_0.

Proof. This follows from the preceding theorem, on applying it to the functions $M_n(\theta) = -\|\Psi_n(\theta)\|$ and $M(\theta) = -\|\Psi(\theta)\|$. ■

The conditions of both theorems consist of a stochastic and a deterministic part. The deterministic condition can be verified by drawing a picture of the graph of the function. A helpful general observation is that, for a compact set Θ and continuous function M or Ψ, uniqueness of θ_0 as a maximizer or zero implies the condition. (See Problem 5.27.)

For $M_n(\theta)$ or $\Psi_n(\theta)$ equal to averages as in (5.1) or (5.2) the uniform convergence required by the stochastic condition is equivalent to the set of functions $\{m_\theta : \theta \in \Theta\}$ or $\{\psi_{\theta,j} : \theta \in \Theta, j = 1, \ldots, k\}$ being *Glivenko-Cantelli*. Glivenko-Cantelli classes of functions are discussed in Chapter 19. One simple set of sufficient conditions is that Θ be compact, that the functions $\theta \mapsto m_\theta(x)$ or $\theta \mapsto \psi_\theta(x)$ are continuous for every x, and that they are dominated by an integrable function.

Uniform convergence of the criterion functions as in the preceding theorems is much stronger than needed for consistency. The following lemma is one of the many possibilities to replace the uniformity by other assumptions.

5.10 ***Lemma.*** *Let Θ be a subset of the real line and let Ψ_n be random functions and Ψ a fixed function of θ such that $\Psi_n(\theta) \to \Psi(\theta)$ in probability for every θ. Assume that each map $\theta \mapsto \Psi_n(\theta)$ is continuous and has exactly one zero $\hat{\theta}_n$, or is nondecreasing with $\Psi_n(\hat{\theta}_n) = o_P(1)$. Let θ_0 be a point such that $\Psi(\theta_0 - \varepsilon) < 0 < \Psi(\theta_0 + \varepsilon)$ for every $\varepsilon > 0$. Then $\hat{\theta}_n \xrightarrow{P} \theta_0$.*

Proof. If the map $\theta \mapsto \Psi_n(\theta)$ is continuous and has a unique zero at $\hat{\theta}_n$, then

$$\mathrm{P}\big(\Psi_n(\theta_0 - \varepsilon) < 0, \Psi_n(\theta_0 + \varepsilon) > 0\big) \le \mathrm{P}(\theta_0 - \varepsilon < \hat{\theta}_n < \theta_0 + \varepsilon).$$

The left side converges to one, because $\Psi_n(\theta_0 \pm \varepsilon) \to \Psi(\theta_0 \pm \varepsilon)$ in probability. Thus the right side converges to one as well, and $\hat{\theta}_n$ is consistent.

If the map $\theta \mapsto \Psi_n(\theta)$ is nondecreasing and $\hat{\theta}_n$ is a zero, then the same argument is valid. More generally, if $\theta \mapsto \Psi_n(\theta)$ is nondecreasing, then $\Psi_n(\theta_0 - \varepsilon) < -\eta$ and $\hat{\theta}_n \le \theta_0 - \varepsilon$ imply $\Psi_n(\hat{\theta}_n) < -\eta$, which has probability tending to zero for every $\eta > 0$ if $\hat{\theta}_n$ is a near zero. This and a similar argument applied to the right tail shows that, for every $\varepsilon, \eta > 0$,

$$\mathrm{P}\big(\Psi_n(\theta_0 - \varepsilon) < -\eta, \Psi_n(\theta_0 + \varepsilon) > \eta\big) \le \mathrm{P}(\theta_0 - \varepsilon < \hat{\theta}_n < \theta_0 + \varepsilon) + o(1).$$

For 2η equal to the smallest of the numbers $-\Psi(\theta_0 - \varepsilon)$ and $\Psi(\theta_0 + \varepsilon)$ the left side still converges to one. ■

5.11 ***Example (Median).*** The sample median $\hat{\theta}_n$ is a (near) zero of the map $\theta \mapsto \Psi_n(\theta) = n^{-1}\sum_{i=1}^n \operatorname{sign}(X_i - \theta)$. By the law of large numbers,

$$\Psi_n(\theta) \xrightarrow{\mathrm{P}} \Psi(\theta) = \mathrm{E}\operatorname{sign}(X - \theta) = \mathrm{P}(X > \theta) - \mathrm{P}(X < \theta),$$

for every fixed θ. Thus, we expect that the sample median converges in probability to a point θ_0 such that $\mathrm{P}(X > \theta_0) = \mathrm{P}(X < \theta_0)$: a population median.

This can be proved rigorously by applying Theorem 5.7 or 5.9. However, even though the conditions of the theorems are satisfied, they are not entirely trivial to verify. (The uniform convergence of Ψ_n to Ψ is proved essentially in Theorem 19.1) In this case it is easier to apply Lemma 5.10. Because the functions $\theta \mapsto \Psi_n(\theta)$ are nonincreasing, it follows that $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$ provided that $\Psi(\theta_0 - \varepsilon) > 0 > \Psi(\theta_0 + \varepsilon)$ for every $\varepsilon > 0$. This is the case if the population median is unique: $\mathrm{P}(X < \theta_0 - \varepsilon) < \frac{1}{2} < \mathrm{P}(X < \theta_0 + \varepsilon)$ for all $\varepsilon > 0$. □

*5.2.1 *Wald's Consistency Proof*

Consider the situation that, for a random sample of variables $X_1, \ldots, X_n$,

$$M_n(\theta) = \mathbb{P}_n m_\theta = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i), \qquad M(\theta) = P m_\theta.$$

In this subsection we consider an alternative set of conditions under which the maximizer $\hat{\theta}_n$ of the process M_n converges in probability to a point of maximum θ_0 of the function M. This "classical" approach to consistency was taken by Wald in 1949 for maximum likelihood estimators. It works best if the parameter set Θ is compact. If not, then the argument must

be complemented by a proof that the estimators are in a compact set eventually or be applied to a suitable compactification of the parameter set.

Assume that the map $\theta \mapsto m_\theta(x)$ is upper-semicontinuous for almost all x: For every θ

$$\limsup_{\theta_n \to \theta} m_{\theta_n}(x) \le m_\theta(x), \qquad \text{a.s..} \tag{5.12}$$

(The exceptional set of x may depend on θ.) Furthermore, assume that for every sufficiently small ball $U \subset \Theta$ the function $x \mapsto \sup_{\theta \in U} m_\theta(x)$ is measurable and satisfies

$$P \sup_{\theta \in U} m_\theta < \infty. \tag{5.13}$$

Typically, the map $\theta \mapsto Pm_\theta$ has a unique global maximum at a point θ_0, but we shall allow multiple points of maximum, and write Θ_0 for the set $\{\theta_0 \in \Theta: Pm_{\theta_0} = \sup_\theta Pm_\theta\}$ of all points at which M attains its global maximum. The set Θ_0 is assumed not empty. The maps $m_\theta: \mathcal{X} \mapsto \overline{\mathbb{R}}$ are allowed to take the value $-\infty$, but the following theorem assumes implicitly that at least Pm_{θ_0} is finite.

5.14 *Theorem.* *Let $\theta \mapsto m_\theta(x)$ be upper-semicontinuous for almost all x and let (5.13) be satisfied. Then for any estimators $\hat\theta_n$ such that $M_n(\hat\theta_n) \ge M_n(\theta_0) - o_P(1)$ for some $\theta_0 \in \Theta_0$, for every $\varepsilon > 0$ and every compact set $K \subset \Theta$,*

$$\mathrm{P}\big(d(\hat\theta_n, \Theta_0) \ge \varepsilon \wedge \hat\theta_n \in K\big) \to 0.$$

Proof. If the function $\theta \mapsto Pm_\theta$ is identically $-\infty$, then $\Theta_0 = \Theta$, and there is nothing to prove. Hence, we may assume that there exists $\theta_0 \in \Theta_0$ such that $Pm_{\theta_0} > -\infty$, whence $P|m_{\theta_0}| < \infty$ by (5.13).

Fix some θ and let $U_l \downarrow \theta$ be a decreasing sequence of open balls around θ of diameter converging to zero. Write $m_U(x)$ for $\sup_{\theta \in U} m_\theta(x)$. The sequence m_{U_l} is decreasing and greater than m_θ for every l. Combination with (5.12) yields that $m_{U_l} \downarrow m_\theta$ almost surely. In view of (5.13), we can apply the monotone convergence theorem and obtain that $Pm_{U_l} \downarrow Pm_\theta$ (which may be $-\infty$).

For $\theta \notin \Theta_0$, we have $Pm_\theta < Pm_{\theta_0}$. Combine this with the preceding paragraph to see that for every $\theta \notin \Theta_0$ there exists an open ball U_θ around θ with $Pm_{U_\theta} < Pm_{\theta_0}$. The set $B = \big\{\theta \in K : d(\theta, \Theta_0) \ge \varepsilon\big\}$ is compact and is covered by the balls $\{U_\theta : \theta \in B\}$. Let $U_{\theta_1}, \ldots, U_{\theta_p}$ be a finite subcover. Then, by the law of large numbers,

$$\sup_{\theta \in B} \mathbb{P}_n m_\theta \le \sup_{j=1,\ldots,p} \mathbb{P}_n m_{U_{\theta_j}} \xrightarrow{\text{as}} \sup_j Pm_{U_{\theta_j}} < Pm_{\theta_0}.$$

If $\hat\theta_n \in B$, then $\sup_{\theta \in B} \mathbb{P}_n m_\theta$ is at least $\mathbb{P}_n m_{\hat\theta_n}$, which by definition of $\hat\theta_n$ is at least $\mathbb{P}_n m_{\theta_0} - o_P(1) = Pm_{\theta_0} - o_P(1)$, by the law of large numbers. Thus

$$\{\hat\theta_n \in B\} \subset \Big\{\sup_{\theta \in B} \mathbb{P}_n m_\theta \ge Pm_{\theta_0} - o_P(1)\Big\}.$$

In view of the preceding display the probability of the event on the right side converges to zero as $n \to \infty$. ■

Even in simple examples, condition (5.13) can be restrictive. One possibility for relaxation is to divide the n observations in groups of approximately the same size. Then (5.13)

may be replaced by, for some k and every $k \le l < 2k$,

$$P^l \sup_{\theta \in U} \sum_{i=1}^{l} m_\theta(x_i) < \infty. \tag{5.15}$$

Surprisingly enough, this simple device may help. For instance, under condition (5.13) the preceding theorem does not apply to yield the asymptotic consistency of the maximum likelihood estimator of (μ, σ) based on a random sample from the $N(\mu, \sigma^2)$ distribution (unless we restrict the parameter set for σ), but under the relaxed condition it does (with $k = 2$). (See Problem 5.25.) The proof of the theorem under (5.15) remains almost the same. Divide the n observations in groups of k observations and, possibly, a remainder group of l observations; next, apply the law of large numbers to the approximately n/k group sums.

5.16 ***Example (Cauchy likelihood).*** The maximum likelihood estimator for θ based on a random sample from the Cauchy distribution with location θ maximizes the map $\theta \mapsto \mathbb{P}_n m_\theta$ for

$$m_\theta(x) = -\log\bigl(1 + (x - \theta)^2\bigr).$$

The natural parameter set $\mathbb{R}$ is not compact, but we can enlarge it to the extended real line, provided that we can define m_θ in a reasonable way for $\theta = \pm\infty$. To have the best chance of satisfying (5.13), we opt for the minimal extension, which in order to satisfy (5.12) is

$$m_{-\infty}(x) = \limsup_{\theta \mapsto -\infty} m_\theta(x) = -\infty; \qquad m_\infty(x) = \limsup_{\theta \mapsto \infty} m_\theta(x) = -\infty.$$

These infinite values should not worry us: They are permitted in the preceding theorem. Moreover, because we maximize $\theta \mapsto \mathbb{P}_n m_\theta$, they ensure that the estimator $\hat\theta_n$ never takes the values $\pm\infty$, which is excellent.

We apply Wald's theorem with $\Theta = \overline{\mathbb{R}}$, equipped with, for instance, the metric $d(\theta_1, \theta_2) = |\arctg \theta_1 - \arctg \theta_2|$. Because the functions $\theta \mapsto m_\theta(x)$ are continuous and nonpositive, the conditions are trivially satisfied. Thus, taking $K = \overline{\mathbb{R}}$, we obtain that $d(\hat\theta_n, \Theta_0) \xrightarrow{\mathrm{P}} 0$. This conclusion is valid for any underlying distribution P of the observations for which the set Θ_0 is nonempty, because so far we have used the Cauchy likelihood only to motivate m_θ.

To conclude that the maximum likelihood estimator in a Cauchy location model is consistent, it suffices to show that $\Theta_0 = \{\theta_0\}$ if P is the Cauchy distribution with center θ_0. This follows most easily from the identifiability of this model, as discussed in Lemma 5.35. □

5.17 ***Example (Current status data).*** Suppose that a "death" that occurs at time T is only observed to have taken place or not at a known "check-up time" C. We model the observations as a random sample $X_1, \ldots, X_n$ from the distribution of $X = \bigl(C, 1\{T \le C\}\bigr)$, where T and C are independent random variables with completely unknown distribution functions F and G, respectively. The purpose is to estimate the "survival distribution" $1 - F$.

If G has a density g with respect to Lebesgue measure λ, then $X = (C, \Delta)$ has a density

$$p_F(c, \delta) = \bigl(\delta F(c) + (1 - \delta)(1 - F)(c)\bigr)g(c)$$

with respect to the product of λ and counting measure on the set $\{0, 1\}$. A maximum likelihood estimator for F can be defined as the distribution function $\hat{F}$ that maximizes the likelihood

$$F \mapsto \prod_{i=1}^{n} \big(\Delta_i F(C_i) + (1 - \Delta_i)(1 - F)(C_i)\big)$$

over all distribution functions on $[0, \infty)$. Because this only involves the numbers $F(C_1)$, $\ldots$, $F(C_n)$, the maximizer of this expression is not unique, but some thought shows that there is a unique maximizer $\hat{F}$ that concentrates on (a subset of) the observation times $C_1, \ldots, C_n$. This is commonly used as an estimator.

We can show the consistency of this estimator by Wald's theorem. By its definition $\hat{F}$ maximizes the function $F \mapsto \mathbb{P}_n \log p_F$, but the consistency proof proceeds in a smoother way by setting

$$m_F = \log \frac{p_F}{p_{(F+F_0)/2}} = \log \frac{2p_F}{p_F + p_{F_0}}.$$

Because the likelihood is bigger at $\hat{F}$ than it is at $\frac{1}{2}\hat{F} + \frac{1}{2}F_0$, it follows that $\mathbb{P}_n m_{\hat{F}} \geq 0 = \mathbb{P}_n m_{F_0}$. (It is not claimed that $\hat{F}$ maximizes $F \mapsto \mathbb{P}_n m_F$; this is not true.)

Condition (5.13) is satisfied trivially, because $m_F \leq \log 2$ for every F. We can equip the set of all distribution functions with the topology of weak convergence. If we restrict the parameter set to distributions on a compact interval $[0, \tau]$, then the parameter set is compact by Prohorov's theorem.† The map $F \mapsto m_F(c, \delta)$ is continuous at F, relative to the weak topology, for every (c, δ) such that c is a continuity point of F. Under the assumption that G has a density, this includes almost every (c, δ), for every given F. Thus, Theorem 5.14 shows that $\hat{F}_n$ converges under F_0 in probability to the set $\mathcal{F}_0$ of all distribution functions that maximize the map $F \mapsto P_{F_0} m_F$, provided $F_0 \in \mathcal{F}_0$. This set always contains F_0, but it does not necessarily reduce to this single point. For instance, if the density g is zero on an interval $[a, b]$, then we receive no information concerning deaths inside the interval $[a, b]$, and there can be no hope that $\hat{F}_n$ converges to F_0 on $[a, b]$. In that case, F_0 is not "identifiable" on the interval $[a, b]$.

We shall show that $\mathcal{F}_0$ is the set of all F such that $F = F_0$ almost everywhere according to G. Thus, the sequence $\hat{F}_n$ is consistent for F_0 "on the set of time points that have a positive probability of occurring."

Because $p_F = p_{F_0}$ under P_{F_0} if and only if $F = F_0$ almost everywhere according to G, it suffices to prove that, for every pair of probability densities p and p_0, $P_0 \log 2p/(p+p_0) \leq 0$ with equality if and only if $p = p_0$ almost surely under P_0. If $P_0(p = 0) > 0$, then $\log 2p/(p + p_0) = -\infty$ with positive probability and hence, because the function is bounded above, $P_0 \log 2p/(p + p_0) = -\infty$. Thus we may assume that $P_0(p = 0) = 0$. Then, with $f(u) = -u \log(\frac{1}{2} + \frac{1}{2}u)$,

$$P_0 \log \frac{2p}{(p + p_0)} = Pf\left(\frac{p_0}{p}\right) \leq f\left(P\frac{p_0}{p}\right) = f(1) = 0,$$

† Alternatively, consider all probability distributions on the compactification $[0, \infty]$ again equipped with the weak topology.

by Jensen's inequality and the concavity of f, with equality only if $p_0/p = 1$ almost surely under P, and then also under P_0. This completes the proof. □

5.3 Asymptotic Normality

Suppose a sequence of estimators $\hat{\theta}_n$ is consistent for a parameter θ that ranges over an open subset of a Euclidean space. The next question of interest concerns the order at which the discrepancy $\hat{\theta}_n - \theta$ converges to zero. The answer depends on the specific situation, but for estimators based on n replications of an experiment the order is often $n^{-1/2}$. Then multiplication with the inverse of this rate creates a proper balance, and the sequence $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution, most often a normal distribution. This is interesting from a theoretical point of view. It also makes it possible to obtain approximate confidence sets. In this section we derive the asymptotic normality of M-estimators.

We can use a characterization of M-estimators either by maximization or by solving estimating equations. Consider the second possibility. Let $X_1, \ldots, X_n$ be a sample from some distribution P, and let a random and a "true" criterion function be of the form:

$$\Psi_n(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n}\psi_\theta(X_i) = \mathbb{P}_n\psi_\theta, \qquad \Psi(\theta) = P\psi_\theta.$$

Assume that the estimator $\hat{\theta}_n$ is a zero of Ψ_n and converges in probability to a zero θ_0 of Ψ. Because $\hat{\theta}_n \to \theta_0$, it makes sense to expand $\Psi_n(\hat{\theta}_n)$ in a Taylor series around θ_0. Assume for simplicity that θ is one-dimensional. Then

$$0 = \Psi_n(\hat{\theta}_n) = \Psi_n(\theta_0) + (\hat{\theta}_n - \theta_0)\dot{\Psi}_n(\theta_0) + \tfrac{1}{2}(\hat{\theta}_n - \theta_0)^2\ddot{\Psi}_n(\tilde{\theta}_n),$$

where $\tilde{\theta}_n$ is a point between $\hat{\theta}_n$ and θ_0. This can be rewritten as

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{-\sqrt{n}\Psi_n(\theta_0)}{\dot{\Psi}_n(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)\ddot{\Psi}_n(\tilde{\theta}_n)}. \tag{5.18}$$

If $P\psi_{\theta_0}^2$ is finite, then the numerator $-\sqrt{n}\Psi_n(\theta_0) = -n^{-1/2}\sum\psi_{\theta_0}(X_i)$ is asymptotically normal by the central limit theorem. The asymptotic mean and variance are $P\psi_{\theta_0} = \Psi(\theta_0) = 0$ and $P\psi_{\theta_0}^2$, respectively. Next consider the denominator. The first term $\dot{\Psi}_n(\theta_0)$ is an average and can be analyzed by the law of large numbers: $\dot{\Psi}_n(\theta_0) \xrightarrow{\mathrm{P}} P\dot{\psi}_{\theta_0}$, provided the expectation exists. The second term in the denominator is a product of $\hat{\theta}_n - \theta = o_P(1)$ and $\ddot{\psi}_n(\tilde{\theta}_n)$ and converges in probability to zero under the reasonable condition that $\ddot{\Psi}_n(\tilde{\theta}_n)$ (which is also an average) is $O_P(1)$. Together with Slutsky's lemma, these observations yield

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N\left(0, \frac{P\psi_{\theta_0}^2}{\left(P\dot{\psi}_{\theta_0}\right)^2}\right). \tag{5.19}$$

The preceding derivation can be made rigorous by imposing appropriate conditions, often called "regularity conditions." The only real challenge is to show that $\ddot{\Psi}_n(\tilde{\theta}_n) = O_P(1)$ (see Problem 5.20 or section 5.6).

The derivation can be extended to higher-dimensional parameters. For a k-dimensional parameter, we use k estimating equations. Then the criterion functions are maps $\Psi_n : \mathbb{R}^k \mapsto$

$\mathbb{R}^k$ and the derivatives $\dot{\Psi}_n(\theta_0)$ are $(k \times k)$-matrices that converge to the $(k \times k)$ matrix $P\dot{\psi}_{\theta_0}$ with entries $P\partial/\partial\theta_j \psi_{\theta_0,i}$. The final statement becomes

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N_k\left(0, \left(P\dot{\psi}_{\theta_0}\right)^{-1} P\psi_{\theta_0}\psi_{\theta_0}^T \left(P\dot{\psi}_{\theta_0}^T\right)^{-1}\right). \tag{5.20}$$

Here the invertibility of the matrix $P\dot{\psi}_{\theta_0}$ is a condition.

In the preceding derivation it is implicitly understood that the function $\theta \mapsto \psi_\theta(x)$ possesses two continuous derivatives with respect to the parameter, for every x. This is true in many examples but fails, for instance, for the function $\psi_\theta(x) = \operatorname{sign}(x-\theta)$, which yields the median. Nevertheless, the median is asymptotically normal. That such a simple, but important, example cannot be treated by the preceding approach has motivated much effort to derive the asymptotic normality of M-estimators by more refined methods. One result is the following theorem, which assumes less than one derivative (a Lipschitz condition) instead of two derivatives.

5.21 *Theorem.* *For each θ in an open subset of Euclidean space, let $x \mapsto \psi_\theta(x)$ be a measurable vector-valued function such that, for every θ_1 and θ_2 in a neighborhood of θ_0 and a measurable function $\dot{\psi}$ with $P\dot{\psi}^2 < \infty$,*

$$\left\|\psi_{\theta_1}(x) - \psi_{\theta_2}(x)\right\| \le \dot{\psi}(x)\,\|\theta_1 - \theta_2\|.$$

Assume that $P\|\psi_{\theta_0}\|^2 < \infty$ and that the map $\theta \mapsto P\psi_\theta$ is differentiable at a zero θ_0, with nonsingular derivative matrix V_{θ_0}. If $\mathbb{P}_n\psi_{\hat{\theta}_n} = o_P(n^{-1/2})$, and $\hat{\theta}_n \xrightarrow{P} \theta_0$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{\theta_0}(X_i) + o_P(1),$$

In particular, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix $V_{\theta_0}^{-1} P\psi_{\theta_0}\psi_{\theta_0}^T (V_{\theta_0}^{-1})^T$.

Proof. For a fixed measurable function f, we abbreviate $\sqrt{n}(\mathbb{P}_n - P)f$ to $\mathbb{G}_n f$, the empirical process evaluated at f. The consistency of $\hat{\theta}_n$ and the Lipschitz condition on the maps $\theta \mapsto \psi_\theta$ imply that

$$\mathbb{G}_n\psi_{\hat{\theta}_n} - \mathbb{G}_n\psi_{\theta_0} \xrightarrow{P} 0. \tag{5.22}$$

For a nonrandom sequence $\hat{\theta}_n$ this is immediate from the fact that the means of these variables are zero, while the variances are bounded by $P\|\psi_{\theta_n} - \psi_{\theta_0}\|^2 \le P\dot{\psi}^2\|\theta_n - \theta_0\|^2$ and hence converge to zero. A proof for estimators $\hat{\theta}_n$ under the present mild conditions takes more effort. The appropriate tools are developed in Chapter 19. In Example 19.7 it is seen that the functions ψ_θ form a Donsker class. Next, (5.22) follows from Lemma 19.24. Here we accept the convergence as a fact and give the remainder of the proof.

By the definitions of $\hat{\theta}_n$ and θ_0, we can rewrite $\mathbb{G}_n\psi_{\hat{\theta}_n}$ as $\sqrt{n}P(\psi_{\theta_0} - \psi_{\hat{\theta}_n}) + o_P(1)$. Combining this with the delta method (or Lemma 2.12) and the differentiability of the map $\theta \mapsto P\psi_\theta$, we find that

$$\sqrt{n}V_{\theta_0}(\theta_0 - \hat{\theta}_n) + \sqrt{n}\,o_P\left(\|\hat{\theta}_n - \theta_0\|\right) = \mathbb{G}_n\psi_{\theta_0} + o_P(1).$$

In particular, by the invertibility of the matrix V_{θ_0},

$$\sqrt{n}\|\hat{\theta}_n - \theta_0\| \leq \left\|V_{\theta_0}^{-1}\right\| \left\|\sqrt{n}\, V_{\theta_0}(\hat{\theta}_n - \theta_0)\right\| = O_P(1) + o_P\bigl(\sqrt{n}\|\hat{\theta}_n - \theta_0\|\bigr).$$

This implies that $\hat{\theta}_n$ is $\sqrt{n}$-consistent: The left side is bounded in probability. Inserting this in the previous display, we obtain that $\sqrt{n}V_{\theta_0}(\hat{\theta}_n - \theta_0) = -\mathbb{G}_n\psi_{\theta_0} + o_P(1)$. We conclude the proof by taking the inverse $V_{\theta_0}^{-1}$ left and right. Because matrix multiplication is a continous map, the inverse of the remainder term still converges to zero in probability. ■

The preceding theorem is a reasonable compromise between simplicity and general applicability, but, unfortunately, it does not cover the sample median. Because the function $\theta \mapsto \operatorname{sign}(x - \theta)$ is not Lipschitz, the Lipschitz condition is apparently still stronger than necessary. Inspection of the proof shows that it is used only to ensure (5.22). It is seen in Lemma 19.24, that (5.22) can be ascertained under the weaker conditions that the collection of functions $x \mapsto \psi_\theta(x)$ are a "Donsker class" and that the map $\theta \mapsto \psi_\theta$ is continuous in probability. The functions $\operatorname{sign}(x - \theta)$ do satisfy these conditions, but a proof and the definition of a Donsker class are deferred to Chapter 19.

If the functions $\theta \mapsto \psi_\theta(x)$ are continuously differentiable, then the natural candidate for $\dot{\psi}(x)$ is $\sup_\theta \|\dot{\psi}_\theta\|$, with the supremum taken over a neighborhood of θ_0. Then the main condition is that the partial derivatives are "locally dominated" by a square-integrable function: There should exist a square-integrable function $\dot{\psi}$ with $\|\dot{\psi}_\theta\| \leq \dot{\psi}$ for every θ close to θ_0. If $\theta \mapsto \dot{\psi}_\theta(x)$ is also continuous at θ_0, then the dominated-convergence theorem readily yields that $V_{\theta_0} = P\dot{\psi}_{\theta_0}$.

The properties of M estimators can typically be obtained under milder conditions by using their characterization as maximizers. The following theorem is in the same spirit as the preceding one but does cover the median. It concerns M-estimators defined as maximizers of a criterion function $\theta \mapsto \mathbb{P}_n m_\theta$, which are assumed to be consistent for a point of maximum θ_0 of the function $\theta \mapsto Pm_\theta$. If the latter function is twice continuously differentiable at θ_0, then, of course, it allows a two-term Taylor expansion of the form

$$Pm_\theta = Pm_{\theta_0} + \tfrac{1}{2}(\theta - \theta_0)^T V_{\theta_0}(\theta - \theta_0) + o\bigl(\|\theta - \theta_0\|^2\bigr).$$

It is this expansion rather than the differentiability that is needed in the following theorem.

5.23 Theorem. *For each θ in an open subset of Euclidean space let $x \mapsto m_\theta(x)$ be a measurable function such that $\theta \mapsto m_\theta(x)$ is differentiable at θ_0 for P-almost every x[†] with derivative $\dot{m}_{\theta_0}(x)$ and such that, for every θ_1 and θ_2 in a neighborhood of θ_0 and a measurable function $\dot{m}$ with $P\dot{m}^2 < \infty$*

$$\bigl|m_{\theta_1}(x) - m_{\theta_2}(x)\bigr| \leq \dot{m}(x)\,\|\theta_1 - \theta_2\|.$$

Furthermore, assume that the map $\theta \mapsto Pm_\theta$ admits a second-order Taylor expansion at a point of maximum θ_0 with nonsingular symmetric second derivative matrix V_{θ_0}. If $\mathbb{P}_n m_{\hat{\theta}_n} \geq \sup_\theta \mathbb{P}_n m_\theta - o_P(n^{-1})$ and $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{m}_{\theta_0}(X_i) + o_P(1).$$

[†] Alternatively, it suffices that $\theta \mapsto m_\theta$ is differentiable at θ_0 in P-probability.

In particular, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix $V_{\theta_0}^{-1} P\dot{m}_{\theta_0}\dot{m}_{\theta_0}^T V_{\theta_0}^{-1}$.

***Proof.** The Lipschitz property and the differentiability of the maps $\theta \mapsto m_\theta$ imply that, for every random sequence $\tilde{h}_n$ that is bounded in probability,

$$\mathbb{G}_n\left[\sqrt{n}\left(m_{\theta_0+\tilde{h}_n/\sqrt{n}} - m_{\theta_0}\right) - \tilde{h}_n^T\dot{m}_{\theta_0}\right] \xrightarrow{\mathrm{P}} 0.$$

For nonrandom sequences $\tilde{h}_n$ this follows, because the variables have zero means, and variances that converge to zero, by the dominated convergence theorem. For general sequences $\tilde{h}_n$ this follows from Lemma 19.31.

A second fact that we need and that is proved subsequently is the $\sqrt{n}$-consistency of the sequence $\hat{\theta}_n$. By Corollary 5.53, the Lipschitz condition, and the twice differentiability of the map $\theta \mapsto Pm_\theta$, the sequence $\sqrt{n}(\hat{\theta}_n - \theta)$ is bounded in probability.

The remainder of the proof is self-contained. In view of the twice differentiability of the map $\theta \mapsto Pm_\theta$, the preceding display can be rewritten as

$$n\mathbb{P}_n\left(m_{\theta_0+\tilde{h}_n/\sqrt{n}} - m_{\theta_0}\right) = \tfrac{1}{2}\tilde{h}_n^T V_{\theta_0}\tilde{h}_n + \tilde{h}_n^T\mathbb{G}_n\dot{m}_{\theta_0} + o_P(1).$$

Because the sequence $\hat{\theta}_n$ is $\sqrt{n}$-consistent, this is valid both for $\tilde{h}_n$ equal to $\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$ and for $\tilde{h}_n = -V_{\theta_0}^{-1}\mathbb{G}_n\dot{m}_{\theta_0}$. After simple algebra in the second case, we obtain the equations

$$n\mathbb{P}_n\left(m_{\theta_0+\hat{h}_n/\sqrt{n}} - m_{\theta_0}\right) = \frac{1}{2}\hat{h}_n^T V_{\theta_0}\hat{h}_n + \hat{h}_n^T\mathbb{G}_n\dot{m}_{\theta_0} + o_P(1),$$

$$n\mathbb{P}_n\left(m_{\theta_0-V_{\theta_0}^{-1}\mathbb{G}_n\dot{m}_{\theta_0}/\sqrt{n}} - m_{\theta_0}\right) = -\frac{1}{2}\mathbb{G}_n\dot{m}_{\theta_0}^T V_{\theta_0}^{-1}\mathbb{G}_n\dot{m}_{\theta_0} + o_P(1).$$

By the definition of $\hat{\theta}_n$, the left side of the first equation is larger than the left side of the second equation (up to $o_P(1)$) and hence the same relation is true for the right sides. Take the difference, complete the square, and conclude that

$$\frac{1}{2}\left(\hat{h}_n + V_{\theta_0}^{-1}\mathbb{G}_n\dot{m}_{\theta_0}\right)^T V_{\theta_0}\left(\hat{h}_n + V_{\theta_0}^{-1}\mathbb{G}_n\dot{m}_{\theta_0}\right) + o_P(1) \geq 0.$$

Because the matrix V_{θ_0} is strictly negative-definite, the quadratic form must converge to zero in probability. The same must be true for $\|\hat{h}_n + V_{\theta_0}^{-1}\mathbb{G}_n\dot{m}_{\theta_0}\|$. ■

The assertions of the preceding theorems must be in agreement with each other and also with the informal derivation leading to (5.20). If $\theta \mapsto m_\theta(x)$ is differentiable, then a maximizer of $\theta \mapsto \mathbb{P}_n m_\theta$ typically solves $\mathbb{P}_n\psi_\theta = 0$ for $\psi_\theta = \dot{m}_\theta$. Then the theorems and (5.20) are in agreement provided that

$$V_\theta = \frac{\partial^2}{\partial\theta^2}Pm_\theta = \frac{\partial}{\partial\theta}P\psi_\theta = P\dot{\psi}_\theta = P\ddot{m}_\theta.$$

This involves changing the order of differentiation (with respect to θ) and integration (with respect to x), and is usually permitted. However, for instance, the second derivative of Pm_θ may exist without $\theta \mapsto m_\theta(x)$ being differentiable for all x, as is seen in the following example.

5.24 ***Example (Median).*** The sample median maximizes the criterion function $\theta \mapsto -\sum_{i=1}^n |X_i - \theta|$. Assume that the distribution function F of the observations is differentiable

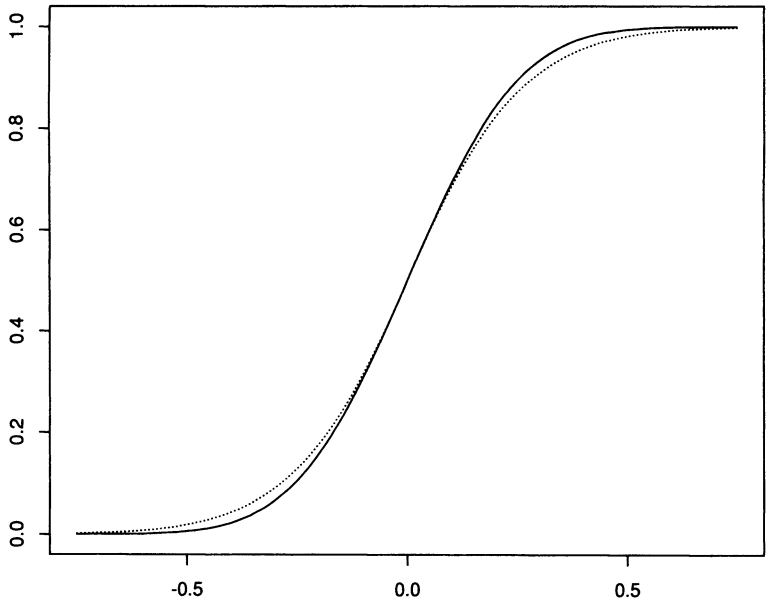

Figure 5.3. The distribution function of the sample median (dotted curve) and its normal approximation for a sample of size 25 from the Laplace distribution.

at its median θ_0 with positive derivative $f(\theta_0)$. Then the sample median is asymptotically normal.

This follows from Theorem 5.23 applied with $m_\theta(x) = |x-\theta| - |x|$. As a consequence of the triangle inequality, this function satisfies the Lipschitz condition with $\dot{m}(x) \equiv 1$. Furthermore, the map $\theta \mapsto m_\theta(x)$ is differentiable at θ_0 except if $x = \theta_0$, with $\dot{m}_{\theta_0}(x) = -\text{sign}(x - \theta_0)$. By partial integration,

$$Pm_\theta = \theta F(0) + \int_{(0,\theta]} (\theta - 2x)\, dF(x) - \theta\big(1 - F(\theta)\big) = 2\int_0^\theta F(x)\, dx - \theta.$$

If F is sufficiently regular around θ_0, then Pm_θ is twice differentiable with first derivative $2F(\theta) - 1$ (which vanishes at θ_0) and second derivative $2f(\theta)$. More generally, under the minimal condition that F is differentiable at θ_0, the function Pm_θ has a Taylor expansion $Pm_{\theta_0} + \frac{1}{2}(\theta - \theta_0)^2 2f(\theta_0) + o\big(|\theta - \theta_0|^2\big)$, so that we set $V_{\theta_0} = 2f(\theta_0)$. Because $P\dot{m}^2_{\theta_0} = \text{E1} = 1$, the asymptotic variance of the median is $1/\big(2f(\theta_0)\big)^2$. Figure 5.3 gives an impression of the accuracy of the approximation. □

5.25 ***Example (Misspecified model).*** Suppose an experimenter postulates a model $\{p_\theta : \theta \in \Theta\}$ for a sample of observations $X_1, \ldots, X_n$. However, the model is misspecified in that the true underlying distribution does not belong to the model. The experimenter decides to use the postulated model anyway, and obtains an estimate $\hat{\theta}_n$ from maximizing the likelihood $\sum \log p_\theta(X_i)$. What is the asymptotic behaviour of $\hat{\theta}_n$?

At first sight, it might appear that $\hat{\theta}_n$ would behave erratically due to the use of the wrong model. However, this is not the case. First, we expect that $\hat{\theta}_n$ is asymptotically consistent for a value θ_0 that maximizes the function $\theta \mapsto P \log p_\theta$, where the expectation is taken under the true underlying distribution P. The density p_{θ_0} can be viewed as the "projection"

of the true underlying distribution P on the model using the *Kullback-Leibler divergence*, which is defined as $-P\log(p_\theta/p)$, as a "distance" measure: p_{θ_0} minimizes this quantity over all densities in the model. Second, we expect that $\sqrt{n}(\hat\theta_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix

$$V_{\theta_0}^{-1} P\dot\ell_{\theta_0}\dot\ell_{\theta_0}^T V_{\theta_0}^{-1}.$$

Here $\ell_\theta = \log p_\theta$, and V_{θ_0} is the second derivative matrix of the map $\theta \mapsto P\log p_\theta$. The preceding theorem with $m_\theta = \log p_\theta$ gives sufficient conditions for this to be true.

The asymptotics give insight into the practical value of the experimenter's estimate $\hat\theta_n$. This depends on the specific situation. However, if the model is not too far off from the truth, then the estimated density $p_{\hat\theta_n}$ may be a reasonable approximation for the true density. □

5.26 ***Example (Exponential frailty model).*** Suppose that the observations are a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of pairs of survival times. For instance, each X_i is the survival time of a "father" and Y_i the survival time of a "son." We assume that given an unobservable value z_i, the survival times X_i and Y_i are independent and exponentially distributed with parameters z_i and θz_i, respectively. The value z_i may be different for each observation. The problem is to estimate the ratio θ of the parameters.

To fit this example into the i.i.d. set-up of this chapter, we assume that the values $z_1, \ldots, z_n$ are realizations of a random sample $Z_1, \ldots, Z_n$ from some given distribution (that we do not have to know or parametrize).

One approach is based on the sufficiency of the variable $X_i + \theta Y_i$ for z_i in the case that θ is known. Given $Z_i = z$, this "statistic" possesses the gamma-distribution with shape parameter 2 and scale parameter z. Corresponding to this, the conditional density of an observation (X, Y) factorizes, for a given z, as $h_\theta(x, y)\, g_\theta(x + \theta y \mid z)$, for $g_\theta(s \mid z) = z^2 s e^{-zs}$ the gamma-density and

$$h_\theta(x, y) = \frac{\theta}{x + \theta y}.$$

Because the density of $X_i + \theta Y_i$ depends on the unobservable value z_i, we might wish to discard the factor $g_\theta(s \mid z)$ from the likelihood and use the factor $h_\theta(x, y)$ only. Unfortunately, this "conditional likelihood" does not behave as an ordinary likelihood, in that the corresponding "conditional likelihood equation," based on the function $\dot h_\theta / h_\theta(x, y) = \partial/\partial\theta \log h_\theta(x, y)$, does not have mean zero under θ. The bias can be corrected by conditioning on the sufficient statistic. Let

$$\psi_\theta(X, Y) = 2\theta\frac{\dot h_\theta}{h_\theta}(X, Y) - 2\theta \mathrm{E}_\theta\left(\frac{\dot h_\theta}{h_\theta}(X, Y) \mid X + \theta Y\right) = \frac{X - \theta Y}{X + \theta Y}.$$

Next define an estimator $\hat\theta_n$ as the solution of $\mathbb{P}_n\psi_\theta = 0$.

This works fairly nicely. Because the function $\theta \mapsto \psi_\theta(x, y)$ is continuous, and decreases strictly from 1 to -1 on $(0, \infty)$ for every $x, y > 0$, the equation $\mathbb{P}_n\psi_\theta = 0$ has a unique solution. The sequence of solutions $\hat\theta_n$ can be seen to be consistent by Lemma 5.10. By straightforward calculation, as $\theta \to \theta_0$,

$$P_{\theta_0}\psi_\theta = -\frac{\theta + \theta_0}{\theta - \theta_0} - \frac{2\theta\theta_0}{(\theta - \theta_0)^2}\log\frac{\theta_0}{\theta} = \frac{1}{3\theta_0}(\theta_0 - \theta) + o(\theta_0 - \theta).$$

Hence the zero of $\theta \mapsto P_{\theta_0}\psi_\theta$ is taken uniquely at $\theta = \theta_0$. Next, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ can be shown to be asymptotically normal by Theorem 5.21. In fact, the functions $\dot{\psi}_\theta(x, y)$ are uniformly bounded in $x, y > 0$ and θ ranging over compacta in $(0, \infty)$, so that, by the mean value theorem, the function $\dot{\psi}$ in this theorem may be taken equal to a constant.

On the other hand, although this estimator is easy to compute, it can be shown that it is not asymptotically optimal. In Chapter 25 on semiparametric models, we discuss estimators with a smaller asymptotic variance. □

5.27 ***Example (Nonlinear least squares).*** Suppose that we observe a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the distribution of a vector (X, Y) that follows the regression model

$$Y = f_{\theta_0}(X) + e, \qquad \mathrm{E}(e \mid X) = 0.$$

Here f_θ is a parametric family of regression functions, for instance $f_\theta(x) = \theta_1 + \theta_2 e^{\theta_3 x}$, and we aim at estimating the unknown vector θ. (We assume that the independent variables are a random sample in order to fit the example in our i.i.d. notation, but the analysis could be carried out conditionally as well.) The least squares estimator that minimizes

$$\theta \mapsto \sum_{i=1}^{n} \big(Y_i - f_\theta(X_i)\big)^2$$

is an M-estimator for $m_\theta(x, y) = \big(y - f_\theta(x)\big)^2$ (or rather minus this function). It should be expected to converge to the minimizer of the limit criterion function

$$\theta \mapsto Pm_\theta = P(f_{\theta_0} - f_\theta)^2 + \mathrm{E}e^2.$$

Thus the least squares estimator should be consistent if θ_0 is identifiable from the model, in the sense that $\theta \neq \theta_0$ implies that $f_\theta(X) \neq f_{\theta_0}(X)$ with positive probability.

For sufficiently regular regression models, we have

$$Pm_\theta \approx P\big((\theta - \theta_0)^T \dot{f}_{\theta_0}\big)^2 + \mathrm{E}e^2.$$

This suggests that the conditions of Theorem 5.23 are satisfied with $V_{\theta_0} = 2P\dot{f}_{\theta_0}\dot{f}_{\theta_0}^T$ and $\dot{m}_{\theta_0}(x, y) = -2\big(y - f_{\theta_0}(x)\big)\dot{f}_{\theta_0}(x)$. If e and X are independent, then this leads to the asymptotic covariance matrix $V_{\theta_0}^{-1} 2\mathrm{E}e^2$. □

Besides giving the asymptotic normality of $\sqrt{n}(\hat{\theta}_n - \theta_0)$, the preceding theorems give an asymptotic representation

$$\hat{\theta}_n = \theta_0 + \frac{1}{n}\sum_{i=1}^{n} V_{\theta_0}^{-1}\psi_{\theta_0}(X_i) + o_P\left(\frac{1}{\sqrt{n}}\right).$$

If we neglect the remainder term,† then this means that $\hat{\theta}_n - \theta_0$ behaves as the average of the variables $V_{\theta_0}^{-1}\psi_{\theta_0}(X_i)$. Then the (asymptotic) "influence" of the nth observation on the

† To make the following derivation rigorous, more information concerning the remainder term would be necessary.

value of $\hat{\theta}_n$ can be computed as

$$\hat{\theta}_n(X_1, \ldots, X_n) - \hat{\theta}_{n-1}(X_1, \ldots, X_{n-1}) \approx \frac{1}{n} V_{\theta_0}^{-1} \psi_{\theta_0}(X_n) - \frac{1}{n(n-1)} \sum_{i=1}^{n-1} V_{\theta_0}^{-1} \psi_{\theta_0}(X_i)$$
$$= \frac{1}{n} V_{\theta_0}^{-1} \psi_{\theta_0}(X_n) + o_P\left(\frac{1}{n}\right).$$

Because the "influence" of an extra observation x is proportional to $V_\theta^{-1}\psi_\theta(x)$, the function $x \mapsto V_\theta^{-1}\psi_\theta(x)$ is called the *asymptotic influence function* of the estimator $\hat{\theta}_n$. Influence functions can be defined for many other estimators as well, but the method of Z-estimation is particularly convenient to obtain estimators with given influence functions. Because V_{θ_0} is a constant (matrix), any shape of influence function can be obtained by simply choosing the right functions ψ_θ.

For the purpose of robust estimation, perhaps the most important aim is to bound the influence of each individual observation. Thus, a Z-estimator is called *B-robust* if the function ψ_θ is bounded.

5.28 ***Example (Robust regression).*** Consider a random sample of observations (X_1, Y_1), $\ldots$, (X_n, Y_n) following the linear regression model

$$Y_i = \theta_0^T X_i + e_i,$$

for i.i.d. errors $e_1, \ldots, e_n$ that are independent of $X_1, \ldots, X_n$. The classical estimator for the regression parameter θ is the least squares estimator, which minimizes $\sum_{i=1}^n (Y_i - \theta^T X_i)^2$. Outlying values of X_i ("leverage points") or extreme values of (X_i, Y_i) jointly ("influence points") can have an arbitrarily large influence on the value of the least-squares estimator, which therefore is nonrobust. As in the case of location estimators, a more robust estimator for θ can be obtained by replacing the square by a function $m(x)$ that grows less rapidly as $x \to \infty$, for instance $m(x) = |x|$ or $m(x)$ equal to the primitive function of Huber's ψ. Usually, minimizing an expression of the type $\sum_{i=1}^n m(Y_i - \theta X_i)$ is equivalent to solving a system of equations

$$\sum_{i=1}^n \psi\big(Y_i - \theta^T X_i\big) X_i = 0.$$

Because $\mathrm{E}\psi(Y - \theta_0^T X)X = \mathrm{E}\psi(e)\mathrm{E}X$, we can expect the resulting estimator to be consistent provided $\mathrm{E}\psi(e) = 0$. Furthermore, we should expect that, for $V_{\theta_0} = \mathrm{E}\psi'(e)XX^T$,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} V_{\theta_0}^{-1} \sum_{i=1}^n \psi\big(Y_i - \theta_0^T X_i\big) X_i + o_P(1).$$

Consequently, even for a bounded function ψ, the influence function $(x, y) \mapsto V_\theta^{-1}\psi(y - \theta^T x)x$ may be unbounded, and an extreme value of an X_i may still have an arbitrarily large influence on the estimate (asymptotically). Thus, the estimators obtained in this way are protected against influence points but may still suffer from leverage points and hence are only partly robust. To obtain fully robust estimators, we can change the estimating

equations to

$$\sum_{i=1}^{n} \psi\big((Y_i - \theta^T X_i)v(X_i)\big)w(X_i) = 0.$$

Here we protect against leverage points by choosing w bounded. For more flexibility we have also allowed a weighting factor $v(X_i)$ inside ψ. The choices $\psi(x) = x$, $v(x) = 1$ and $w(x) = x$ correspond to the (nonrobust) least-squares estimator.

The solution $\hat{\theta}_n$ of our final estimating equation should be expected to be consistent for the solution of

$$0 = \mathrm{E}\psi\big((Y - \theta^T X)v(X)\big)w(X) = \mathrm{E}\psi\Big(\big(e + \theta_0^T X - \theta^T X\big)v(X)\Big)w(X).$$

If the function ψ is odd and the error symmetric, then the true value θ_0 will be a solution whenever e is symmetric about zero, because then $\mathrm{E}\psi(e\sigma) = 0$ for every σ.

Precise conditions for the asymptotic normality of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ can be obtained from Theorems 5.21 and 5.9. The verification of the conditions of Theorem 5.21, which are "local" in nature, is relatively easy, and, if necessary, the Lipschitz condition can be relaxed by using results on empirical processes introduced in Chapter 19 directly. Perhaps proving the consistency of $\hat{\theta}_n$ is harder. The biggest technical problem may be to show that $\hat{\theta}_n = O_P(1)$, so it would help if θ could a priori be restricted to a bounded set. On the other hand, for bounded functions ψ, the case of most interest in the present context, the functions $(x, y) \mapsto \psi\big((y - \theta^T x)v(x)\big)w(x)$ readily form a Glivenko-Cantelli class when θ ranges freely, so that verification of the strong uniqueness of θ_0 as a zero becomes the main challenge when applying Theorem 5.9. This leads to a combination of conditions on ψ, v, w, and the distributions of e and X. □

5.29 ***Example (Optimal robust estimators).*** Every sufficiently regular function ψ defines a location estimator $\hat{\theta}_n$ through the equation $\sum_{i=1}^{n} \psi(X_i - \theta) = 0$. In order to choose among the different estimators, we could compare their asymptotic variances and use the one with the smallest variance under the postulated (or estimated) distribution P of the observations. On the other hand, if we also wish to guard against extreme obervations, then we should find a balance between robustness and asymptotic variance. One possibility is to use the estimator with the smallest asymptotic variance at the postulated, ideal distribution P under the side condition that its influence function be uniformly bounded by some constant c. In this example we show that for P the normal distribution, this leads to the Huber estimator.

The Z-estimator is consistent for the solution θ_0 of the equation $P\psi(\cdot - \theta) = \mathrm{E}\psi(X_1 - \theta) = 0$. Suppose that we fix an underlying, ideal P whose "location" θ_0 is zero. Then the problem is to find ψ that minimizes the asymptotic variance $P\psi^2/(P\psi')^2$ under the two side conditions, for a given constant c,

$$\sup_x \left|\frac{\psi(x)}{P\psi'}\right| \le c, \quad \text{and} \quad P\psi = 0.$$

The problem is homogeneous in ψ, and hence we may assume that $P\psi' = 1$ without loss of generality. Next, minimization of $P\psi^2$ under the side conditions $P\psi = 0$, $P\psi' = 1$ and $\|\psi\|_\infty \le c$ can be achieved by using Lagrange multipliers, as in problem 14.6 This leads to minimizing

$$P\psi^2 + \lambda P\psi + \mu(P\psi' - 1) = P\Big(\psi^2 + \psi\big(\lambda + \mu(p'/p)\big) - \mu\Big)$$

for fixed "multipliers" λ and μ under the side condition $\|\psi\|_\infty \le c$ with respect to ψ. This expectation is minimized by minimizing the integrand pointwise, for every fixed x. Thus the minimizing ψ has the property that, for every x separately, $y = \psi(x)$ minimizes the parabola $y^2 + \lambda y + \mu y(p'/p)(x)$ over $y \in [-c, c]$. This readily gives the solution, with $[y]_c^d$ the value y truncated to the interval $[c, d]$,

$$\psi(x) = \left[-\frac{1}{2}\lambda - \frac{1}{2}\mu\frac{p'}{p}(x)\right]_{-c}^{c}.$$

The constants λ and μ can be solved from the side conditions $P\psi = 0$ and $P\psi' = 1$. The normal distribution $P = \Phi$ has location score function $p'/p(x) = -x$, and by symmetry it follows that $\lambda = 0$ in this case. Then the optimal ψ reduces to Huber's ψ function. □

*5.4 Estimated Parameters

In many situations, the estimating equations for the parameters of interest contain preliminary estimates for "nuisance parameters." For example, many robust location estimators are defined as the solutions of equations of the type

$$\sum_{i=1}^{n} \psi\left(\frac{X_i - \theta}{\hat{\sigma}}\right) = 0. \tag{5.30}$$

Here $\hat{\sigma}$ is an initial (robust) estimator of scale, which is meant to stabilize the robustness of the location estimator. For instance, the "cut-off" parameter k in Huber's ψ-function determines the amount of robustness of Huber's estimator, but the effect of a particular choice of k on bounding the influence of outlying observations is relative to the range of the observations. If the observations are concentrated in the interval $[-k, k]$, then Huber's ψ yields nothing else but the sample mean, if all observations are outside $[-k, k]$, we get the median. Scaling the observations to a standard scale gives a clear meaning to the value of k. The use of the *median absolute deviation from the median* (see. section 21.3) is often recommended for this purpose.

If the scale estimator is itself a Z-estimator, then we can treat the pair $(\hat{\theta}, \hat{\sigma})$ as a Z-estimator for a system of equations, and next apply the preceding theorems. More generally, we can apply the following result. In this subsection we allow a condition in terms of Donsker classes, which are discussed in Chapter 19. The proof of the following theorem follows the same steps as the proof of Theorem 5.21.

5.31 *Theorem.* *For each θ in an open subset of $\mathbb{R}^k$ and each η in a metric space, let $x \mapsto \psi_{\theta,\eta}(x)$ be an $\mathbb{R}^k$-valued measurable function such that the class of functions $\{\psi_{\theta,\eta} : \|\theta - \theta_0\| < \delta, d(\eta, \eta_o) < \delta\}$ is Donsker for some $\delta > 0$, and such that $P\|\psi_{\theta,\eta} - \psi_{\theta_0,\eta_0}\|^2 \to 0$ as $(\theta, \eta) \to (\theta_0, \eta_0)$. Assume that $P\psi_{\theta_0,\eta_0} = 0$, and that the maps $\theta \mapsto P\psi_{\theta,\eta}$ are differentiable at θ_0, uniformly in η in a neighborhood of η_0 with nonsingular derivative matrices $V_{\theta_0,\eta}$ such that $V_{\theta_0,\eta} \to V_{\theta_0,\eta_0}$. If $\sqrt{n}\,\mathbb{P}_n\psi_{\hat{\theta}_n,\hat{\eta}_n} = o_P(1)$ and $(\hat{\theta}_n, \hat{\eta}_n) \xrightarrow{P} (\theta_0, \eta_0)$, then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0,\eta_0}^{-1}\sqrt{n}\,P\psi_{\theta_0,\hat{\eta}_n} - V_{\theta_0,\eta_0}^{-1}\mathbb{G}_n\psi_{\theta_0,\eta_0} + o_P(1 + \sqrt{n}\|P\psi_{\theta_0,\hat{\eta}_n}\|).$$

Under the conditions of this theorem, the limiting distribution of the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ depends on the estimator $\hat{\eta}_n$ through the "drift" term $\sqrt{n}P\psi_{\theta_0,\hat{\eta}_n}$. In general, this gives a contribution to the limiting distribution, and $\hat{\eta}_n$ must be chosen with care. If $\hat{\eta}_n$ is $\sqrt{n}$-consistent and the map $\eta \mapsto P\psi_{\theta_0,\eta}$ is differentiable, then the drift term can be analyzed using the delta-method.

It may happen that the drift term is zero. If the parameters θ and η are "orthogonal" in this sense, then the auxiliary estimators $\hat{\eta}_n$ may converge at an arbitrarily slow rate and affect the limit distribution of $\hat{\theta}_n$ only through their limiting value η_0.

5.32 ***Example (Symmetric location).*** Suppose that the distribution of the observations is symmetric about θ_0. Let $x \mapsto \psi(x)$ be an antisymmetric function, and consider the Z-estimators that solve equation (5.30). Because $P\psi\big((X - \theta_0)/\sigma\big) = 0$ for every σ, by the symmetry of P and the antisymmetry of ψ, the "drift term" due to $\hat{\eta}$ in the preceding theorem is identically zero. The estimator $\hat{\theta}_n$ has the same limiting distribution whether we use an arbitrary consistent estimator of a "true scale" σ_0 or σ_0 itself. □

5.33 ***Example (Robust regression).*** In the linear regression model considered in Example 5.28, suppose that we choose the weight functions v and w dependent on the data and solve the robust estimator $\hat{\theta}_n$ of the regression parameters from

$$0 = \frac{1}{n}\sum_{i=1}^{n} \psi\big((Y_i - \theta^T X_i)\hat{v}_n(X_i)\big)\hat{w}_n(X_i).$$

This corresponds to defining a nuisance parameter $\eta = (v, w)$ and setting $\psi_{\theta,v,w}(x, y) = \psi\big((y - \theta^T x)v(x)\big)w(x)$. If the functions $\psi_{\theta,v,w}$ run through a Donsker class (and they easily do), and are continuous in (θ, v, w), and the map $\theta \mapsto P\psi_{\theta,v,w}$ is differentiable at θ_0 uniformly in (v, w), then the preceding theorem applies. If $\mathrm{E}\psi(e\sigma) = 0$ for every σ, then $P\psi_{\theta_0,v,w} = 0$ for any v and w, and the limit distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is the same, whether we use the random weight functions $(\hat{v}_n, \hat{w}_n)$ or their limit (v_0, w_0) (assuming that this exists).

The purpose of using random weight functions could be, besides stabilizing the robustness, to improve the asymptotic efficiency of $\hat{\theta}_n$. The limit (v_0, w_0) typically is not the same for every underlying distribution P, and the estimators $(\hat{v}_n, \hat{w}_n)$ can be chosen in such a way that the asymptotic variance is minimal. □

5.5 Maximum Likelihood Estimators

Maximum likelihood estimators are examples of M-estimators. In this section we specialize the consistency and the asymptotic normality results of the preceding sections to this important special case. Our approach reverses the historical order. Maximum likelihood estimators were shown to be asymptotically normal first by Fisher in the 1920s and rigorously by Cramér, among others, in the 1940s. General M-estimators were not introduced and studied systematically until the 1960s, when they became essential in the development of robust estimators.

If $X_1, \ldots, X_n$ are a random sample from a density p_θ, then the maximum likelihood estimator $\hat{\theta}_n$ maximizes the function $\theta \mapsto \sum \log p_\theta(X_i)$, or equivalently, the function

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_\theta}{p_{\theta_0}}(X_i) = \mathbb{P}_n \log \frac{p_\theta}{p_{\theta_0}}.$$

(Subtraction of the "constant" $\sum \log p_{\theta_0}(X_i)$ turns out to be mathematically convenient.) If we agree that $\log 0 = -\infty$, then this expression is with probability 1 well defined if p_{θ_0} is the true density. The asymptotic function corresponding to M_n is[†]

$$M(\theta) = \mathrm{E}_{\theta_0} \log \frac{p_\theta}{p_{\theta_0}}(X) = P_{\theta_0} \log \frac{p_\theta}{p_{\theta_0}}.$$

The number $-M(\theta)$ is called the *Kullback-Leibler divergence* of p_θ and p_{θ_0}; it is often considered a measure of "distance" between p_θ and p_{θ_0}, although it does not have the properties of a mathematical distance. Based on the results of the previous sections, we may expect the maximum likelihood estimator to converge to a point of maximum of $M(\theta)$. Is the true value θ_0 always a point of maximum? The answer is affirmative, and, moreover, the true value is a unique point of maximum if the true measure is *identifiable*:

$$P_\theta \neq P_{\theta_0}, \qquad \text{every } \theta \neq \theta_0. \tag{5.34}$$

This requires that the model for the observations is not the same under the parameters θ and θ_0. Identifiability is a natural and even a necessary condition: If the parameter is not identifiable, then consistent estimators cannot exist.

5.35 ***Lemma.*** *Let $\{p_\theta : \theta \in \Theta\}$ be a collection of subprobability densities such that* (5.34) *holds and such that P_{θ_0} is a probability measure. Then $M(\theta) = P_{\theta_0} \log p_\theta / p_{\theta_0}$ attains its maximum uniquely at θ_0.*

Proof. First note that $M(\theta_0) = P_{\theta_0} \log 1 = 0$. Hence we wish to show that $M(\theta)$ is strictly negative for $\theta \neq \theta_0$.

Because $\log x \leq 2(\sqrt{x} - 1)$ for every $x \geq 0$, we have, writing μ for the dominating measure,

$$\begin{aligned} P_{\theta_0} \log \frac{p_\theta}{p_{\theta_0}} &\leq 2P_{\theta_0}\left(\sqrt{\frac{p_\theta}{p_{\theta_0}}} - 1\right) = 2\int \sqrt{p_\theta p_{\theta_0}}\, d\mu - 2 \\ &\leq -\int (\sqrt{p_\theta} - \sqrt{p_{\theta_0}})^2\, d\mu. \end{aligned}$$

(The last inequality is an equality if $\int p_\theta\, d\mu = 1$.) This is always nonpositive, and is zero only if p_θ and p_{θ_0} are equal. By assumption the latter happens only if $\theta = \theta_0$. ■

Thus, under conditions such as in section 5.2 and identifiability, the sequence of maximum likelihood estimators is consistent for the true parameter.

[†] Presently we take the expectation P_{θ_0} under the parameter θ_0, whereas the derivation in section 5.3 is valid for a generic underlying probability structure and does not conceptually require that the set of parameters θ indexes a set of underlying distributions.

This conclusion is derived from viewing the maximum likelihood estimator as an M-estimator for $m_\theta = \log p_\theta$. Sometimes it is technically advantageous to use a different starting point. For instance, consider the function

$$m_\theta = \log \frac{p_\theta + p_{\theta_0}}{2p_{\theta_0}}.$$

By the concavity of the logarithm, the maximum likelihood estimator $\hat{\theta}$ satisfies

$$\mathbb{P}_n m_{\hat{\theta}} \geq \mathbb{P}_n \frac{1}{2} \log \frac{p_{\hat{\theta}}}{p_{\theta_0}} + \mathbb{P}_n \frac{1}{2} \log 1 \geq 0 = \mathbb{P}_n m_{\theta_0}.$$

Even though $\hat{\theta}$ does not maximize $\theta \mapsto \mathbb{P}_n m_\theta$, this inequality can be used as the starting point for a consistency proof, since Theorem 5.7 requires that $M_n(\hat{\theta}) \geq M_n(\theta_0) - o_P(1)$ only. The true parameter is still identifiable from this criterion function, because, by the preceding lemma, $P_{\theta_0} m_\theta = 0$ implies that $(p_\theta + p_{\theta_0})/2 = p_{\theta_0}$, or $p_\theta = p_{\theta_0}$. A technical advantage is that $m_\theta \geq \log(1/2)$. For another variation, see Example 5.17.

Consider asymptotic normality. The maximum likelihood estimator solves the likelihood equations

$$\frac{\partial}{\partial \theta} \sum_{i=1}^{n} \log p_\theta(X_i) = 0.$$

Hence it is a Z-estimator for ψ_θ equal to the *score function* $\dot{\ell}_\theta = \partial/\partial\theta \log p_\theta$ of the model. In view of the results of section 5.3, we expect that the sequence $\sqrt{n}(\hat{\theta}_n - \theta)$ is, under θ, asymptotically normal with mean zero and covariance matrix

$$(P_\theta \ddot{\ell}_\theta)^{-1} P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T \left(P_\theta \ddot{\ell}_\theta^T\right)^{-1}. \tag{5.36}$$

Under regularity conditions, this reduces to the inverse of the Fisher information matrix

$$I_\theta = P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T.$$

To see this in the case of a one-dimensional parameter, differentiate the identity $\int p_\theta \, d\mu \equiv 1$ twice with respect to θ. Assuming that the order of differentiation and integration can be reversed, we obtain $\int \dot{p}_\theta \, d\mu \equiv \int \ddot{p}_\theta \, d\mu \equiv 0$. Together with the identities

$$\dot{\ell}_\theta = \frac{\dot{p}_\theta}{p_\theta}; \qquad \ddot{\ell}_\theta = \frac{\ddot{p}_\theta}{p_\theta} - \left(\frac{\dot{p}_\theta}{p_\theta}\right)^2,$$

this implies that $P_\theta \dot{\ell}_\theta = 0$ (scores have mean zero), and $P_\theta \ddot{\ell}_\theta = -I_\theta$ (the curvature of the likelihood is equal to minus the Fisher information). Consequently, (5.36) reduces to I_θ^{-1}. The higher-dimensional case follows in the same way, in which we should interpret the identities $P_\theta \dot{\ell}_\theta = 0$ and $P_\theta \ddot{\ell}_\theta = -I_\theta$ as a vector and a matrix identity, respectively.

We conclude that maximum likelihood estimators typically satisfy

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{\theta}{\rightsquigarrow} N\left(0, I_\theta^{-1}\right).$$

This is a very important result, as it implies that maximum likelihood estimators are asymptotically optimal. The convergence in distribution means roughly that the maximum likelihood estimator $\hat{\theta}_n$ is $N\left(\theta, (nI_\theta)^{-1}\right)$-distributed for every θ, for large n. Hence, it is asymptotically unbiased and asymptotically of variance $(nI_\theta)^{-1}$. According to the Cramér-Rao

theorem, the variance of an unbiased estimator is at least $(nI_\theta)^{-1}$. Thus, we could infer that the maximum likelihood estimator is asymptotically uniformly minimum-variance unbiased, and in this sense optimal. We write "could" because the preceding reasoning is informal and unsatisfying. The asymptotic normality does not warrant any conclusion about the convergence of the moments $\mathrm{E}_\theta \hat{\theta}_n$ and $\mathrm{var}_\theta\, \hat{\theta}_n$; we have not introduced an asymptotic version of the Cramér-Rao theorem; and the Cramér-Rao bound does not make any assertion concerning asymptotic normality. Moreover, the unbiasedness required by the Cramér-Rao theorem is restrictive and can be relaxed considerably in the asymptotic situation.

However, the message that maximum likelihood estimators are *asymptotically efficient* is correct. We give a precise discussion in Chapter 8. The justification through asymptotics appears to be the only general justification of the method of maximum likelihood. In some form, this result was found by Fisher in the 1920s, but a better and more general insight was only obtained in the period from 1950 through 1970 through the work of Le Cam and others.

In the preceding informal derivations and discussion, it is implicitly understood that the density p_θ possesses at least two derivatives with respect to the parameter. Although this can be relaxed considerably, a certain amount of smoothness of the dependence $\theta \mapsto p_\theta$ is essential for the asymptotic normality. Compare the behavior of the maximum likelihood estimators in the case of uniformly distributed observations: They are neither asymptotically normal nor asymptotically optimal.

5.37 ***Example (Uniform distribution).*** Let $X_1, \ldots, X_n$ be a sample from the uniform distribution on $[0, \theta]$. Then the maximum likelihood estimator is the maximum $X_{(n)}$ of the observations. Because the variance of $X_{(n)}$ is of the order $O(n^{-2})$, we expect that a suitable norming rate in this case is not $\sqrt{n}$, but n. Indeed, for each $x < 0$

$$\mathrm{P}_\theta\big(n(X_{(n)} - \theta) \le x\big) = \mathrm{P}_\theta\left(X_1 \le \theta + \frac{x}{n}\right)^n = \left(\frac{\theta + x/n}{\theta}\right)^n \to e^{x/\theta}.$$

Thus, the sequence $-n(X_{(n)} - \theta)$ converges in distribution to an exponential distribution with mean θ. Consequently, the sequence $\sqrt{n}(X_{(n)} - \theta)$ converges to zero in probability.

Note that most of the informal operations in the preceding introduction are illegal or not even defined for the uniform distribution, starting with the definition of the likelihood equations. The informal conclusion that the maximum likelihood estimator is asymptotically optimal is also wrong in this case; see section 9.4. □

We conclude this section with a theorem that establishes the asymptotic normality of maximum likelihood estimators rigorously. Clearly, the asymptotic normality follows from Theorem 5.23 applied to $m_\theta = \log p_\theta$, or from Theorem 5.21 applied with $\psi_\theta = \dot{\ell}_\theta$ equal to the score function of the model. The following result is a minor variation on the first theorem. Its conditions somehow also ensure the relationship $P_\theta \ddot{\ell}_\theta = -I_\theta$ and the twice-differentiability of the map $\theta \mapsto P_{\theta_0} \log p_\theta$, even though the existence of second derivatives is not part of the assumptions. This remarkable phenomenon results from the trivial fact that square roots of probability densities have squares that integrate to 1. To exploit this, we require the differentiability of the maps $\theta \mapsto \sqrt{p_\theta}$, rather than of the maps $\theta \mapsto \log p_\theta$. A statistical model $(P_\theta : \theta \in \Theta)$ is called *differentiable in quadratic mean* if there exists a

measurable vector-valued function $\dot{\ell}_{\theta_0}$ such that, as $\theta \to \theta_0$,

$$\int \left[\sqrt{p_\theta} - \sqrt{p_{\theta_0}} - \frac{1}{2}(\theta - \theta_0)^T \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}} \right]^2 d\mu = o\big(\|\theta - \theta_0\|^2\big), \tag{5.38}$$

This property also plays an important role in asymptotic optimality theory. A discussion, including simple conditions for its validity, is given in Chapter 7. It should be noted that

$$\frac{\partial}{\partial\theta}\sqrt{p_\theta} = \frac{1}{2\sqrt{p_\theta}}\frac{\partial}{\partial\theta}p_\theta = \frac{1}{2}\left(\frac{\partial}{\partial\theta}\log p_\theta\right)\sqrt{p_\theta}.$$

Thus, the function $\dot{\ell}_{\theta_0}$ in the integral really is the score function of the model (as the notation suggests), and the expression $I_{\theta_0} = P_{\theta_0}\dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}^T$ defines the Fisher information matrix. However, condition (5.38) does not require existence of $\partial/\partial\theta\, p_\theta(x)$ for every x.

5.39 *Theorem.* *Suppose that the model $(P_\theta : \theta \in \Theta)$ is differentiable in quadratic mean at an inner point θ_0 of $\Theta \subset \mathbb{R}^k$. Furthermore, suppose that there exists a measurable function $\dot{\ell}$ with $P_{\theta_0}\dot{\ell}^2 < \infty$ such that, for every θ_1 and θ_2 in a neighborhood of θ_0,*

$$\big|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)\big| \le \dot{\ell}(x)\,\|\theta_1 - \theta_2\|.$$

If the Fisher information matrix I_{θ_0} is nonsingular and $\hat{\theta}_n$ is consistent, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = I_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) + o_{P_{\theta_0}}(1).$$

In particular, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix $I_{\theta_0}^{-1}$.

Proof. This theorem is a corollary of Theorem 5.23. We shall show that the conditions of the latter theorem are satisfied for $m_\theta = \log p_\theta$ and $V_{\theta_0} = -I_{\theta_0}$.

Fix an arbitrary converging sequence of vectors $h_n \to h$, and set

$$W_n = 2\left(\sqrt{\frac{p_{\theta_0 + h_n/\sqrt{n}}}{p_{\theta_0}}} - 1\right).$$

By the differentiability in quadratic mean, the sequence $\sqrt{n}W_n$ converges in $L_2(P_{\theta_0})$ to the function $h^T\dot{\ell}_{\theta_0}$. In particular, it converges in probability, whence by a delta method

$$\sqrt{n}\big(\log p_{\theta_0 + h_n/\sqrt{n}} - \log p_{\theta_0}\big) = 2\sqrt{n}\log\big(1 + \tfrac{1}{2}W_n\big) \xrightarrow{P} h^T\dot{\ell}_{\theta_0}.$$

In view of the Lipschitz condition on the map $\theta \mapsto \log p_\theta$, we can apply the dominated-convergence theorem to strengthen this to convergence in $L_2(P_{\theta_0})$. This shows that the map $\theta \mapsto \log p_\theta$ is differentiable in probability, as required in Theorem 5.23. (The preceding argument considers only sequences θ_n of the special form $\theta_0 + h_n/\sqrt{n}$ approaching θ_0. Because h_n can be any converging sequence and $\sqrt{n+1}/\sqrt{n} \to 1$, these sequences are actually not so special. By re-indexing the result can be seen to be true for any $\theta_n \to \theta_0$.)

Next, by computing means (which are zero) and variances, we see that

$$\mathbb{G}_n\left[\sqrt{n}\big(\log p_{\theta_0 + h_n/\sqrt{n}} - \log p_{\theta_0}\big) - h^T\dot{\ell}_{\theta_0}\right] \xrightarrow{P} 0.$$

Equating this result to the expansion given by Theorem 7.2, we see that

$$nP_{\theta_0}\left(\log p_{\theta_0+h_n/\sqrt{n}} - \log p_{\theta_0}\right) \to -\tfrac{1}{2}h^T I_{\theta_0} h.$$

Hence the map $\theta \mapsto P_{\theta_0} \log p_\theta$ is twice-differentiable with second derivative matrix $-I_{\theta_0}$, or at least permits the corresponding Taylor expansion of order 2. ■

5.40 ***Example (Binary regression).*** Suppose that we observe a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ consisting of k-dimensional vectors of "covariates" X_i, and 0-1 "response variables" Y_i, following the model

$$P_\theta(Y_i = 1 \mid X_i = x) = \Psi(\theta^T x).$$

Here $\Psi: \mathbb{R} \mapsto [0, 1]$ is a known continuously differentiable, monotone function. The choices $\Psi(\theta) = 1/(1+e^{-\theta})$ (the logistic distribution function) and $\Psi = \Phi$ (the normal distribution function) correspond to the *logit model* and *probit model*, respectively. The maximum likelihood estimator $\hat{\theta}_n$ maximizes the (conditional) likelihood function

$$\theta \mapsto \prod_{i=1}^n p_\theta(Y_i \mid X_i) := \prod_{i=1}^n \Psi(\theta^T X_i)^{Y_i}\left(1 - \Psi(\theta^T X_i)\right)^{1-Y_i}.$$

The consistency and asymptotic normality of $\hat{\theta}_n$ can be proved, for instance, by combining Theorems 5.7 and 5.39. (Alternatively, we may follow the classical approach given in section 5.6. The latter is particularly attractive for the logit model, for which the log likelihood is strictly concave in θ, so that the point of maximum is unique.) For identifiability of θ we must assume that the distribution of the X_i is not concentrated on a $(k-1)$-dimensional affine subspace of $\mathbb{R}^k$. For simplicity we assume that the range of X_i is bounded.

The consistency can be proved by applying Theorem 5.7 with $m_\theta = \log(p_\theta + p_{\theta_0})/2$. Because p_{θ_0} is bounded away from 0 (and ∞), the function m_θ is somewhat better behaved than the function $\log p_\theta$.

By Lemma 5.35, the parameter θ is identifiable from the density p_θ. We can redo the proof to see that, with $\lesssim$ meaning "less than up to a constant,"

$$\begin{aligned} P_{\theta_0}(m_\theta - m_{\theta_0}) &\lesssim -\int \left(\left(\frac{1}{2}(p_\theta + p_{\theta_0})\right)^{1/2} - p_{\theta_0}^{1/2}\right)^2 d\mu \\ &\lesssim -\mathrm{E}\left(\Psi(\theta^T X) - \Psi\left(\theta_0^T X\right)\right)^2. \end{aligned}$$

This shows that θ_0 is the unique point of maximum of $\theta \mapsto P_{\theta_0} m_\theta$. Furthermore, if $P_{\theta_0} m_{\theta_k} \to P_{\theta_0} m_{\theta_0}$, then $\theta_k^T X \xrightarrow{\mathrm{P}} \theta_0^T X$. If the sequence θ_k is also bounded, then $\mathrm{E}\left((\theta_k - \theta_0)^T X\right)^2 \to 0$, whence $\theta_k \mapsto \theta_0$ by the nonsingularity of the matrix $\mathrm{E}XX^T$. On the other hand, $\|\theta_k\|$ cannot have a diverging subsequence, because in that case $\theta_k^T/\|\theta_k\| X \xrightarrow{\mathrm{P}} 0$ and hence $\theta_k/\|\theta_k\| \to 0$ by the same argument. This verifies condition (5.8).

Checking the uniform convergence to zero of $\sup_\theta |\mathbb{P}_n m_\theta - P m_\theta|$ is not trivial, but it becomes an easy exercise if we employ the Glivenki-Cantelli theorem, as discussed in Chapter 19. The functions $x \mapsto \Psi(\theta^T x)$ form a VC-class, and the functions m_θ take the form $m_\theta(x, y) = \phi\left(\Psi(\theta^T x), y, \Psi(\theta_0^T x)\right)$, where the function $\phi(\gamma, y, \eta)$ is Lipschitz in its first argument with Lipschitz constant bounded above by $1/\eta + 1/(1-\eta)$. This is enough to

ensure that the functions m_θ form a Donsker class and hence certainly a Glivenko-Cantelli class, in view of Example 19.20.

The asymptotic normality of $\sqrt{n}(\hat{\theta}_n - \theta)$ is now a consequence of Theorem 5.39. The score function

$$\dot{\ell}_\theta(y \mid x) = \frac{y - \Psi(\theta^T x)}{\Psi(\theta^T x)(1 - \Psi)(\theta^T x)} \Psi'(\theta^T x) x$$

is uniformly bounded in x, y and θ ranging over compacta, and continuous in θ for every x and y. The Fisher information matrix

$$I_\theta = \mathrm{E} \frac{\Psi'(\theta^T X)^2}{\Psi(\theta^T X)(1 - \Psi)(\theta^T X)} X X^T$$

is continuous in θ, and is bounded below by a multiple of $\mathrm{E} X X^T$ and hence is nonsingular. The differentiability in quadratic mean follows by calculus, or by Lemma 7.6. □

*5.6 Classical Conditions

In this section we discuss the "classical conditions" for asymptotic normality of M-estimators. These conditions were formulated in the 1930s and 1940s to make the informal derivations of the asymptotic normality of maximum likelihood estimators, for instance by Fisher, mathematically rigorous. Although Theorem 5.23 requires less than a first derivative of the criterion function, the "classical conditions" require existence of third derivatives. It is clear that the classical conditions are too stringent, but they are still of interest, because they are simple, lead to simple proofs, and nevertheless apply to many examples. The classical conditions also ensure existence of Z-estimators and have a little to say about their consistency.

We describe the classical approach for general Z-estimators and vector-valued parameters. The higher-dimensional case requires more skill in calculus and matrix algebra than is necessary for the one-dimensional case. When simplified to dimension one the arguments do not go much beyond making the informal derivation leading from (5.18) to (5.19) rigorous.

Let the observations $X_1, \ldots, X_n$ be a sample from a distribution P, and consider the estimating equations

$$\Psi_n(\theta) = \frac{1}{n} \sum_{i=1}^n \psi_\theta(X_i) = \mathbb{P}_n \psi_\theta, \qquad \Psi(\theta) = P \psi_\theta.$$

The estimator $\hat{\theta}_n$ is a zero of Ψ_n, and the true value θ_0 a zero of Ψ. The essential condition of the following theorem is that the second-order partial derivatives of $\psi_\theta(x)$ with respect to θ exist for every x and satisfy

$$\left| \frac{\partial^2 \psi_{\theta,h}(x)}{\partial \theta_i \theta_j} \right| \le \ddot{\psi}(x),$$

for some integrable measurable function $\ddot{\psi}$. This should be true at least for every θ in a neighborhood of θ_0.

5.41 Theorem. *For each θ in an open subset of Euclidean space, let $\theta \mapsto \psi_\theta(x)$ be twice continuously differentiable for every x. Suppose that $P\psi_{\theta_0} = 0$, that $P\|\psi_{\theta_0}\|^2 < \infty$ and that the matrix $P\dot{\psi}_{\theta_0}$ exists and is nonsingular. Assume that the second-order partial derivatives are dominated by a fixed integrable function $\ddot{\psi}(x)$ for every θ in a neighborhood of θ_0. Then every consistent estimator sequence $\hat{\theta}_n$ such that $\Psi_n(\hat{\theta}_n) = 0$ for every n satisfies*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\big(P\dot{\psi}_{\theta_0}\big)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\theta_0}(X_i) + o_P(1).$$

In particular, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix $(P\dot{\psi}_{\theta_0})^{-1}P\psi_{\theta_0}\psi_{\theta_0}^T(P\dot{\psi}_{\theta_0})^{-1}$.

Proof. By Taylor's theorem there exists a (random) vector $\tilde{\theta}_n$ on the line segment between θ_0 and $\hat{\theta}_n$ such that

$$0 = \Psi_n(\hat{\theta}_n) = \Psi_n(\theta_0) + \dot{\Psi}_n(\theta_0)(\hat{\theta}_n - \theta_0) + \tfrac{1}{2}(\hat{\theta}_n - \theta_0)^T\ddot{\Psi}_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0).$$

The first term on the right $\Psi_n(\theta_0)$ is an average of the i.i.d. random vectors $\psi_{\theta_0}(X_i)$, which have mean $P\psi_{\theta_0} = 0$. By the central limit theorem, the sequence $\sqrt{n}\Psi_n(\theta_0)$ converges in distribution to a multivariate normal distribution with mean 0 and covariance matrix $P\psi_{\theta_0}\psi_{\theta_0}^T$. The derivative $\dot{\Psi}_n(\theta_0)$ in the second term is an average also. By the law of large numbers it converges in probability to the matrix $V = P\dot{\psi}_{\theta_0}$. The second derivative $\ddot{\Psi}_n(\tilde{\theta}_n)$ is a k-vector of $(k \times k)$ matrices depending on the second-order derivatives $\ddot{\psi}_\theta$. By assumption, there exists a ball B around θ_0 such that $\ddot{\psi}_\theta$ is dominated by $\|\ddot{\psi}\|$ for every $\theta \in B$. The probability of the event $\{\hat{\theta}_n \in B\}$ tends to 1. On this event

$$\big\|\ddot{\Psi}_n(\tilde{\theta}_n)\big\| = \left\|\frac{1}{n}\sum_{i=1}^{n}\ddot{\psi}_{\tilde{\theta}_n}(X_i)\right\| \le \frac{1}{n}\sum_{i=1}^{n}\big\|\ddot{\psi}(X_i)\big\|.$$

This is bounded in probability by the law of large numbers. Combination of these facts allows us to rewrite the preceding display as

$$-\Psi_n(\theta_0) = \big(V + o_P(1) + \tfrac{1}{2}(\hat{\theta}_n - \theta_0)\,O_P(1)\big)(\hat{\theta}_n - \theta_0) = \big(V + o_P(1)\big)(\hat{\theta}_n - \theta_0),$$

because the sequence $(\hat{\theta}_n - \theta_0)\,O_P(1) = o_P(1)O_P(1)$ converges to 0 in probability if $\hat{\theta}_n$ is consistent for θ_0. The probability that the matrix $V_{\theta_0} + o_P(1)$ is invertible tends to 1. Multiply the preceding equation by $\sqrt{n}$ and apply $\big(V + o_P(1)\big)^{-1}$ left and right to complete the proof. ■

In the preceding sections, the existence and consistency of solutions $\hat{\theta}_n$ of the estimating equations is assumed from the start. The present smoothness conditions actually ensure the existence of solutions. (Again the conditions could be significantly relaxed, as shown in the next proof.) Moreover, provided there exists a consistent estimator sequence at all, it is always possible to select a consistent sequence of solutions.

5.42 Theorem. *Under the conditions of the preceding theorem, the probability that the equation $\mathbb{P}_n\psi_\theta = 0$ has at least one root tends to 1, as $n \to \infty$, and there exists a sequence of roots $\hat{\theta}_n$ such that $\hat{\theta}_n \to \theta_0$ in probability. If $\psi_\theta = \dot{m}_\theta$ is the gradient of some function*

m_θ and θ_0 is a point of local maximum of $\theta \mapsto Pm_\theta$, then the sequence $\hat\theta_n$ can be chosen to be local maxima of the maps $\theta \mapsto \mathbb{P}_n m_\theta$.

Proof. Integrate the Taylor expansion of $\theta \mapsto \psi_\theta(x)$ with respect to x to find that, for a point $\tilde\theta = \tilde\theta(x)$ on the line segment between θ_0 and θ,

$$P\psi_\theta = P\psi_{\theta_0} + P\dot\psi_{\theta_0}(\theta - \theta_0) + \tfrac{1}{2}(\theta - \theta_0)^T P\ddot\psi_{\tilde\theta}(\theta - \theta_0).$$

By the domination condition, $\|P\ddot\psi_{\tilde\theta}\|$ is bounded by $P\|\ddot\psi\| < \infty$ if θ is sufficiently close to θ_0. Thus, the map $\Psi(\theta) = P\psi_\theta$ is differentiable at θ_0. By the same argument Ψ is differentiable throughout a small neighborhood of θ_0, and by a similar expansion (but now to first order) the derivative $P\dot\psi_\theta$ can be seen to be continuous throughout this neighborhood. Because $P\dot\psi_{\theta_0}$ is nonsingular by assumption, we can make the neighborhood still smaller, if necessary, to ensure that the derivative of Ψ is nonsingular throughout the neighborhood. Then, by the inverse function theorem, there exists, for every sufficiently small $\delta > 0$, an open neighborhood G_δ of θ_0 such that the map $\Psi\colon \overline{G}_\delta \mapsto \overline{\text{ball}}(0, \delta)$ is a homeomorphism. The diameter of $\overline{G}_\delta$ is bounded by a multiple of δ, by the mean-value theorem and the fact that the norms of the derivatives $(P\dot\psi_\theta)^{-1}$ of the inverse Ψ^{-1} are bounded.

Combining the preceding Taylor expansion with a similar expansion for the sample version $\Psi_n(\theta) = \mathbb{P}_n\psi_\theta$, we see

$$\sup_{\theta\in\overline{G}_\delta} \big\|\Psi_n(\theta) - \Psi(\theta)\big\| \le o_P(1) + \delta o_P(1) + \delta^2 O_P(1),$$

where the $o_P(1)$ terms and the $O_P(1)$ term result from the law of large numbers, and are uniform in small δ. Because $\mathrm{P}\big(o_P(1) + \delta o_P(1) > \frac{1}{2}\delta\big) \to 0$ for every $\delta > 0$, there exists $\delta_n \downarrow 0$ such that $\mathrm{P}\big(o_P(1) + \delta_n o_P(1) > \frac{1}{2}\delta_n\big) \to 0$. If $K_{n,\delta}$ is the event where the left side of the preceding display is bounded above by δ, then $\mathrm{P}(K_{n,\delta_n}) \to 1$ as $n \to \infty$.

On the event $K_{n,\delta}$ the map $\theta \mapsto \theta - \Psi_n \circ \Psi^{-1}(\theta)$ maps $\overline{\text{ball}}(0, \delta)$ into itself, by the definitions of $\overline{G}_\delta$ and $K_{n,\delta}$. Because the map is also continuous, it possesses a fixed-point in $\overline{\text{ball}}(0, \delta)$, by Brouwer's fixed point theorem. This yields a zero of Ψ_n in the set $\overline{G}_\delta$, whence the first assertion of the theorem.

For the final assertion, first note that the Hessian $P\dot\psi_{\theta_0}$ of $\theta \mapsto Pm_\theta$ at θ_0 is negative-definite, by assumption. A Taylor expansion as in the proof of Theorem 5.41 shows that $\mathbb{P}_n\dot\psi_{\hat\theta_n} - \mathbb{P}_n\dot\psi_{\theta_0} \xrightarrow{\mathrm{P}} 0$ for every $\hat\theta_n \xrightarrow{\mathrm{P}} \theta_0$. Hence the Hessian $\mathbb{P}_n\dot\psi_{\hat\theta_n}$ of $\theta \mapsto \mathbb{P}_n m_\theta$ at any consistent zero $\hat\theta_n$ converges in probability to the negative-definite matrix $P\dot\psi_{\theta_0}$ and is negative-definite with probability tending to 1. ■

The assertion of the theorem that there exists a consistent sequence of roots of the estimating equations is easily misunderstood. It does not guarantee the existence of an asymptotically consistent sequence of estimators. The only claim is that a clairvoyant statistician (with preknowledge of θ_0) can choose a consistent sequence of roots. In reality, it may be impossible to choose the right solutions based only on the data (and knowledge of the model). In this sense the preceding theorem, a standard result in the literature, looks better than it is.

The situation is not as bad as it seems. One interesting situation is if the solution of the estimating equation is unique for every n. Then our solutions must be the same as those of the clairvoyant statistician and hence the sequence of solutions is consistent.

In general, the deficit can be repaired with the help of a preliminary sequence of estimators $\tilde{\theta}_n$. If the sequence $\tilde{\theta}_n$ is consistent, then it works to choose the root $\hat{\theta}_n$ of $\mathbb{P}_n\psi_\theta = 0$ that is closest to $\tilde{\theta}_n$. Because $\|\hat{\theta}_n - \tilde{\theta}_n\|$ is smaller than the distance $\|\hat{\theta}_n^c - \tilde{\theta}_n\|$ between the clairvoyant sequence $\hat{\theta}_n^c$ and $\tilde{\theta}_n$, both distances converge to zero in probability. Thus the sequence of closest roots is consistent.

The assertion of the theorem can also be used in a negative direction. The point θ_0 in the theorem is required to be a zero of $\theta \mapsto P\psi_\theta$, but, apart from that, it may be arbitrary. Thus, the theorem implies at the same time that a malicious statistician can always choose a sequence of roots $\hat{\theta}_n$ that converges to any given zero. These may include other points besides the "true" value of θ. Furthermore, inspection of the proof shows that the sequence of roots can also be chosen to jump back and forth between two (or more) zeros. If the function $\theta \mapsto P\psi_\theta$ has multiple roots, we must exercise care. We can be sure that certain roots of $\theta \mapsto \mathbb{P}_n\psi_\theta$ are bad estimators.

Part of the problem here is caused by using estimating equations, rather than maximization to find estimators, which blurs the distinction between points of absolute maximum, local maximum, and even minimum. In the light of the results on consistency in section 5.2, we may expect the location of the point of absolute maximum of $\theta \mapsto \mathbb{P}_n m_\theta$ to converge to a point of absolute maximum of $\theta \mapsto Pm_\theta$. As long as this is unique, the absolute maximizers of the criterion function are typically consistent.

5.43 ***Example (Weibull distribution).*** Let $X_1, \ldots, X_n$ be a sample from the Weibull distribution with density

$$p_{\theta,\sigma}(x) = \frac{\theta}{\sigma} x^{\theta-1} e^{-x^\theta/\sigma}, \qquad x > 0,\ \theta > 0,\ \sigma > 0.$$

(Then $\sigma^{1/\theta}$ is a scale parameter.) The score function is given by the partial derivatives of the log density with respect to θ and σ:

$$\dot{\ell}_{\theta,\sigma}(x) = \left(\frac{1}{\theta} + \log x - \frac{x^\theta}{\sigma}\log x,\ w - \frac{1}{\sigma} + \frac{x^\theta}{\sigma^2}\right).$$

The likelihood equations $\sum \dot{\ell}_{\theta,\sigma}(x_i) = 0$ reduce to

$$\sigma = \frac{1}{n}\sum_{i=1}^n x_i^\theta; \qquad \frac{1}{\theta} + \frac{1}{n}\sum_{i=1}^n \log x_i - \frac{\sum_{i=1}^n x_i^\theta \log x_i}{\sum_{i=1}^n x_i^\theta} = 0.$$

The second equation is strictly decreasing in θ, from ∞ at $\theta = 0$ to $\overline{\log x} - \log x_{(n)}$ at $\theta = \infty$. Hence a solution exists, and is unique, unless all x_i are equal. Provided the higher-order derivatives of the score function exist and can be dominated, the sequence of maximum likelihood estimators $(\hat{\theta}_n, \hat{\sigma}_n)$ is asymptotically normal by Theorems 5.41 and 5.42. There exist four different third-order derivatives, given by

$$\begin{aligned}
\frac{\partial^3 \ell_{\theta,\sigma}(x)}{\partial\theta^3} &= \frac{2}{\theta^3} - \frac{x^\theta}{\sigma}\log^3 x \\
\frac{\partial^3 \ell_{\theta,\sigma}(x)}{\partial\theta^2\partial\sigma} &= \frac{x^\theta}{\sigma^2}\log^2 x \\
\frac{\partial^3 \ell_{\theta,\sigma}(x)}{\partial\theta\partial\sigma^2} &= -\frac{2x^\theta}{\sigma^3}\log x \\
\frac{\partial^3 \ell_{\theta,\sigma}(x)}{\partial\sigma^3} &= -\frac{2}{\sigma^3} + \frac{6x^\theta}{\sigma^4}.
\end{aligned}$$

For θ and σ ranging over sufficiently small neighborhoods of θ_0 and σ_0, these functions are dominated by a function of the form

$$M(x) = A(1 + x^B)\big(1 + |\log x| + \cdots + |\log x|^3\big),$$

for sufficiently large A and B. Because the Weibull distribution has an exponentially small tail, the mixed moment $\mathrm{E}_{\theta_0,\sigma_0} X^p |\log X|^q$ is finite for every $p, q \geq 0$. Thus, all moments of $\dot{\ell}_\theta$ and $\ddot{\ell}_\theta$ exist and M is integrable. □

*5.7 One-Step Estimators

The method of Z-estimation as discussed so far has two disadvantages. First, it may be hard to find the roots of the estimating equations. Second, for the roots to be consistent, the estimating equation needs to behave well throughout the parameter set. For instance, existence of a second root close to the boundary of the parameter set may cause trouble. The *one-step method* overcomes these problems by building on and improving a preliminary estimator $\tilde{\theta}_n$.

The idea is to solve the estimator from a linear approximation to the original estimating equation $\Psi_n(\theta) = 0$. Given a preliminary estimator $\tilde{\theta}_n$, the one-step estimator is the solution (in θ) to

$$\Psi_n(\tilde{\theta}_n) + \dot{\Psi}_n(\tilde{\theta}_n)(\theta - \tilde{\theta}_n) = 0.$$

This corresponds to replacing $\Psi_n(\theta)$ by its tangent at $\tilde{\theta}_n$, and is known as the method of *Newton-Rhapson* in numerical analysis. The solution $\theta = \hat{\theta}_n$ is

$$\hat{\theta}_n = \tilde{\theta}_n - \dot{\Psi}_n(\tilde{\theta}_n)^{-1}\, \Psi_n(\tilde{\theta}_n).$$

In numerical analysis this procedure is iterated a number of times, taking $\hat{\theta}_n$ as the new preliminary guess, and so on. Provided that the starting point $\tilde{\theta}_n$ is well chosen, the sequence of solutions converges to a root of Ψ_n. Our interest here goes in a different direction. We suppose that the preliminary estimator $\tilde{\theta}_n$ is already within range $n^{-1/2}$ of the true value of θ. Then, as we shall see, just one iteration of the Newton-Rhapson scheme produces an estimator $\hat{\theta}_n$ that is as good as the Z-estimator defined by Ψ_n. In fact, it is better in that its consistency is guaranteed, whereas the true Z-estimator may be inconsistent or not uniquely defined.

In this way consistency and asymptotic normality are effectively separated, which is useful because these two aims require different properties of the estimating equations. Good initial estimators can be constructed by ad-hoc methods and take care of consistency. Next, these initial estimators can be improved by the one-step method. Thus, for instance, the good properties of maximum likelihood estimation can be retained, even in cases in which the consistency fails.

In this section we impose the following condition on the random criterion functions Ψ_n. For every constant M and a given nonsingular matrix $\dot{\Psi}_0$,

$$\sup_{\sqrt{n}\|\theta-\theta_0\|<M} \Big\| \sqrt{n}\big(\Psi_n(\theta) - \Psi_n(\theta_0)\big) - \dot{\Psi}_0 \sqrt{n}(\theta - \theta_0) \Big\| \xrightarrow{\mathrm{P}} 0. \tag{5.44}$$

Condition (5.44) suggests that Ψ_n is differentiable at θ_0, with derivative tending to $\dot{\Psi}_0$, but this is not an assumption. We do not require that a derivative $\dot{\Psi}_n$ exists, and introduce

a further refinement of the Newton-Rhapson scheme by replacing $\dot{\Psi}_n(\tilde{\theta}_n)$ by arbitrary estimators. Given nonsingular, random matrices $\dot{\Psi}_{n,0}$ that converge in probability to $\dot{\Psi}_0$ define the *one-step estimator*

$$\hat{\theta}_n = \tilde{\theta}_n - \dot{\Psi}_{n,0}^{-1}\Psi_n(\tilde{\theta}_n).$$

Call an estimator sequence $\tilde{\theta}_n$ *$\sqrt{n}$-consistent* if the sequence $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ is uniformly tight. The interpretation is that $\tilde{\theta}_n$ already determines the value θ_0 within $n^{-1/2}$-range.

5.45 ***Theorem (One-step estimation).*** *Let $\sqrt{n}\Psi_n(\theta_0) \rightsquigarrow Z$ and let (5.44) hold. Then the one-step estimator $\hat{\theta}_n$, for a given $\sqrt{n}$-consistent estimator sequence $\tilde{\theta}_n$ and estimators $\dot{\Psi}_{n,0} \xrightarrow{P} \dot{\Psi}_0$, satisfies*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\dot{\Psi}_0^{-1}\sqrt{n}\Psi_n(\theta_0) + o_P(1).$$

5.46 ***Addendum.*** *For $\Psi_n(\theta) = \mathbb{P}_n\psi_\theta$ condition (5.44) is satisfied under the conditions of Theorem 5.21 with $\dot{\Psi}_0 = V_{\theta_0}$, and under the conditions of Theorem 5.41 with $\dot{\Psi}_0 = P\dot{\psi}_{\theta_0}$.*

Proof. The standardized estimator $\dot{\Psi}_{n,0}\sqrt{n}(\hat{\theta}_n - \theta_0)$ equals

$$\dot{\Psi}_{n,0}\sqrt{n}(\tilde{\theta}_n - \theta_0) - \sqrt{n}\big(\Psi_n(\tilde{\theta}_n) - \Psi_n(\theta_0)\big) - \dot{\Psi}_{n,0}^{-1}\sqrt{n}\Psi_n(\theta_0).$$

By (5.44) the second term can be replaced by $-\dot{\Psi}_0\sqrt{n}(\tilde{\theta}_n - \theta_0) + o_P(1)$. Thus the expression can be rewritten as

$$(\dot{\Psi}_{n,0} - \dot{\Psi}_0)\sqrt{n}(\tilde{\theta}_n - \theta_0) - \sqrt{n}\Psi_n(\theta_0) + o_P(1).$$

The first term converges to zero in probability, and the theorem follows after application of Slutsky's lemma.

For a proof of the addendum, see the proofs of the corresponding theorems. ■

If the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution, then it is certainly uniformly tight. Consequently, a sequence of one-step estimators is $\sqrt{n}$-consistent and can itself be used as preliminary estimator for a second iteration of the modified Newton-Rhapson algorithm. Presumably, this would give a value closer to a root of Ψ_n. However, the limit distribution of this "two-step estimator" is the same, so that repeated iteration does not give asymptotic improvement. In practice a multistep method may nevertheless give better results.

We close this section with a discussion of the *discretization* trick. This device is mostly of theoretical value and has been introduced to relax condition (5.44) to the following. For every *nonrandom* sequence $\theta_n = \theta_0 + O(n^{-1/2})$,

$$\left\|\sqrt{n}\big(\Psi_n(\theta_n) - \Psi_n(\theta_0)\big) - \dot{\Psi}_0\sqrt{n}(\theta_n - \theta_0)\right\| \xrightarrow{P} 0. \tag{5.47}$$

This new condition is less stringent and much easier to check. It is sufficiently strong if the preliminary estimators $\tilde{\theta}_n$ are *discretized* on grids of mesh width $n^{-1/2}$. For instance, $\tilde{\theta}_n$ is suitably discretized if all its realizations are points of the grid $n^{-1/2}\mathbb{Z}^k$ (consisting of the points $n^{-1/2}(i_1, \ldots, i_k)$ for integers $i_1, \ldots, i_k$). This is easy to achieve, but perhaps unnatural. Any preliminary estimator sequence $\tilde{\theta}_n$ can be discretized by replacing its values

by the closest points of the grid. Because this changes each coordinate by at most $n^{-1/2}$, $\sqrt{n}$-consistency of $\tilde{\theta}_n$ is retained by discretization.

Define a one-step estimator $\hat{\theta}_n$ as before, but now use a discretized version of the preliminary estimator.

5.48 ***Theorem (Discretized one-step estimation).*** *Let $\sqrt{n}\Psi_n(\theta_0) \rightsquigarrow Z$ and let (5.47) hold. Then the one-step estimator $\hat{\theta}_n$, for a given $\sqrt{n}$-consistent, discretized estimator sequence $\tilde{\theta}_n$ and estimators $\dot{\Psi}_{n,0} \xrightarrow{\mathrm{P}} \dot{\Psi}_0$, satisfies*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\dot{\Psi}_0^{-1}\sqrt{n}\Psi_n(\theta_0) + o_P(1).$$

5.49 ***Addendum.*** *For $\Psi_n(\theta) = \mathbb{P}_n\psi_\theta$ and $\mathbb{P}_n$ the empirical measure of a random sample from a density p_θ that is differentiable in quadratic mean (5.38), condition (5.47), is satisfied, with $\dot{\Psi}_0 = -P_{\theta_0}\psi_{\theta_0}\dot{\ell}_{\theta_0}^T$, if, as $\theta \to \theta_0$,*

$$\int \left[\psi_\theta\sqrt{p_\theta} - \psi_{\theta_0}\sqrt{p_{\theta_0}}\right]^2 d\mu \to 0.$$

Proof. The arguments of the previous proof apply, except that it must be shown that

$$R(\tilde{\theta}_n) := \sqrt{n}\big(\Psi_n(\tilde{\theta}_n) - \Psi_n(\theta_0)\big) - \dot{\Psi}_0^{-1}\sqrt{n}(\tilde{\theta}_n - \theta_0)$$

converges to zero in probability. Fix $\varepsilon > 0$. By the $\sqrt{n}$-consistency, there exists M with $\mathrm{P}\big(\sqrt{n}\|\tilde{\theta}_n - \theta_0\| > M\big) < \varepsilon$. If $\sqrt{n}\|\tilde{\theta}_n - \theta_0\| \le M$, then $\tilde{\theta}_n$ equals one of the values in the set $S_n = \big\{\theta \in n^{-1/2}\mathbb{Z}^k : \|\theta - \theta_0\| \le n^{-1/2}M\big\}$. For each M and n there are only finitely many elements in this set. Moreover, for fixed M the number of elements is bounded independently of n. Thus

$$\begin{aligned}\mathrm{P}\big(\|R(\tilde{\theta}_n)\| > \varepsilon\big) &\le \varepsilon + \sum_{\theta_n \in S_n} \mathrm{P}\big(\|R(\theta_n)\| > \varepsilon \wedge \tilde{\theta}_n = \theta_n\big) \\ &\le \varepsilon + \sum_{\theta_n \in S_n} \mathrm{P}\big(\|R(\theta_n)\| > \varepsilon\big).\end{aligned}$$

The maximum of the terms in the sum corresponds to a sequence of nonrandom vectors θ_n with $\theta_n = \theta_0 + O(n^{-1/2})$. It converges to zero by (5.47). Because the number of terms in the sum is bounded independently of n, the sum converges to zero.

For a proof of the addendum, see proposition A.10 in [139]. ■

If the score function $\dot{\ell}_\theta$ of the model also satisfies the conditions of the addendum, then the estimators $\dot{\Psi}_{n,0} = -P_{\tilde{\theta}_n}\psi_{\tilde{\theta}_n}\dot{\ell}_{\tilde{\theta}_n}^T$ are consistent for $\dot{\Psi}_0$. This shows that discretized one-step estimation can be carried through under very mild regularity conditions. Note that the addendum requires only continuity of $\theta \mapsto \psi_\theta$, whereas (5.47) appears to require differentiability.

5.50 ***Example (Cauchy distribution).*** Suppose $X_1, \ldots, X_n$ are a sample from the Cauchy location family $p_\theta(x) = \pi^{-1}\big(1 + (x - \theta)^2\big)^{-1}$. Then the score function is given by

$$\dot{\ell}_\theta(x) = \frac{2(x - \theta)}{1 + (x - \theta)^2}.$$

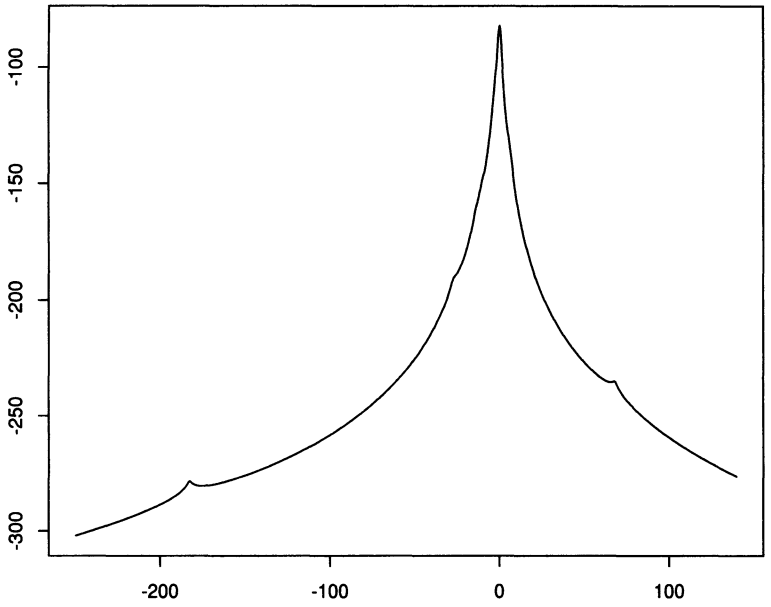

Figure 5.4. Cauchy log likelihood function of a sample of 25 observations, showing three local maxima. The value of the absolute maximum is well-separated from the other maxima, and its location is close to the true value zero of the parameter.

This function behaves like $1/x$ for $x \to \pm\infty$ and is bounded in between. The second moment of $\dot{\ell}_\theta(X_1)$ therefore exists, unlike the moments of the distribution itself. Because the sample mean possesses the same (Cauchy) distribution as a single observation X_1, the sample mean is a very inefficient estimator. Instead we could use the median, or another M-estimator. However, the asymptotically best estimator should be based on maximum likelihood. We have

$$\dot{\ell}_\theta(x) = \frac{4(x-\theta)\big((x-\theta)^2 - 3\big)}{\big(1+(x-\theta)^2\big)^3}.$$

The tails of this function are of the order $1/x^3$, and the function is bounded in between. These bounds are uniform in θ varying over a compact interval. Thus the conditions of Theorems 5.41 and 5.42 are satisfied. Since the consistency follows from Example 5.16, the sequence of maximum likelihood estimators is asymptotically normal.

The Cauchy likelihood estimator has gained a bad reputation, because the likelihood equation $\sum \dot{\ell}_\theta(X_i) = 0$ typically has several roots. The number of roots behaves asymptotically as two times a Poisson$(1/\pi)$ variable plus 1. (See [126].) Therefore, the one-step (or possibly multi-step method) is often recommended, with, for instance, the median as the initial estimator. Perhaps a better solution is not to use the likelihood equations, but to determine the maximum likelihood estimator by, for instance, visual inspection of a graph of the likelihood function, as in Figure 5.4. This is particularly appropriate because the difficulty of multiple roots does not occur in the two parameter location-scale model. In the model with density $p_\theta(x/\sigma)/\sigma$, the maximum likelihood estimator for (θ, σ) is unique. (See [25].) □

5.51 ***Example (Mixtures).*** Let f and g be given, positive probability densities on the real line. Consider estimating the parameter $\theta = (\mu, \nu, \sigma, \tau, p)$ based on a random sample from

the mixture density

$$x \mapsto pf\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma} + (1-p)g\left(\frac{x-\nu}{\tau}\right)\frac{1}{\tau}.$$

If f and g are sufficiently regular, then this is a smooth five-dimensional parametric model, and the standard theory should apply. Unfortunately, the supremum of the likelihood over the natural parameter space is ∞, and there exists no maximum likelihood estimator. This is seen, for instance, from the fact that the likelihood is bigger than

$$pf\left(\frac{x_1-\mu}{\sigma}\right)\frac{1}{\sigma}\prod_{i=2}^{n}(1-p)g\left(\frac{x_i-\nu}{\tau}\right)\frac{1}{\tau}.$$

If we set $\mu = x_1$ and next maximize over $\sigma > 0$, then we obtain the value ∞ whenever $p > 0$, irrespective of the values of ν and τ.

A one-step estimator appears reasonable in this example. In view of the smoothness of the likelihood, the general theory yields the asymptotic efficiency of a one-step estimator if started with an initial $\sqrt{n}$-consistent estimator. Moment estimators could be appropriate initial estimators. □

*5.8 Rates of Convergence

In this section we discuss some results that give the rate of convergence of M-estimators. These results are useful as intermediate steps in deriving a limit distribution, but also of interest on their own. Applications include both classical estimators of "regular" parameters and estimators that converge at a slower than $\sqrt{n}$-rate. The main result is simple enough, but its conditions include a maximal inequality, for which results such as in Chapter 19 are needed.

Let $\mathbb{P}_n$ be the empirical distribution of a random sample of size n from a distribution P, and, for every θ in a metric space Θ, let $x \mapsto m_\theta(x)$ be a measurable function. Let $\hat{\theta}_n$ (nearly) maximize the criterion function $\theta \mapsto \mathbb{P}_n m_\theta$.

The criterion function may be viewed as the sum of the deterministic map $\theta \mapsto Pm_\theta$ and the random fluctations $\theta \mapsto \mathbb{P}_n m_\theta - Pm_\theta$. The rate of convergence of $\hat{\theta}_n$ depends on the combined behavior of these maps. If the deterministic map changes rapidly as θ moves away from the point of maximum and the random fluctuations are small, then $\hat{\theta}_n$ has a high rate of convergence. For convenience of notation we measure the fluctuations in terms of the empirical process $\mathbb{G}_n m_\theta = \sqrt{n}(\mathbb{P}_n m_\theta - Pm_\theta)$.

5.52 ***Theorem (Rate of convergence).*** *Assume that for fixed constants C and $\alpha > \beta$, for every n, and for every sufficiently small $\delta > 0$,*

$$\sup_{d(\theta,\theta_0)<\delta} P(m_\theta - m_{\theta_0}) \le -C\delta^\alpha,$$

$$\mathrm{E}^* \sup_{d(\theta,\theta_0)<\delta} \left|\mathbb{G}_n(m_\theta - m_{\theta_0})\right| \le C\delta^\beta.$$

If the sequence $\hat{\theta}_n$ satisfies $\mathbb{P}_n m_{\hat{\theta}_n} \ge \mathbb{P}_n m_{\theta_0} - O_P\left(n^{\alpha/(2\beta-2\alpha)}\right)$ and converges in outer probability to θ_0, then $n^{1/(2\alpha-2\beta)}d(\hat{\theta}_n, \theta_0) = O_P^*(1)$.

Proof. Set $r_n = n^{1/(2\alpha-2\beta)}$ and suppose that $\hat{\theta}_n$ maximizes the map $\theta \mapsto \mathbb{P}_n m_\theta$ up to a variable $R_n = O_P(r_n^{-\alpha})$.

For each n, the parameter space minus the point θ_0 can be partitioned into the "shells" $S_{j,n} = \{\theta: 2^{j-1} < r_n d(\theta, \theta_0) \le 2^j\}$, with j ranging over the integers. If $r_n\, d(\hat{\theta}_n, \theta_0)$ is larger than 2^M for a given integer M, then $\hat{\theta}_n$ is in one of the shells $S_{j,n}$ with $j \ge M$. In that case the supremum of the map $\theta \mapsto \mathbb{P}_n m_\theta - \mathbb{P}_n m_{\theta_0}$ over this shell is at least $-R_n$ by the property of $\hat{\theta}_n$. Conclude that, for every $\varepsilon > 0$,

$$\mathrm{P}^*\big(r_n\, d(\hat{\theta}_n, \theta_0) > 2^M\big) \le \sum_{\substack{j \ge M \\ 2^j \le \varepsilon r_n}} \mathrm{P}^*\left(\sup_{\theta \in S_{j,n}} (\mathbb{P}_n m_\theta - \mathbb{P}_n m_{\theta_0}) \ge -\frac{K}{r_n^\alpha} \right) + \mathrm{P}^*\big(2d(\hat{\theta}_n, \theta_0) \ge \varepsilon\big) + \mathrm{P}\big(r_n^\alpha R_n \ge K\big).$$

If the sequence $\hat{\theta}_n$ is consistent for θ_0, then the second probability on the right converges to 0 as $n \to \infty$, for every fixed $\varepsilon > 0$. The third probability on the right can be made arbitrarily small by choice of K, uniformly in n. Choose $\varepsilon > 0$ small enough to ensure that the conditions of the theorem hold for every $\delta \le \varepsilon$. Then for every j involved in the sum, we have

$$\sup_{\theta \in S_{j,n}} P(m_\theta - m_{\theta_0}) \le -C \frac{2^{(j-1)\alpha}}{r_n^\alpha}.$$

For $\frac{1}{2}C2^{(M-1)\alpha} \ge K$, the series can be bounded in terms of the empirical process $\mathbb{G}_n$ by

$$\sum_{\substack{j \ge M \\ 2^j \le \varepsilon r_n}} \mathrm{P}^*\left(\big\|\mathbb{G}_n(m_\theta - m_{\theta_0})\big\|_{S_{j,n}} \ge C\sqrt{n} \frac{2^{(j-1)\alpha}}{2r_n^\alpha} \right) \le \sum_{j \ge M} \frac{(2^j/r_n)^\beta\, 2r_n^\alpha}{\sqrt{n} 2^{(j-1)\alpha}},$$

by Markov's inequality and the definition of r_n. The right side converges to zero for every $M = M_n \to \infty$. ■

Consider the special case that the parameter θ is a Euclidean vector. If the map $\theta \mapsto Pm_\theta$ is twice-differentiable at the point of maximum θ_0, then its first derivative at θ_0 vanishes and a Taylor expansion of the limit criterion function takes the form

$$P(m_\theta - m_{\theta_0}) = \tfrac{1}{2}(\theta - \theta_0)^T V (\theta - \theta_0) + o\big(\|\theta - \theta_0\|^2\big).$$

Then the first condition of the theorem holds with $\alpha = 2$ provided that the second-derivative matrix V is nonsingular.

The second condition of the theorem is a *maximal inequality* and is harder to verify. In "regular" cases it is valid with $\beta = 1$ and the theorem yields the "usual" rate of convergence $\sqrt{n}$. The theorem also applies to nonstandard situations and yields, for instance, the rate $n^{1/3}$ if $\alpha = 2$ and $\beta = \frac{1}{2}$. Lemmas 19.34, 19.36 and 19.38 and corollary 19.35 are examples of maximal inequalities that can be appropriate for the present purpose. They give bounds in terms of the entropies of the classes of functions $\{m_\theta - m_{\theta_0}: d(\theta, \theta_0) < \delta\}$.

A Lipschitz condition on the maps $\theta \mapsto m_\theta$ is one possibility to obtain simple estimates on these entropies and is applicable in many applications. The result of the following corollary is used earlier in this chapter.

5.53 *Corollary.* *For each θ in an open subset of Euclidean space let $x \mapsto m_\theta(x)$ be a measurable function such that, for every θ_1 and θ_2 in a neighborhood of θ_0 and a measurable function $\dot{m}$ such that $P\dot{m}^2 < \infty$,*

$$\left|m_{\theta_1}(x) - m_{\theta_2}(x)\right| \le \dot{m}(x)\,\|\theta_1 - \theta_2\|.$$

Furthermore, suppose that the map $\theta \mapsto Pm_\theta$ admits a second-order Taylor expansion at the point of maximum θ_0 with nonsingular second derivative. If $\mathbb{P}_n m_{\hat{\theta}_n} \ge \mathbb{P}_n m_{\theta_0} - O_P(n^{-1})$, then $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_P(1)$, provided that $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Proof. By assumption, the first condition of Theorem 5.52 is valid with $\alpha = 2$. To see that the second one is valid with $\beta = 1$, we apply Corollary 19.35 to the class of functions $\mathcal{F} = \{m_\theta - m_{\theta_0} : \|\theta - \theta_0\| < \delta\}$. This class has envelope function $F = \dot{m}\delta$, whence

$$\mathrm{E}^* \sup_{\|\theta-\theta_0\|<\delta} \left|\mathbb{G}_n(m_\theta - m_{\theta_0})\right| \lesssim \int_0^{\|\dot{m}\|_{P,2}\delta} \sqrt{\log N_{[]}\big(\varepsilon, \mathcal{F}, L_2(P)\big)}\, d\varepsilon.$$

The bracketing entropy of the class $\mathcal{F}$ is estimated in Example 19.7. Inserting the upper bound obtained there into the integral, we obtain that the preceding display is bounded above by a multiple of

$$\int_0^{\|\dot{m}\|_{P,2}\delta} \sqrt{\log\left(\frac{\delta}{\varepsilon}\right)}\, d\varepsilon.$$

Change the variables in the integral to see that this is a multiple of δ. ■

Rates of convergence different from $\sqrt{n}$ are quite common for M-estimators of infinite-dimensional parameters and may also be obtained through the application of Theorem 5.52. See Chapters 24 and 25 for examples. Rates slower than $\sqrt{n}$ may also arise for fairly simple parametric estimates.

5.54 *Example (Modal interval).* Suppose that we define an estimator $\hat{\theta}_n$ of location as the center of an interval of length 2 that contains the largest possible fraction of the observations. This is an M-estimator for the functions $m_\theta = 1_{[\theta-1,\theta+1]}$.

For many underlying distributions the first condition of Theorem 5.52 holds with $\alpha = 2$. It suffices that the map $\theta \mapsto Pm_\theta = P[\theta - 1, \theta + 1]$ is twice-differentiable and has a proper maximum at some point θ_0. Using the maximal inequality Corollary 19.35 (or Lemma 19.38), we can show that the second condition is valid with $\beta = \frac{1}{2}$. Indeed, the bracketing entropy of the intervals in the real line is of the order δ/ε^2, and the envelope function of the class of functions $1_{[\theta-1,\theta+1]} - 1_{[\theta_0-1,\theta_0+1]}$ as θ ranges over $(\theta_0 - \delta, \theta_0 + \delta)$ is bounded by $1_{[\theta_0-1-\delta,\theta_0-1+\delta]} + 1_{[\theta_0+1-\delta,\theta_0+1+\delta]}$, whose squared L_2-norm is bounded by $\|p\|_\infty 2\delta$.

Thus Theorem 5.52 applies with $\alpha = 2$ and $\beta = \frac{1}{2}$ and yields the rate of convergence $n^{1/3}$. The resulting location estimator is very robust against outliers. However, in view of its slow convergence rate, one should have good reasons to use it.

The use of an interval of length 2 is somewhat awkward. Every other fixed length would give the same result. More interestingly, we can also replace the fixed-length interval by the smallest interval that contains a fixed fraction, for instance 1/2, of the observations. This

still yields a rate of convergence of $n^{1/3}$. The intuitive reason for this is that the length of a "shorth" settles down by a $\sqrt{n}$-rate and hence its randomness is asymptotically negligible relative to its center. □

The preceding theorem requires the consistency of $\hat{\theta}_n$ as a condition. This consistency is implied if the other conditions are valid for every $\delta > 0$, not just for small values of δ. This can be seen from the proof or the more general theorem in the next section. Because the conditions are not natural for large values of δ, it is usually better to argue the consistency by other means.

5.8.1 *Nuisance Parameters*

In Chapter 25 we need an extension of Theorem 5.52 that allows for a "smoothing" or "nuisance" parameter. We also take the opportunity to insert a number of other refinements, which are sometimes useful.

Let $x \mapsto m_{\theta,\eta}(x)$ be measurable functions indexed by parameters (θ, η), and consider estimators $\hat{\theta}_n$ contained in a set Θ_n that, for a given $\hat{\eta}_n$ contained in a set H_n, maximize the map

$$\theta \mapsto \mathbb{P}_n m_{\theta,\hat{\eta}_n}.$$

The sets Θ_n and H_n need not be metric spaces, but instead we measure the discrepancies between $\hat{\theta}_n$ and θ_0, and $\hat{\eta}_n$ and a limiting value η_0, by nonnegative functions $\theta \mapsto d_\eta(\theta, \theta_0)$ and $\eta \mapsto d(\eta, \eta_0)$, which may be arbitrary.

5.55 ***Theorem.*** *Assume that, for arbitrary functions $e_n: \Theta_n \times H_n \mapsto \mathbb{R}$ and $\phi_n: (0, \infty) \mapsto \mathbb{R}$ such that $\delta \mapsto \phi_n(\delta)/\delta^\beta$ is decreasing for some $\beta < 2$, every $(\theta, \eta) \in \Theta_n \times H_n$, and every $\delta > 0$,*

$$P(m_{\theta,\eta} - m_{\theta_0,\eta}) + e_n(\theta, \eta) \le -d_\eta^2(\theta, \theta_0) + d^2(\eta, \eta_0),$$
$$\mathrm{E}^* \sup_{\substack{d_\eta(\theta,\theta_0)<\delta \\ (\theta,\eta)\in\Theta_n\times H_n}} \left|\mathbb{G}_n(m_{\theta,\eta} - m_{\theta_0,\eta}) - \sqrt{n}\, e_n(\theta, \eta)\right| \le \phi_n(\delta).$$

Let $\delta_n > 0$ satisfy $\phi_n(\delta_n) \le \sqrt{n}\,\delta_n^2$ for every n. If $\mathrm{P}(\hat{\theta}_n \in \Theta_n, \hat{\eta}_n \in H_n) \to 1$ and $\mathbb{P}_n m_{\hat{\theta}_n,\hat{\eta}_n} \ge \mathbb{P}_n m_{\theta_0,\hat{\eta}_n} - O_P(\delta_n^2)$, then $d_{\hat{\eta}_n}(\hat{\theta}_n, \theta_0) = O_P^*\big(\delta_n + d(\hat{\eta}_n, \eta_0)\big)$.

Proof. For simplicity assume that $\mathbb{P}_n m_{\hat{\theta}_n,\hat{\eta}_n} \ge \mathbb{P}_n m_{\theta_0,\hat{\eta}_n}$, without a tolerance term. For each $n \in \mathbb{N}$, $j \in \mathbb{Z}$ and $M > 0$, let $S_{n,j,M}$ be the set

$$\big\{(\theta, \eta) \in \Theta_n \times H_n : 2^{j-1}\delta_n < d_\eta(\theta, \theta_0) \le 2^j\delta_n, d(\eta, \eta_0) \le 2^{-M} d_\eta(\theta, \theta_0)\big\}.$$

Then the intersection of the events $(\hat{\theta}_n, \hat{\eta}_n) \in \Theta_n \times H_n$, and $d_{\hat{\eta}_n}(\hat{\theta}_n, \theta_0) \ge 2^M\big(\delta_n + d(\hat{\eta}_n, \eta_0)\big)$ is contained in the union of the events $\big\{(\hat{\theta}_n, \hat{\eta}_n) \in S_{n,j,M}\big\}$ over $j \ge M$. By the definition of $\hat{\theta}_n$, the supremum of $\mathbb{P}_n(m_{\theta,\eta} - m_{\theta_0,\eta})$ over the set of parameters $(\theta, \eta) \in S_{n,j,M}$ is nonnegative on the event $\big\{(\hat{\theta}_n, \hat{\eta}_n) \in S_{n,j,M}\big\}$. Conclude that

$$\mathrm{P}^*\Big(d_{\hat{\eta}_n}(\hat{\theta}_n, \theta_0) \ge 2^M\big(\delta_n + d(\hat{\eta}_n, \eta_0)\big), (\hat{\theta}_n, \hat{\eta}_n) \in \Theta_n \times H_n\Big)$$
$$\le \sum_{j \ge M} \mathrm{P}^*\Big(\sup_{(\theta,\eta)\in S_{n,j,M}} \mathbb{P}_n(m_{\theta,\eta} - m_{\theta_0,\eta}) \ge 0\Big).$$

For every j, $(\theta, \eta) \in S_{n,j,M}$, and every sufficiently large M,

$$P(m_{\theta,\eta} - m_{\theta_0,\eta}) + e_n(\theta, \eta) \le -d_\eta^2(\theta, \theta_0) + d^2(\eta, \eta_0)$$
$$\le -(1 - 2^{-2M})\, d_\eta^2(\theta, \theta_0) \le -2^{2j-4}\delta_n^2.$$

From here on the proof is the same as the proof of Theorem 5.52, except that we use that $\phi_n(c\delta) \le c^\beta \phi_n(\delta)$ for every $c > 1$, by the assumption on ϕ_n. ■

*5.9 Argmax Theorem

The consistency of a sequence of M-estimators can be understood as the points of maximum $\hat\theta_n$ of the criterion functions $\theta \mapsto M_n(\theta)$ converging in probability to a point of maximum of the limit criterion function $\theta \mapsto M(\theta)$. So far we have made no attempt to understand the distributional limit properties of a sequence of M-estimators in a similar way. This is possible, but it is somewhat more complicated and is perhaps best studied after developing the theory of weak convergence of stochastic processes, as in Chapters 18 and 19.

Because the estimators $\hat\theta_n$ typically converge to constants, it is necessary to rescale them before studying distributional limit properties. Thus, we start by searching for a sequence of numbers $r_n \mapsto \infty$ such that the sequence $\hat h_n = r_n(\hat\theta_n - \theta)$ is uniformly tight. The results of the preceding section should be useful. If $\hat\theta_n$ maximizes the function $\theta \mapsto M_n(\theta)$, then the rescaled estimators $\hat h_n$ are maximizers of the *local criterion functions*

$$h \mapsto M_n\left(\theta + \frac{h}{r_n}\right) - M_n(\theta_0).$$

Suppose that these, if suitably normed, converge to a limit process $h \mapsto M(h)$. Then the general principle is that the sequence $\hat h_n$ converges in distribution to the maximizer of this limit process.

For simplicity of notation we shall write the local criterion functions as $h \mapsto M_n(h)$. Let $\{M_n(h): h \in H_n\}$ be arbitrary stochastic processes indexed by subsets H_n of a given metric space. We wish to prove that the argmax-functional is continuous: If $M_n \rightsquigarrow M$ and $H_n \to H$ in a suitable sense, then the (near) maximizers $\hat h_n$ of the random maps $h \mapsto M_n(h)$ converge in distribution to the maximizer $\hat h$ of the limit process $h \mapsto M(h)$. It is easy to find examples in which this is not true, but given the right definitions it is, under some conditions. Given a set B, set

$$M(B) = \sup_{h \in B} M(h).$$

Then convergence in distribution of the vectors $\big(M_n(A), M_n(B)\big)$ for given pairs of sets A and B is an appropriate form of convergence of M_n to M. The following theorem gives some flexibility in the choice of the indexing sets. We implicitly either assume that the suprema $M_n(B)$ are measurable or understand the weak convergence in terms of outer probabilities, as in Chapter 18.

The result we are looking for is not likely to be true if the maximizer of the limit process is not well defined. Exactly as in Theorem 5.7, the maximum should be "well separated." Because in the present case the limit is a stochastic process, we require that every sample path $h \mapsto M(h)$ possesses a well-separated maximum (condition (5.57)).

5.56 ***Theorem (Argmax theorem).*** *Let M_n and M be stochastic processes indexed by subsets H_n and H of a given metric space such that, for every pair of a closed set F and a set K in a given collection $\mathcal{K}$,*

$$\big(M_n(F \cap K \cap H_n),\, M_n(K \cap H_n)\big) \rightsquigarrow \big(M(F \cap K \cap H),\, M(K \cap H)\big).$$

Furthermore, suppose that every sample path of the process $h \mapsto M(h)$ possesses a well-separated point of maximum $\hat{h}$ in that, for every open set G and every $K \in \mathcal{K}$,

$$M(\hat{h}) > M(G^c \cap K \cap H), \qquad \text{if } \hat{h} \in G, \qquad \text{a.s..} \tag{5.57}$$

If $M_n(\hat{h}_n) \geq M_n(H_n) - o_P(1)$ and for every $\varepsilon > 0$ there exists $K \in \mathcal{K}$ such that $\sup_n \mathrm{P}(\hat{h}_n \notin K) < \varepsilon$ and $\mathrm{P}(\hat{h} \notin K) < \varepsilon$, then $\hat{h}_n \rightsquigarrow \hat{h}$.

Proof. If $\hat{h}_n \in F \cap K$, then $M_n(F \cap K \cap H_n) \geq M_n(B) - o_P(1)$ for any set B. Hence, for every closed set F and every $K \in \mathcal{K}$,

$$\begin{aligned} \mathrm{P}(\hat{h}_n \in F \cap K) &\leq \mathrm{P}\big(M_n(F \cap K \cap H_n) \geq M_n(K \cap H_n) - o_P(1)\big) \\ &\leq \mathrm{P}\big(M(F \cap K \cap H) \geq M(K \cap H)\big) + o(1), \end{aligned}$$

by Slutsky's lemma and the portmanteau lemma. If $\hat{h} \in F^c$, then $M(F \cap K \cap H)$ is strictly smaller than $M(\hat{h})$ by (5.57) and hence on the intersection with the event in the far right side $\hat{h}$ cannot be contained in $K \cap H$. It follows that

$$\limsup \mathrm{P}(\hat{h}_n \in F \cap K) \leq \mathrm{P}(\hat{h} \in F) + \mathrm{P}(\hat{h} \notin K \cap H).$$

By assumption we can choose K such that the left and right sides change by less than ε if we replace K by the whole space. Hence $\hat{h}_n \rightsquigarrow \hat{h}$ by the portmanteau lemma. ■

The theorem works most smoothly if we can take $\mathcal{K}$ to consist only of the whole space. However, then we are close to assuming some sort of global uniform convergence of M_n to M, and this may not hold or be hard to prove. It is usually more economical in terms of conditions to show that the maximizers $\hat{h}_n$ are contained in certain sets K, with high probability. Then uniform convergence of M_n to M on K is sufficient. The choice of compact sets K corresponds to establishing the uniform tightness of the sequence $\hat{h}_n$ before applying the argmax theorem.

If the sample paths of the processes M_n are bounded on K and $H_n = H$ for every n, then the weak convergence of the processes M_n viewed as elements of the space $\ell^\infty(K)$ implies the convergence condition of the argmax theorem. This follows by the continuous-mapping theorem, because the map

$$z \mapsto \big(z(A \cap K), z(B \cap K)\big)$$

from $\ell^\infty(K)$ to $\mathbb{R}^2$ is continuous, for every pair of sets A and B. The weak convergence in $\ell^\infty(K)$ remains sufficient if the sets H_n depend on n but converge in a suitable way. Write $H_n \to H$ if H is the set of all limits $\lim h_n$ of converging sequences h_n with $h_n \in H_n$ for every n and, moreover, the limit $h = \lim_i h_{n_i}$ of every converging sequence h_{n_i} with $h_{n_i} \in H_{n_i}$ for every i is contained in H.

5.58 *Corollary.* *Suppose that $M_n \rightsquigarrow M$ in $\ell^\infty(K)$ for every compact subset K of $\mathbb{R}^k$, for a limit process M with continuous sample paths that have unique points of maxima $\hat{h}$. If $H_n \to H$, $M_n(\hat{h}_n) \geq M_n(H_n) - o_P(1)$, and the sequence $\hat{h}_n$ is uniformly tight, then $\hat{h}_n \rightsquigarrow \hat{h}$.*

Proof. The compactness of K and the continuity of the sample paths $h \mapsto M(h)$ imply that the (unique) points of maximum $\hat{h}$ are automatically well separated in the sense of (5.57). Indeed, if this fails for a given open set $G \ni \hat{h}$ and K (and a given ω in the underlying probability space), then there exists a sequence h_m in $G^c \cap K \cap H$ such that $M(h_m) \to M(\hat{h})$. If K is compact, then this sequence can be chosen convergent. The limit h_0 must be in the closed set G^c and hence cannot be $\hat{h}$. By the continuity of M it also has the property that $M(h_0) = \lim M(h_m) = M(\hat{h})$. This contradicts the assumption that $\hat{h}$ is a unique point of maximum.

If we can show that $\big(M_n(F \cap H_n), M_n(K \cap H_n)\big)$ converges to the corresponding limit for every compact sets $F \subset K$, then the theorem is a corollary of Theorem 5.56. If $H_n = H$ for every n, then this convergence is immediate from the weak convergence of M_n to M in $\ell^\infty(K)$, by the continuous-mapping theorem. For H_n changing with n this convergence may fail, and we need to refine the proof of Theorem 5.56. This goes through with minor changes if

$$\limsup_{n\to\infty} \mathrm{P}\big(M_n(F \cap H_n) - M_n(\mathring{K} \cap H_n) \geq x\big) \leq \mathrm{P}\big(M(F \cap H) - M(\mathring{K} \cap H) \geq x\big),$$

for every x, every compact set F and every large closed ball K. Define functions $g_n\colon \ell^\infty(K) \mapsto \mathbb{R}$ by

$$g_n(z) = \sup_{h \in F \cap H_n} z(h) - \sup_{h \in \mathring{K} \cap H_n} z(h),$$

and g similarly, but with H replacing H_n. By an argument as in the proof of Theorem 18.11, the desired result follows if $\limsup g_n(z_n) \leq g(z)$ for every sequence $z_n \to z$ in $\ell^\infty(K)$ and continuous function z. (Then $\limsup \mathrm{P}(g_n(M_n) \geq x) \leq \mathrm{P}(g(M) \geq x)$ for every x, for any weakly converging sequence $M_n \rightsquigarrow M$ with a limit with continuous sample paths.) This in turn follows if for every precompact set $B \subset K$,

$$\sup_{h \in \mathring{B} \cap H} z(h) \leq \overline{\lim_{n\to\infty}} \sup_{h \in B \cap H_n} z_n(h) \leq \sup_{h \in \overline{B} \cap H} z(h).$$

To prove the upper inequality, select $h_n \in B \cap H_n$ such that

$$\sup_{h \in B \cap H_n} z_n(h) = z_n(h_n) + o(1) = z(h_n) + o(1).$$

Because $\overline{B}$ is compact, every subsequence of h_n has a converging subsequence. Because $H_n \to H$, the limit h must be in $\overline{B} \cap H$. Because $z(h_n) \to z(h)$, the upper bound follows.

To prove the lower inequality, select for given $\varepsilon > 0$ an element $h \in \mathring{B} \cap H$ such that

$$\sup_{h \in \mathring{B} \cap H} z(h) \leq z(h) + \varepsilon.$$

Because $H_n \to H$, there exists $h_n \in H_n$ with $h_n \to h$. This sequence must be in $\mathring{B} \subset B$ eventually, whence $z(h) = \lim z(h_n) = \lim z_n(h_n)$ is bounded above by $\liminf \sup_{h \in B \cap H_n} z_n(h)$. ■

The argmax theorem can also be used to prove consistency, by applying it to the original criterion functions $\theta \mapsto M_n(\theta)$. Then the limit process $\theta \mapsto M(\theta)$ is degenerate, and has a fixed point of maximum θ_0. Weak convergence becomes convergence in probability, and the theorem now gives conditions for the consistency $\hat{\theta}_n \xrightarrow{P} \theta_0$. Condition (5.57) reduces to the well-separation of θ_0, and the convergence

$$\sup_{\theta \in F \cap K \cap \Theta_n} M_n(\theta) \xrightarrow{P} \sup_{\theta \in F \cap K \cap \Theta} M_n(\theta)$$

is, apart from allowing Θ_n to depend on n, weaker than the uniform convergence of M_n to M.

Notes

In the section on consistency we have given two main results (uniform convergence and Wald's proof) that have proven their value over the years, but there is more to say on this subject. The two approaches can be unified by replacing the uniform convergence by "one-sided uniform convergence," which in the case of i.i.d. observations can be established under the conditions of Wald's theorem by a bracketing approach as in Example 19.8 (but then one-sided). Furthermore, the use of special properties, such as convexity of the ψ or m functions, is often helpful. Examples such as Lemma 5.10, or the treatment of maximum likelihood estimators in exponential families in Chapter 4, appear to indicate that no single approach can be satisfactory.

The study of the asymptotic properties of maximum likelihood estimators and other M-estimators has a long history. Fisher [48], [50] was a strong advocate of the method of maximum likelihood and noted its asymptotic optimality as early as the 1920s. What we have labelled the classical conditions correspond to the rigorous treatment given by Cramér [27] in his authoritative book. Huber initiated the systematic study of M-estimators, with the purpose of developing robust statistical procedures. His paper [78] contains important ideas that are precursors for the application of techniques from the theory of empirical processes by, among others, Pollard, as in [117], [118], and [120]. For one-dimensional parameters these empirical process methods can be avoided by using a maximal inequality based on the L_2-norm (see, e.g., Theorem 2.2.4 in [146]). Surprisingly, then a Lipschitz condition on the Hellinger distance (an integrated quantity) suffices; see for example, [80] or [94]. For higher-dimensional parameters the results are also not the best possible, but I do not know of any simple better ones.

The books by Huber [79] and by Hampel, Ronchetti, Rousseeuw, and Stahel [73] are good sources for applications of M-estimators in robust statistics. These references also discuss the relative efficiency of the different M-estimators, which motivates, for instance, the use of Huber's ψ-function. In this chapter we have derived Huber's estimator as the solution of the problem of minimizing the asymptotic variance under the side condition of a uniformly bounded influence function. Originally Huber derived it as the solution to the problem of minimizing the maximum asymptotic variance $\sup_P \sigma_P^2$ for P ranging over a contamination neighborhood: $P = (1 - \varepsilon)\Phi + \varepsilon Q$ with Q arbitrary. For M-estimators these two approaches turn out to be equivalent.

The one-step method can be traced back to numerical schemes for solving the likelihood equations, including Fisher's method of scoring. One-step estimators were introduced for

their asymptotic efficiency by Le Cam in 1956, who later developed them for general locally asymptotically quadratic models, and also introduced the discretization device, (see [93]).

PROBLEMS

1. Let $X_1, \ldots, X_n$ be a sample from a density that is strictly positive and symmetric about some point. Show that the Huber M-estimator for location is consistent for the symmetry point.
2. Find an expression for the asymptotic variance of the Huber estimator for location if the observations are normally distributed.
3. Define $\psi(x) = 1 - p, 0, p$ if $x < 0, 0, > 0$. Show that $\mathrm{E}\psi(X - \theta) = 0$ implies that $\mathrm{P}(X < \theta) \leq p \leq \mathrm{P}(X \leq \theta)$.
4. Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$-distributed. Derive the maximum likelihood estimator for (μ, σ^2) and show that it is asymptotically normal. Calculate the Fisher information matrix for this parameter and its inverse.
5. Let $X_1, \ldots, X_n$ be i.i.d. Poisson$(1/\theta)$-distributed. Derive the maximum likelihood estimator for θ and show that it is asymptotically normal.
6. Let $X_1, \ldots, X_n$ be i.i.d. $N(\theta, \theta)$-distributed. Derive the maximum likelihood estimator for θ and show that it is asymptotically normal.
7. Find a sequence of fixed (nonrandom) functions $M_n: \mathbb{R} \mapsto \mathbb{R}$ that converges pointwise to a limit M_0 and such that each M_n has a unique maximum at a point θ_n, but the sequence θ_n does not converge to θ_0. Can you also find a sequence M_n that converges uniformly?
8. Find a sequence of fixed (nonrandom) functions $M_n: \mathbb{R} \mapsto \mathbb{R}$ that converges pointwise but not uniformly to a limit M_0 such that each M_n has a unique maximum at a point θ_n and the sequence θ_n converges to θ_0.
9. Let $X_1, \ldots, X_n$ be i.i.d. observations from a uniform distribution on $[0, \theta]$. Show that the sequence of maximum likelihood estimators is asymptotically consistent. Show that it is not asymptotically normal.
10. Let $X_1, \ldots, X_n$ be i.i.d. observations from an exponential density $\theta \exp(-\theta x)$. Show that the sequence of maximum likelihood estimators is asymptotically normal.
11. Let $\mathbb{F}_n^{-1}(p)$ be a pth sample quantile of a sample from a cumulative distribution F on $\mathbb{R}$ that is differentiable with positive derivative at the population pth-quantile $F^{-1}(p) = \inf\{x: F(x) \geq p\}$. Show that $\sqrt{n}\big(\mathbb{F}_n^{-1}(p) - F^{-1}(p)\big)$ is asymptotically normal with mean zero and variance $p(1-p)/f\big(F^{-1}(p)\big)^2$.
12. Derive a minimal condition on the distribution function F that guarantees the consistency of the sample pth quantile.
13. Calculate the asymptotic variance of $\sqrt{n}(\hat{\theta}_n - \theta)$ in Example 5.26.
14. Suppose that we observe a random sample from the distribution of (X, Y) in the following *errors-in-variables* model:

$$X = Z + e$$
$$Y_i = \alpha + \beta Z + f,$$

where (e, f) is bivariate normally distributed with mean 0 and covariance matrix $\sigma^2 I$ and is independent from the unobservable variable Z. In analogy to Example 5.26, construct a system of estimating equations for (α, β) based on a conditional likelihood, and study the limit properties of the corresponding estimators.

15. In Example 5.27, for what point is the least squares estimator $\hat{\theta}_n$ consistent if we drop the condition that $\mathrm{E}(e \mid X) = 0$? Derive an (implicit) solution in terms of the function $\mathrm{E}(e \mid X)$. Is it necessarily θ_0 if $\mathrm{E}e = 0$?

16. In Example 5.27, consider the asymptotic behavior of the least absolute-value estimator $\hat{\theta}$ that minimizes $\sum_{i=1}^{n} |Y_i - \phi_\theta(X_i)|$.

17. Let $X_1, \ldots, X_n$ be i.i.d. with density $f_{\lambda,a}(x) = \lambda e^{-\lambda(x-a)} 1\{x \geq a\}$, where the parameters $\lambda > 0$ and $a \in \mathbb{R}$ are unknown. Calculate the maximum likelihood estimator $(\hat{\lambda}_n, \hat{a}_n)$ of (λ, a) and derive its asymptotic properties.

18. Let X be Poisson-distributed with density $p_\theta(x) = \theta^x e^{-\theta}/x!$. Show by direct calculation that $\mathrm{E}_\theta \dot{\ell}_\theta(X) = 0$ and $\mathrm{E}_\theta \ddot{\ell}_\theta(X) = -\mathrm{E}_\theta \dot{\ell}_\theta^2(X)$. Compare this with the assertions in the introduction. Apparently, differentiation under the integral (sum) is permitted in this case. Is that obvious from results from measure theory or (complex) analysis?

19. Let $X_1, \ldots, X_n$ be a sample from the $N(\theta, 1)$ distribution, where it is known that $\theta \geq 0$. Show that the maximum likelihood estimator is not asymptotically normal under $\theta = 0$. Why does this not contradict the theorems of this chapter?

20. Show that $(\tilde{\theta} - \theta_0)\Psi_n(\tilde{\theta}_n)$ in formula (5.18) converges in probability to zero if $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$, and that there exists an integrable function M and $\delta > 0$ with $|\dot{\psi}_\theta(x)| \leq M(x)$ for every x and every $\|\theta - \theta_0\| < \delta$.

21. If $\hat{\theta}_n$ maximizes M_n, then it also maximizes M_n^+. Show that this may be used to relax the conditions of Theorem 5.7 to $\sup_\theta |M_n^+ - M^+|(\theta) \to 0$ in probability $\big(\text{if } M(\theta_0) > 0\big)$.

22. Suppose that for every $\varepsilon > 0$ there exists a set Θ_ε with $\liminf \mathrm{P}(\hat{\theta}_n \in \Theta_\varepsilon) \geq 1 - \varepsilon$. Then uniform convergence of M_n to M in Theorem 5.7 can be relaxed to uniform convergence on every Θ_ε.

23. Show that Wald's consistency proof yields almost sure convergence of $\hat{\theta}_n$, rather than convergence in probability if the parameter space is compact and $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o(1)$.

24. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. and satisfy the linear regression relationship $Y_i = \theta^T X_i + e_i$ for (unobservable) errors $e_1, \ldots, e_n$ independent of $X_1, \ldots, X_n$. Show that the mean absolute deviation estimator, which minimizes $\sum |Y_i - \theta X_i|$, is asymptotically normal under a mild condition on the error distribution.

25. (i) Verify the conditions of Wald's theorem for m_θ the log likelihood function of the $N(\mu, \sigma^2)$-distribution if the parameter set for $\theta = (\mu, \sigma^2)$ is a compact subset of $\mathbb{R} \times \mathbb{R}^+$.

(ii) Extend m_θ by continuity to the compactification of $\mathbb{R} \times \mathbb{R}^+$. Show that the conditions of Wald's theorem fail at the points $(\mu, 0)$.

(iii) Replace m_θ by the log likelihood function of a pair of two independent observations from the $N(\mu, \sigma^2)$-distribution. Show that Wald's theorem now does apply, also with a compactified parameter set.

26. A distribution on $\mathbb{R}^k$ is called *ellipsoidally symmetric* if it has a density of the form $x \mapsto g\big((x-\mu)^T \Sigma^{-1}(x-\mu)\big)$ for a function $g: [0, \infty) \mapsto [0, \infty)$, a vector μ, and a symmetric positive-definite matrix Σ. Study the Z-estimators for location $\hat{\mu}$ that solve an equation of the form

$$\sum_{i=1}^{n} \psi\big((X_i - \mu)^T \hat{\Sigma}_n^{-1}(X_i - \mu)\big),$$

for given estimators $\hat{\Sigma}_n$ and, for instance, Huber's ψ-function. Is the asymptotic distribution of $\hat{\Sigma}_n$ important?

27. Suppose that Θ is a compact metric space and $M: \Theta \to \mathbb{R}$ is continuous. Show that (5.8) is equivalent to the point θ_0 being a point of unique global maximum. Can you relax the continuity of M to some form of "semi-continuity"?