7

Local Asymptotic Normality

A sequence of statistical models is "locally asymptotically normal" if, asymptotically, their likelihood ratio processes are similar to those for a normal location parameter. Technically, this is if the likelihood ratio processes admit a certain quadratic expansion. An important example in which this arises is repeated sampling from a smooth parametric model. Local asymptotic normality implies convergence of the models to a Gaussian model after a rescaling of the parameter.

7.1 Introduction

Suppose we observe a sample $X_1, \ldots, X_n$ from a distribution P_θ on some measurable space $(\mathcal{X}, \mathcal{A})$ indexed by a parameter θ that ranges over an open subset Θ of $\mathbb{R}^k$. Then the full observation is a single observation from the product P_θ^n of n copies of P_θ, and the statistical model is completely described as the collection of probability measures $\{P_\theta^n : \theta \in \Theta\}$ on the sample space $(\mathcal{X}^n, \mathcal{A}^n)$. In the context of the present chapter we shall speak of a *statistical experiment*, rather than of a statistical model. In this chapter it is shown that many statistical experiments can be approximated by Gaussian experiments after a suitable reparametrization.

The reparametrization is centered around a fixed parameter θ_0, which should be regarded as known. We define a *local parameter* $h = \sqrt{n}(\theta - \theta_0)$, rewrite P_θ^n as $P_{\theta_0 + h/\sqrt{n}}^n$, and thus obtain an experiment with parameter h. In this chapter we show that, for large n, the experiments

$$\left(P_{\theta_0+h/\sqrt{n}}^n : h \in \mathbb{R}^k\right) \quad \text{and} \quad \left(N\left(h, I_{\theta_0}^{-1}\right) : h \in \mathbb{R}^k\right)$$

are similar in statistical properties, whenever the original experiments $\theta \mapsto P_\theta$ are "smooth" in the parameter. The second experiment consists of observing a single observation from a normal distribution with mean h and known covariance matrix (equal to the inverse of the Fisher information matrix). This is a simple experiment, which is easy to analyze, whence the approximation yields much information about the asymptotic properties of the original experiments. This information is extracted in several chapters to follow and concerns both asymptotic optimality theory and the behavior of statistical procedures such as the maximum likelihood estimator and the likelihood ratio test.

We have taken the local parameter set equal to $\mathbb{R}^k$, which is not correct if the parameter set Θ is a true subset of $\mathbb{R}^k$. If θ_0 is an inner point of the original parameter set, then the vector $\theta = \theta_0 + h/\sqrt{n}$ is a parameter in Θ for a given h, for every sufficiently large n, and the local parameter set converges to the whole of $\mathbb{R}^k$ as $n \to \infty$. Then taking the local parameter set equal to $\mathbb{R}^k$ does not cause errors. To give a meaning to the results of this chapter, the measure $P_{\theta_0+h/\sqrt{n}}$ may be defined arbitrarily if $\theta_0 + h/\sqrt{n} \notin \Theta$.

7.2 Expanding the Likelihood

The convergence of the local experiments is defined and established later in this chapter. First, we discuss the technical tool: a Taylor expansion of the logarithm of the likelihood. Let p_θ be a density of P_θ with respect to some measure μ. Assume for simplicity that the parameter is one-dimensional and that the log likelihood $\ell_\theta(x) = \log p_\theta(x)$ is twice-differentiable with respect to θ, for every x, with derivatives $\dot{\ell}_\theta(x)$ and $\ddot{\ell}_\theta(x)$. Then, for every fixed x,

$$\log \frac{p_{\theta+h}}{p_\theta}(x) = h\dot{\ell}_\theta(x) + \frac{1}{2}h^2\ddot{\ell}_\theta(x) + o_x(h^2).$$

The subscript x in the remainder term is a reminder of the fact that this term depends on x as well as on h. It follows that

$$\log \prod_{i=1}^n \frac{p_{\theta+h/\sqrt{n}}}{p_\theta}(X_i) = \frac{h}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_\theta(X_i) + \frac{1}{2}\frac{h^2}{n} \sum_{i=1}^n \ddot{\ell}_\theta(X_i) + \mathrm{Rem}_n.$$

Here the score has mean zero, $P_\theta \dot{\ell}_\theta = 0$, and $-P_\theta \ddot{\ell}_\theta = P_\theta \dot{\ell}_\theta^2 = I_\theta$ equals the Fisher information for θ (see, e.g., section 5.5). Hence the first term can be rewritten as $h\Delta_{n,\theta}$, where $\Delta_{n,\theta} = n^{-1/2} \sum_{i=1}^n \dot{\ell}_\theta(X_i)$ is asymptotically normal with mean zero and variance I_θ, by the central limit theorem. Furthermore, the second term in the expansion is asymptotically equivalent to $-\frac{1}{2}h^2 I_\theta$, by the law of large numbers. The remainder term should behave as $o(1/n)$ times a sum of n terms and hopefully is asymptotically negligible. Consequently, under suitable conditions we have, for every h,

$$\log \prod_{i=1}^n \frac{p_{\theta+h/\sqrt{n}}}{p_\theta}(X_i) = h\Delta_{n,\theta} - \frac{1}{2}I_\theta h^2 + o_{P_\theta}(1).$$

In the next section we see that this is similar in form to the likelihood ratio process of a Gaussian experiment. Because this expansion concerns the likelihood process in a neighborhood of θ, we speak of "local asymptotic normality" of the sequence of models $\{P_\theta^n : \theta \in \Theta\}$.

The preceding derivation can be made rigorous under moment or continuity conditions on the second derivative of the log likelihood. Local asymptotic normality was originally deduced in this manner. Surprisingly, it can also be established under a single condition that only involves a first derivative: differentiability of the root density $\theta \mapsto \sqrt{p_\theta}$ in quadratic mean. This entails the existence of a vector of measurable functions $\dot{\ell}_\theta = (\dot{\ell}_{\theta,1}, \ldots, \dot{\ell}_{\theta,k})^T$ such that

$$\int \left[\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2}h^T \dot{\ell}_\theta \sqrt{p_\theta} \right]^2 d\mu = o\big(\|h\|^2\big), \qquad h \to 0. \tag{7.1}$$

If this condition is satisfied, then the model $(P_\theta : \theta \in \Theta)$ is called *differentiable in quadratic mean* at θ.

Usually, $\frac{1}{2}h^T \dot{\ell}_\theta(x)\sqrt{p_\theta(x)}$ is the derivative of the map $h \mapsto \sqrt{p_{\theta+h}(x)}$ at $h = 0$ for (almost) every x. In this case

$$\dot{\ell}_\theta(x) = 2\frac{1}{\sqrt{p_\theta(x)}}\frac{\partial}{\partial\theta}\sqrt{p_\theta(x)} = \frac{\partial}{\partial\theta}\log p_\theta(x).$$

Condition (7.1) does not require differentiability of the map $\theta \mapsto p_\theta(x)$ for any single x, but rather differentiability in (quadratic) mean. Admittedly, the latter is typically established by pointwise differentiability plus a convergence theorem for integrals. Because the condition is exactly right for its purpose, we establish in the following theorem local asymptotic normality under (7.1). A lemma following the theorem gives easily verifiable conditions in terms of pointwise derivatives.

7.2 *Theorem.* *Suppose that Θ is an open subset of $\mathbb{R}^k$ and that the model $(P_\theta : \theta \in \Theta)$ is differentiable in quadratic mean at θ. Then $P_\theta\dot{\ell}_\theta = 0$ and the Fisher information matrix $I_\theta = P_\theta\dot{\ell}_\theta\dot{\ell}_\theta^T$ exists. Furthermore, for every converging sequence $h_n \to h$, as $n \to \infty$,*

$$\log\prod_{i=1}^n \frac{p_{\theta+h_n/\sqrt{n}}}{p_\theta}(X_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^n h^T\dot{\ell}_\theta(X_i) - \frac{1}{2}h^T I_\theta h + o_{P_\theta}(1).$$

Proof. Given a converging sequence $h_n \to h$, we use the abbreviations p_n, p, and g for $p_{\theta+h_n/\sqrt{n}}$, p_θ, and $h^T\dot{\ell}_\theta$, respectively. By (7.1) the sequence $\sqrt{n}(\sqrt{p_n} - \sqrt{p})$ converges in quadratic mean (i.e., in $L_2(\mu)$) to $\frac{1}{2}g\sqrt{p}$. This implies that the sequence $\sqrt{p_n}$ converges in quadratic mean to $\sqrt{p}$. By the continuity of the inner product,

$$Pg = \int \frac{1}{2}g\sqrt{p}\,2\sqrt{p}\,d\mu = \lim\int \sqrt{n}(\sqrt{p_n} - \sqrt{p})\,(\sqrt{p_n} + \sqrt{p})\,d\mu.$$

The right side equals $\sqrt{n}(1-1) = 0$ for every n, because both probability densities integrate to 1. Thus $Pg = 0$.

The random variable $W_{ni} = 2\big[\sqrt{p_n/p}(X_i) - 1\big]$ is with P-probability 1 well defined. By (7.1)

$$\begin{aligned} \operatorname{var}\left(\sum_{i=1}^n W_{ni} - \frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i)\right) &\le \mathrm{E}\big(\sqrt{n}W_{ni} - g(X_i)\big)^2 \to 0, \\ \mathrm{E}\sum_{i=1}^n W_{ni} = 2n\left(\int\sqrt{p_n}\sqrt{p}\,d\mu - 1\right) &= -n\int[\sqrt{p_n} - \sqrt{p}]^2\,d\mu \to -\frac{1}{4}Pg^2. \end{aligned} \tag{7.3}$$

Here $Pg^2 = \int g^2\,dP = h^T I_\theta h$ by the definitions of g and I_θ. If both the means and the variances of a sequence of random variables converge to zero, then the sequence converges to zero in probability. Therefore, combining the preceding pair of displayed equations, we find

$$\sum_{i=1}^n W_{ni} = \frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i) - \frac{1}{4}Pg^2 + o_P(1). \tag{7.4}$$

Next, we express the log likelihood ratio in $\sum_{i=1}^n W_{ni}$ through a Taylor expansion of the logarithm. If we write $\log(1+x) = x - \frac{1}{2}x^2 + x^2 R(2x)$, then $R(x) \to 0$ as $x \to 0$, and

$$\log \prod_{i=1}^n \frac{p_n}{p}(X_i) = 2\sum_{i=1}^n \log\left(1 + \frac{1}{2}W_{ni}\right)$$
$$= \sum_{i=1}^n W_{ni} - \frac{1}{4}\sum_{i=1}^n W_{ni}^2 + \frac{1}{2}\sum_{i=1}^n W_{ni}^2 R(W_{ni}). \tag{7.5}$$

As a consequence of the right side of (7.3), it is possible to write $nW_{ni}^2 = g^2(X_i) + A_{ni}$ for random variables A_{ni} such that $\mathrm{E}|A_{ni}| \to 0$. The averages $\bar{A}_n$ converge in mean and hence in probability to zero. Combination with the law of large numbers yields

$$\sum_{i=1}^n W_{ni}^2 = \overline{(g^2)}_n + \bar{A}_n \xrightarrow{P} Pg^2.$$

By the triangle inequality followed by Markov's inequality,

$$n\mathrm{P}\big(|W_{ni}| > \varepsilon\sqrt{2}\big) \le n\mathrm{P}\big(g^2(X_i) > n\varepsilon^2\big) + n\mathrm{P}\big(|A_{ni}| > n\varepsilon^2\big)$$
$$\le \varepsilon^{-2}Pg^2\{g^2 > n\varepsilon^2\} + \varepsilon^{-2}\mathrm{E}|A_{ni}| \to 0.$$

The left side is an upper bound for $\mathrm{P}\big(\max_{1\le i\le n}|W_{ni}| > \varepsilon\sqrt{2}\big)$. Thus the sequence $\max_{1\le i\le n}|W_{ni}|$ converges to zero in probability. By the property of the function R, the sequence $\max_{1\le i\le n}\big|R(W_{ni})\big|$ converges in probability to zero as well. The last term on the right in (7.5) is bounded by $\max_{1\le i\le n}\big|R(W_{ni})\big|\sum_{i=1}^n W_{ni}^2$. Thus it is $o_P(1)O_P(1)$, and converges in probability to zero. Combine to obtain that

$$\log \prod_{i=1}^n \frac{p_n}{p}(X_i) = \sum_{i=1}^n W_{ni} - \frac{1}{4}Pg^2 + o_P(1).$$

Together with (7.4) this yields the theorem. ■

To establish the differentiability in quadratic mean of specific models requires a convergence theorem for integrals. Usually one proceeds by showing differentiability of the map $\theta \mapsto p_\theta(x)$ for almost every x plus μ-equi-integrability (e.g., domination). The following lemma takes care of most examples.

7.6 *Lemma.* *For every θ in an open subset of $\mathbb{R}^k$ let p_θ be a μ-probability density. Assume that the map $\theta \mapsto s_\theta(x) = \sqrt{p_\theta(x)}$ is continuously differentiable for every x. If the elements of the matrix $I_\theta = \int (\dot{p}_\theta/p_\theta)(\dot{p}_\theta^T/p_\theta)\, p_\theta\, d\mu$ are well defined and continuous in θ, then the map $\theta \mapsto \sqrt{p_\theta}$ is differentiable in quadratic mean (7.1) with $\dot{\ell}_\theta$ given by $\dot{p}_\theta/p_\theta$.*

Proof. By the chain rule, the map $\theta \mapsto p_\theta(x) = s_\theta^2(x)$ is differentiable for every x with gradient $\dot{p}_\theta = 2s_\theta\dot{s}_\theta$. Because s_θ is nonnegative, its gradient $\dot{s}_\theta$ at a point at which $s_\theta = 0$ must be zero. Conclude that we can write $\dot{s}_\theta = \frac{1}{2}(\dot{p}_\theta/p_\theta)\sqrt{p_\theta}$, where the quotient $\dot{p}_\theta/p_\theta$ may be defined arbitrarily if $p_\theta = 0$. By assumption, the map $\theta \mapsto I_\theta = 4\int \dot{s}_\theta\dot{s}_\theta^T\, d\mu$ is continuous.

Because the map $\theta \mapsto s_\theta(x)$ is continuously differentiable, the difference $s_{\theta+h}(x) - s_\theta(x)$ can be written as the integral $\int_0^1 h^T\dot{s}_{\theta+uh}(x)\, du$ of its derivative. By Jensen's (or Cauchy-Schwarz's) inequality, the square of this integral is bounded by the integral $\int_0^1 \big(h^T\dot{s}_{\theta+uh}(x)\big)^2$

du of the square. Conclude that

$$\int \left(\frac{s_{\theta+th_t} - s_\theta}{t} \right)^2 d\mu \le \int\int_0^1 \left(h_t^T \dot{s}_{\theta+uth_t} \right)^2 du\, d\mu = \frac{1}{4} \int_0^1 h_t^T I_{\theta+uth_t} h_t \, du,$$

where the last equality follows by Fubini's theorem and the definition of I_θ. For $h_t \to h$ the right side converges to $\frac{1}{4} h^T I_\theta h = \int (h^T \dot{s}_\theta)^2 \, d\mu$ by the continuity of the map $\theta \mapsto I_\theta$.

By the differentiability of the map $\theta \mapsto s_\theta(x)$ the integrand in

$$\int \left[\frac{s_{\theta+th_t} - s_\theta}{t} - h^T \dot{s}_\theta \right]^2 d\mu$$

converges pointwise to zero. The result of the preceding paragraph combined with Proposition 2.29 shows that the integral converges to zero. ■

7.7 ***Example (Exponential families).*** The preceding lemma applies to most exponential family models

$$p_\theta(x) = d(\theta) h(x) e^{Q(\theta)^T t(x)}.$$

An exponential family model is smooth in its natural parameter (away from the boundary of the natural parameter space). Thus the maps $\theta \mapsto \sqrt{p_\theta(x)}$ are continuously differentiable if the maps $\theta \mapsto Q(\theta)$ are continuously differentiable and map the parameter set Θ into the interior of the natural parameter space. The score function and information matrix equal

$$\dot{\ell}_\theta(x) = Q_\theta' \big(t(x) - \mathrm{E}_\theta t(X) \big), \qquad I_\theta = Q_\theta' \mathrm{cov}_\theta \, t(X) (Q_\theta')^T.$$

Thus the asymptotic expansion of the local log likelihood is valid for most exponential families. □

7.8 ***Example (Location models).*** The preceding lemma also includes all location models $\{ f(x - \theta) : \theta \in \mathbb{R} \}$ for a positive, continuously differentiable density f with finite *Fisher information for location*

$$I_f = \int \left(\frac{f'}{f} \right)^2 (x) \, f(x) \, dx.$$

The score function $\dot{\ell}_\theta(x)$ can be taken equal to $-(f'/f)(x - \theta)$. The Fisher information is equal to I_f for every θ and hence certainly continuous in θ.

By a refinement of the lemma, differentiability in quadratic mean can also be established for slightly irregular shapes, such as the Laplace density $f(x) = \frac{1}{2} e^{-|x|}$. For the Laplace density the map $\theta \mapsto \log f(x - \theta)$ fails to be differentiable at the single point $\theta = x$. At other points the derivative exists and equals $\mathrm{sign}(x - \theta)$. It can be shown that the Laplace location model is differentiable in quadratic mean with score function $\dot{\ell}_\theta(x) = \mathrm{sign}(x - \theta)$. This may be proved by writing the difference $\sqrt{f(x - h)} - \sqrt{f(x)}$ as the integral $\int_0^1 \frac{1}{2} h \, \mathrm{sign}(x - uh) \sqrt{f(x - uh)} \, du$ of its derivative, which is possible even though the derivative does not exist everywhere. Next the proof of the preceding lemma applies. □

7.9 ***Counterexample (Uniform distribution).*** The family of uniform distributions on $[0, \theta]$ is nowhere differentiable in quadratic mean. The reason is that the support of the

uniform distribution depends too much on the parameter. Differentiability in quadratic mean (7.1) does not require that all densities p_θ have the same support. However, restriction of the integral (7.1) to the set $\{p_\theta = 0\}$ yields

$$P_{\theta+h}(p_\theta = 0) = \int_{p_\theta=0} p_{\theta+h}\, d\mu = o(h^2).$$

Thus, under (7.1) the total mass $P_{\theta+h}(p_\theta = 0)$ of $P_{\theta+h}$ that is orthogonal to P_θ must "disappear" as $h \to 0$ at a rate faster than h^2.

This is not true for the uniform distribution, because, for $h \geq 0$,

$$P_{\theta+h}(p_\theta = 0) = \int_{[0,\theta]^c} \frac{1}{\theta+h} 1_{[0,\theta+h]}(x)\, dx = \frac{h}{\theta+h}.$$

The orthogonal part does converge to zero, but only at the rate $O(h)$. □

7.3 Convergence to a Normal Experiment

The true meaning of local asymptotic normality is convergence of the local statistical experiments to a normal experiment. In Chapter 9 the notion of convergence of statistical experiments is introduced in general. In this section we bypass this general theory and establish a direct relationship between the local experiments and a normal limit experiment.

The limit experiment is the experiment that consists of observing a single observation X with the $N(h, I_\theta^{-1})$-distribution. The log likelihood ratio process of this experiment equals

$$\log \frac{dN(h, I_\theta^{-1})}{dN(0, I_\theta^{-1})}(X) = h^T I_\theta X - \frac{1}{2} h^T I_\theta h.$$

The right side is very similar in form to the right side of the expansion of the log likelihood ratio $\log dP^n_{\theta+h/\sqrt{n}}/dP^n_\theta$ given in Theorem 7.2. In view of the similarity, the possibility of a normal approximation is not a complete surprise. The approximation in this section is "local" in nature: We fix θ and think of

$$\left(P^n_{\theta+h/\sqrt{n}} : h \in \mathbb{R}^k\right)$$

as a statistical model with parameter h, for "known" θ. We show that this can be approximated by the statistical model $\left(N(h, I_\theta^{-1}) : h \in \mathbb{R}^k\right)$.

A motivation for studying a local approximation is that, usually, asymptotically, the "true" parameter can be known with unlimited precision. The true statistical difficulty is therefore determined by the nature of the measures P_θ for θ in a small neighbourhood of the true value. In the present situation "small" turns out to be "of size $O(1/\sqrt{n})$."

A relationship between the models that can be statistically interpreted will be described through the possible (limit) distributions of statistics. For each n, let $T_n = T_n(X_1, \ldots, X_n)$ be a statistic in the experiment $(P^n_{\theta+h/\sqrt{n}} : h \in \mathbb{R}^k)$ with values in a fixed Euclidean space. Suppose that the sequence of statistics T_n converges in distribution under every possible (local) parameter:

$$T_n \overset{h}{\rightsquigarrow} L_{\theta,h}, \qquad \text{every } h.$$

Here $\overset{h}{\rightsquigarrow}$ means convergence in distribution under the parameter $\theta + h/\sqrt{n}$, and $L_{\theta,h}$ may be any probability distribution. According to the following theorem, the distributions $\{L_{\theta,h} : h \in \mathbb{R}^k\}$ are necessarily the distributions of a statistic T in the normal experiment $\left(N(h, I_\theta^{-1}) : h \in \mathbb{R}^k\right)$. Thus, every weakly converging sequence of statistics is "matched" by a statistic in the limit experiment. (In the present set-up the vector θ is considered known and the vector h is the statistical parameter. Consequently, by "statistics" T_n and T are understood measurable maps that do not depend on h but may depend on θ.)

This principle of matching estimators is a method to give the convergence of models a statistical interpretation. Most measures of quality of a statistic can be expressed in the distribution of the statistic under different parameters. For instance, if a certain hypothesis is rejected for values of a statistic T_n exceeding a number c, then the power function $h \mapsto \mathrm{P}_h(T_n > c)$ is relevant; alternatively, if T_n is an estimator of h, then the mean square error $h \mapsto \mathrm{E}_h(T_n - h)^2$, or a similar quantity, determines the quality of T_n. Both quality measures depend on the laws of the statistics only. The following theorem asserts that as a function of h the law of a statistic T_n can be well approximated by the law of some statistic T. Then the quality of the approximating T is the same as the "asymptotic quality" of the sequence T_n. Investigation of the possible T should reveal the asymptotic performance of possible sequences T_n. Concrete applications of this principle to testing and estimation are given in later chapters.

A minor technical complication is that it is necessary to allow randomized statistics in the limit experiment. A *randomized statistic* T based on the observation X is defined as a measurable map $T = T(X, U)$ that depends on X but may also depend on an independent variable U with a uniform distribution on $[0, 1]$. Thus, the statistician working in the limit experiment is allowed to base an estimate or test on both the observation and the outcome of an extra experiment that can be run without knowledge of the parameter. In most situations such randomization is not useful, but the following theorem would not be true without it.[†]

7.10 ***Theorem.*** *Assume that the experiment $(P_\theta : \theta \in \Theta)$ is differentiable in quadratic mean (7.1) at the point θ with nonsingular Fisher information matrix I_θ. Let T_n be statistics in the experiments $(P^n_{\theta+h/\sqrt{n}} : h \in \mathbb{R}^k)$ such that the sequence T_n converges in distribution under every h. Then there exists a randomized statistic T in the experiment $\left(N(h, I_\theta^{-1}) : h \in \mathbb{R}^k\right)$ such that $T_n \overset{h}{\rightsquigarrow} T$ for every h.*

Proof. For later reference, it is useful to use the abbreviations

$$P_{n,h} = P^n_{\theta+h/\sqrt{n}}, \qquad J = I_\theta, \qquad \Delta_n = \frac{1}{\sqrt{n}} \sum \dot{\ell}_\theta(X_i).$$

By assumption, the marginals of the sequence (T_n, Δ_n) converge in distribution under $h = 0$; hence they are uniformly tight by Prohorov's theorem. Because marginal tightness implies joint tightness, Prohorov's theorem can be applied in the other direction to see the existence of a subsequence of $\{n\}$ along which

$$(T_n, \Delta_n) \overset{0}{\rightsquigarrow} (S, \Delta),$$

[†] It is not important that U is uniformly distributed. Any randomization mechanism that is sufficiently rich will do.

jointly, for some random vector (S, Δ). The vector Δ is necessarily a marginal weak limit of the sequence Δ_n and hence it is $N(0, J)$-distributed. Combination with Theorem 7.2 yields

$$\left(T_n, \log \frac{dP_{n,h}}{dP_{n,0}}\right) \overset{0}{\rightsquigarrow} \left(S, h^T\Delta - \frac{1}{2}h^T Jh\right).$$

In particular, the sequence $\log dP_{n,h}/dP_{n,0}$ converges to the normal $N(-\frac{1}{2}h^T Jh, h^T Jh)$-distribution. By Example 6.5, the sequences $P_{n,h}$ and $P_{n,0}$ are contiguous. The limit law L_h of T_n under h can therefore be expressed in the joint law on the right, by the general form of Le Cam's third lemma: For each Borel set B

$$L_h(B) = \mathrm{E}1_B(S)e^{h^T\Delta - \frac{1}{2}h^T Jh}.$$

We need to find a statistic T in the normal experiment having this law under h (for every h), using only the knowledge that Δ is $N(0, J)$-distributed.

By the lemma below there exists a randomized statistic T such that, with U uniformly distributed and independent of Δ,[†]

$$\big(T(\Delta, U), \Delta\big) \sim (S, \Delta).$$

Because the random vectors on the left and right sides have the same second marginal distribution, this is the same as saying that $T(\delta, U)$ is distributed according to the conditional distribution of S given $\Delta = \delta$, for almost every δ. As shown in the next lemma, this can be achieved by using the quantile transformation.

Let X be an observation in the limit experiment $\big(N(h, J^{-1}) : h \in \mathbb{R}^k\big)$. Then JX is under $h = 0$ normally $N(0, J)$-distributed and hence it is equal in distribution to Δ. Furthermore, by Fubini's theorem,

$$\begin{aligned}\mathrm{P}_h\big(T(JX, U) \in B\big) &= \int \mathrm{P}\big(T(Jx, U) \in B\big)\, e^{-\frac{1}{2}(x-h)^T J(x-h)} \sqrt{\frac{\det J}{(2\pi)^k}}\, dx \\ &= \mathrm{E}_0 1_B\big(T(JX, U)\big)\, e^{h^T JX - \frac{1}{2}h^T Jh}.\end{aligned}$$

This equals $L_h(B)$, because, by construction, the vector $\big(T(JX, U), JX\big)$ has the same distribution under $h = 0$ as (S, Δ). The randomized statistic $T(JX, U)$ has law L_h under h and hence satisfies the requirements. ■

7.11 ***Lemma.*** *Given a random vector (S, Δ) with values in $\mathbb{R}^d \times \mathbb{R}^k$ and an independent uniformly $[0, 1]$ random variable U (defined on the same probability space), there exists a jointly measurable map T on $\mathbb{R}^k \times [0, 1]$ such that $\big(T(\Delta, U), \Delta\big)$ and (S, Δ) are equal in distribution.*

Proof. For simplicity of notation we only give a construction for $d = 2$. It is possible to produce two independent uniform $[0, 1]$ variables U_1 and U_2 from one given $[0, 1]$ variable U. (For instance, construct U_1 and U_2 from the even and odd numbered digits in the decimal expansion of U.) Therefore it suffices to find a statistic $T = T(\Delta, U_1, U_2)$ such that (T, Δ) and (S, Δ) are equal in law. Because the second marginals are equal, it

[†] The symbol $\sim$ means "equal-in-law."

suffices to construct T such that $T(\delta, U_1, U_2)$ is equal in distribution to S given $\Delta = \delta$, for every $\delta \in \mathbb{R}^k$. Let $Q_1(u_1 \mid \delta)$ and $Q_2(u_2 \mid \delta, s_1)$ be the quantile functions of the conditional distributions

$$P^{S_1 \mid \Delta = \delta} \qquad \text{and} \qquad P^{S_2 \mid \Delta = \delta, S_1 = s_1},$$

respectively. These are measurable functions in their two and three arguments, respectively. Furthermore, $Q_1(U_1|\delta)$ has law $P^{S_1|\Delta=\delta}$ and $Q_2(U_2|\delta, s_1)$ has law $P^{S_2|\Delta=\delta, S_1=s_1}$, for every δ and s_1. Set

$$T(\delta, U_1, U_2) = \Big(Q_1(U_1|\delta), \quad Q_2\big(U_2|\delta, Q_1(U_1|\delta)\big)\Big).$$

Then the first coordinate $Q_1(U_1|\delta)$ of $T(\delta, U_1, U_2)$ possesses the distribution $P^{S_1|\Delta=\delta}$. Given that this first coordinate equals s_1, the second coordinate is distributed as $Q_2(U_2|\delta, s_1)$, which has law $P^{S_2|\Delta=\delta, S_1=s_1}$ by construction. Thus T satisfies the requirements. ■

7.4 Maximum Likelihood

Maximum likelihood estimators in smooth parametric models were shown to be asymptotically normal in Chapter 5. The convergence of the local experiments to a normal limit experiment gives an insightful explanation of this fact.

By the representation theorem, Theorem 7.10, every sequence of statistics in the local experiments $(P^n_{\theta+h/\sqrt{n}} : h \in \mathbb{R}^k)$ is matched in the limit by a statistic in the normal experiment. Although this does not follow from this theorem, a sequence of maximum likelihood estimators is typically matched by the maximum likelihood estimator in the limit experiment. Now the maximum likelihood estimator for h in the experiment $\big(N(h, I_\theta^{-1}) : h \in \mathbb{R}^k\big)$ is the observation X itself (the mean of a sample of size one), and this is normally distributed. Thus, we should expect that the maximum likelihood estimators $\hat{h}_n$ for the local parameter h in the experiments $(P^n_{\theta+h/\sqrt{n}} : h \in \mathbb{R}^k)$ converge in distribution to X. In terms of the original parameter θ, the local maximum likelihood estimator $\hat{h}_n$ is the standardized maximum likelihood estimator $\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta)$. Furthermore, the local parameter $h = 0$ corresponds to the value θ of the original parameter. Thus, we should expect that under θ the sequence $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution to X under $h = 0$, that is, to the $N(0, I_\theta^{-1})$-distribution.

As a heuristic explanation of the asymptotic normality of maximum likelihood estimators the preceding argument is much more insightful than the proof based on linearization of the score equation. It also explains why, or in what sense, the maximum likelihood estimator is asymptotically optimal: in the same sense as the maximum likelihood estimator of a Gaussian location parameter is optimal.

This heuristic argument cannot be justified under just local asymptotic normality, which is too weak a connection between the sequence of local experiments and the normal limit experiment for this purpose. Clearly, the argument is valid under the conditions of Theorem 5.39, because the latter theorem guarantees the asymptotic normality of the maximum likelihood estimator. This theorem adds a Lipschitz condition on the maps $\theta \mapsto \log p_\theta(x)$, and the "global" condition that $\hat{\theta}_n$ is consistent to differentiability in quadratic mean. In the following theorem, we give a direct argument, and also allow that θ is not an inner point of the parameter set, so that the local parameter spaces may not converge to the full space $\mathbb{R}^k$.

Then the maximum likelihood estimator in the limit experiment is a "projection" of X and the limit distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ may change accordingly.

Let Θ be an arbitrary subset of $\mathbb{R}^k$ and define H_n as the *local parameter space* $H_n = \sqrt{n}(\Theta - \theta)$. Then $\hat{h}_n$ is the maximizer over H_n of the random function (or "process")

$$h \mapsto \log \frac{dP^n_{\theta+h/\sqrt{n}}}{dP^n_\theta}.$$

If the experiment $(P_\theta : \theta \in \Theta)$ is differentiable in quadratic mean, then this sequence of processes converges (marginally) in distribution to the process

$$h \mapsto \log \frac{dN\big(h, I_\theta^{-1}\big)}{dN\big(0, I_\theta^{-1}\big)}(X) = -\frac{1}{2}(X-h)^T I_\theta (X-h) + \frac{1}{2}X^T I_\theta X.$$

If the sequence of sets H_n converges in a suitable sense to a set H, then we should expect, under regularity conditions, that the sequence $\hat{h}_n$ converges to the maximizer $\hat{h}$ of the latter process over H. This maximizer is the projection of the vector X onto the set H relative to the metric $d(x, y) = (x-y)^T I_\theta (x-y)$ (where a "projection" means a closest point); if $H = \mathbb{R}^k$, this projection reduces to X itself.

An appropriate notion of *convergence of sets* is the following. 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.[†]

7.12 ***Theorem.*** *Suppose that the experiment $(P_\theta : \theta \in \Theta)$ is differentiable in quadratic mean at θ_0 with nonsingular Fisher information matrix I_{θ_0}. Furthermore, suppose that for every θ_1 and θ_2 in a neighborhood of θ_0 and a measurable function $\dot{\ell}$ with $P_{\theta_0}\dot{\ell}^2 < \infty$,*

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

If the sequence of maximum likelihood estimators $\hat{\theta}_n$ is consistent and the sets $H_n = \sqrt{n}(\Theta - \theta_0)$ converge to a nonempty, convex set H, then the sequence $I_{\theta_0}^{1/2}\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges under θ_0 in distribution to the projection of a standard normal vector onto the set $I_{\theta_0}^{1/2}H$.

Proof. Let $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_{\theta_0})$ be the empirical process. In the proof of Theorem 5.39 it is shown that the map $\theta \mapsto \log p_\theta$ is differentiable at θ_0 in $L_2(P_{\theta_0})$ with derivative $\dot{\ell}_{\theta_0}$ and that the map $\theta \mapsto P_{\theta_0} \log p_\theta$ permits a Taylor expansion of order 2 at θ_0, with "second-derivative matrix" $-I_{\theta_0}$. Therefore, the conditions of Lemma 19.31 are satisfied for $m_\theta = \log p_\theta$, whence, for every M,

$$\sup_{\|h\| \le M} \left| n\mathbb{P}_n \log \frac{p_{\theta_0+h/\sqrt{n}}}{p_{\theta_0}} - h^T \mathbb{G}_n \dot{\ell}_{\theta_0} + \frac{1}{2}h^T I_{\theta_0} h \right| \xrightarrow{P} 0.$$

By Corollary 5.53 the estimators $\hat{\theta}_n$ are $\sqrt{n}$-consistent under θ_0.

The preceding display is also valid for every sequence M_n that diverges to ∞ sufficiently slowly. Fix such a sequence. By the $\sqrt{n}$-consistency of $\hat{\theta}_n$, the local maximum likelihood

[†] See Chapter 16 for examples.

estimators $\hat{h}_n$ are bounded in probability and hence belong to the balls of radius M_n with probability tending to 1. Furthermore, the sequence of intersections $H_n \cap$ ball $(0, M_n)$ converges to H, as the original sets H_n. Thus, we may assume that the $\hat{h}_n$ are the maximum likelihood estimators relative to local parameter sets H_n that are contained in the balls of radius M_n. Fix an arbitrary closed set F. If $\hat{h}_n \in F$, then the log likelihood is maximal on F. Hence $\mathrm{P}(\hat{h}_n \in F)$ is bounded above by

$$\begin{aligned}
&\mathrm{P}\left(\sup_{h \in F \cap H_n} \mathbb{P}_n \log \frac{p_{\theta_0 + h/\sqrt{n}}}{p_{\theta_0}} \geq \sup_{h \in H_n} \mathbb{P}_n \log \frac{p_{\theta_0 + h/\sqrt{n}}}{p_{\theta_0}} \right) \\
&\quad = \mathrm{P}\left(\sup_{h \in F \cap H_n} h^T \mathbb{G}_n \dot{\ell}_{\theta_0} - \frac{1}{2} h^T I_{\theta_0} h \geq \sup_{h \in H_n} h^T \mathbb{G}_n \dot{\ell}_{\theta_0} - \frac{1}{2} h^T I_{\theta_0} h + o_P(1) \right) \\
&\quad = \mathrm{P}\Big(\big\| I_{\theta_0}^{-1/2} \mathbb{G}_n \dot{\ell}_{\theta_0} - I_{\theta_0}^{1/2}(F \cap H_n) \big\| \leq \big\| I_{\theta_0}^{-1/2} \mathbb{G}_n \dot{\ell}_{\theta_0} - I_{\theta_0}^{1/2} H_n \big\| + o_P(1) \Big),
\end{aligned}$$

by completing the square. By Lemma 7.13 (ii) and (iii) ahead, we can replace H_n by H on both sides, at the cost of adding a further $o_P(1)$-term and increasing the probability. Next, by the continuous mapping theorem and the continuity of the map $z \mapsto \|z - A\|$ for every set A, the probability is asymptotically bounded above by, with Z a standard normal vector,

$$\mathrm{P}\Big(\big\| Z - I_{\theta_0}^{1/2}(F \cap H) \big\| \leq \big\| Z - I_{\theta_0}^{1/2} H \big\| \Big).$$

The projection ΠZ of the vector Z on the set $I_{\theta_0}^{1/2} H$ is unique, because the latter set is convex by assumption and automatically closed. If the distance of Z to $I_{\theta_0}^{1/2}(F \cap H)$ is smaller than its distance to the set $I_{\theta_0}^{1/2} H$, then ΠZ must be in $I_{\theta_0}^{1/2}(F \cap H)$. Consequently, the probability in the last display is bounded by $\mathrm{P}(\Pi Z \in I_{\theta_0}^{1/2} F)$. The theorem follows from the portmanteau lemma. ■

7.13 ***Lemma.*** *If the sequence of subsets H_n of $\mathbb{R}^k$ converges to a nonempty set H and the sequence of random vectors X_n converges in distribution to a random vector X, then*

(i) $\|X_n - H_n\| \rightsquigarrow \|X - H\|$.

(ii) $\|X_n - H_n \cap F\| \geq \|X_n - H \cap F\| + o_P(1)$, *for every closed set F.*

(iii) $\|X_n - H_n \cap G\| \leq \|X_n - H \cap G\| + o_P(1)$, *for every open set G.*

Proof. (i). Because the map $x \mapsto \|x - H\|$ is (Lipschitz) continuous for any set H, we have that $\|X_n - H\| \rightsquigarrow \|X - H\|$ by the continuous-mapping theorem. If we also show that $\|X_n - H_n\| - \|X_n - H\| \xrightarrow{P} 0$, then the proof is complete after an application of Slutsky's lemma. By the uniform tightness of the sequence X_n, it suffices to show that $\|x - H_n\| \to \|x - H\|$ uniformly for x ranging over compact sets, or equivalently that $\|x_n - H_n\| \to \|x - H\|$ for every converging sequence $x_n \to x$.

For every fixed vector x_n, there exists a vector $h_n \in H_n$ with $\|x_n - H_n\| \geq \|x_n - h_n\| - 1/n$. Unless $\|x_n - H_n\|$ is unbounded, we can choose the sequence h_n bounded. Then every subsequence of h_n has a further subsequence along which it converges, to a limit h in H. Conclude that, in any case,

$$\liminf \|x_n - H_n\| \geq \liminf \|x_n - h_n\| \geq \|x - h\| \geq \|x - H\|.$$

Conversely, for every $\varepsilon > 0$ there exists $h \in H$ and a sequence $h_n \to h$ with $h_n \in H_n$ and

$$\|x - H\| \geq \|x - h\| - \varepsilon = \lim \|x_n - h_n\| - \varepsilon \geq \limsup \|x_n - H_n\| - \varepsilon.$$

Combination of the last two displays yields the desired convergence of the sequence $\|x_n - H_n\|$ to $\|x - H\|$.

(ii). The assertion is equivalent to the statement $P\big(\|X_n - H_n \cap F\| - \|X_n - H \cap F\| > -\varepsilon\big) \to 1$ for every $\varepsilon > 0$. In view of the uniform tightness of the sequence X_n, this follows if $\liminf \|x_n - H_n \cap F\| \geq \|x - H \cap F\|$ for every converging sequence $x_n \to x$. We can prove this by the method of the first half of the proof of (i), replacing H_n by $H_n \cap F$.

(iii). Analogously to the situation under (ii), it suffices to prove that $\limsup \|x_n - H_n \cap G\| \leq \|x - H \cap G\|$ for every converging sequence $x_n \to x$. This follows as the second half of the proof of (i). ■

*7.5 Limit Distributions under Alternatives

Local asymptotic normality is a convenient tool in the study of the behavior of statistics under "contiguous alternatives." Under local asymptotic normality,

$$\log \frac{dP^n_{\theta+h/\sqrt{n}}}{dP^n_\theta} \overset{\theta}{\rightsquigarrow} N\left(-\frac{1}{2}h^T I_\theta h, h^T I_\theta h\right).$$

Therefore, in view of Example 6.5 the sequences of distributions $P^n_{\theta+h/\sqrt{n}}$ and P^n_θ are mutually contiguous. This is of great use in many proofs. With the help of Le Cam's third lemma it also allows to obtain limit distributions of statistics under the parameters $\theta + h/\sqrt{n}$, once the limit behavior under θ is known. Such limit distributions are of interest, for instance, in studying the asymptotic efficiency of estimators or tests.

The general scheme is as follows. Many sequences of statistics T_n allow an approximation by an average of the type

$$\sqrt{n}(T_n - \mu_\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_\theta(X_i) + o_{P_\theta}(1).$$

According to Theorem 7.2, the sequence of log likelihood ratios can be approximated by an average as well: It is asymptotically equivalent to an affine transformation of $n^{-1/2} \sum \dot{\ell}_\theta(X_i)$. The sequence of joint averages $n^{-1/2} \sum \big(\psi_\theta(X_i), \dot{\ell}_\theta(X_i)\big)$ is asymptotically multivariate normal under θ by the central limit theorem (provided ψ_θ has mean zero and finite second moment). With the help of Slutsky's lemma we obtain the joint limit distribution of T_n and the log likelihood ratios under θ:

$$\left(\sqrt{n}(T_n - \mu_\theta), \log \frac{dP^n_{\theta+h/\sqrt{n}}}{dP^n_\theta}\right) \overset{\theta}{\rightsquigarrow} N\left(\begin{pmatrix} 0 \\ -\frac{1}{2}h^T I_\theta h \end{pmatrix}, \begin{pmatrix} P_\theta \psi_\theta \psi_\theta^T & P_\theta \psi_\theta h^T \dot{\ell}_\theta \\ P_\theta \psi_\theta^T h^T \dot{\ell}_\theta & h^T I_\theta h \end{pmatrix}\right).$$

Finally we can apply Le Cam's third Example 6.7 to obtain the limit distribution of $\sqrt{n}(T_n - \mu_\theta)$ under $\theta + h/\sqrt{n}$. Concrete examples of this scheme are discussed in later chapters.

*7.6 Local Asymptotic Normality

The preceding sections of this chapter are restricted to the case of independent, identically distributed observations. However, the general ideas have a much wider applicability. A

wide variety of models satisfy a general form of local asymptotic normality and for that reason allow a unified treatment. These include models with independent, not identically distributed observations, but also models with dependent observations, such as used in time series analysis or certain random fields. Because local asymptotic normality underlies a large part of asymptotic optimality theory and also explains the asymptotic normality of certain estimators, such as maximum likelihood estimators, it is worthwhile to formulate a general concept.

Suppose the observation at "time" n is distributed according to a probability measure $P_{n,\theta}$, for a parameter θ ranging over an open subset Θ of $\mathbb{R}^k$.

7.14 Definition. The sequence of statistical models $(P_{n,\theta} : \theta \in \Theta)$ is *locally asymptotically normal* (*LAN*) at θ if there exist matrices r_n and I_θ and random vectors $\Delta_{n,\theta}$ such that $\Delta_{n,\theta} \overset{\theta}{\rightsquigarrow} N(0, I_\theta)$ and for every converging sequence $h_n \to h$

$$\log \frac{dP_{n,\theta+r_n^{-1}h_n}}{dP_{n,\theta}} = h^T \Delta_{n,\theta} - \frac{1}{2} h^T I_\theta h + o_{P_{n,\theta}}(1).$$

7.15 ***Example.*** If the experiment $(P_\theta : \theta \in \Theta)$ is differentiable in quadratic mean, then the sequence of models $(P_\theta^n : \theta \in \Theta)$ is locally asymptotically normal with norming matrices $r_n = \sqrt{n} I$. □

An inspection of the proof of Theorem 7.10 readily reveals that this depends on the local asymptotic normality property only. Thus, the local experiments

$$\left(P_{n,\theta+r_n^{-1}h} : h \in \mathbb{R}^k\right)$$

of a locally asymptotically normal sequence converge to the experiment $\left(N(h, I_\theta^{-1}) : h \in \mathbb{R}^k\right)$, in the sense of this theorem. All results for the case of i.i.d. observations that are based on this approximation extend to general locally asymptotically normal models. To illustrate the wide range of applications we include, without proof, three examples, two of which involve dependent observations.

7.16 ***Example (Autoregressive processes).*** An autoregressive process $\{X_t : t \in \mathbb{Z}\}$ of order 1 satisfies the relationship $X_t = \theta X_{t-1} + Z_t$ for a sequence of independent, identically distributed variables $\ldots, Z_{-1}, Z_0, Z_1, \ldots$ with mean zero and finite variance. There exists a stationary solution $\ldots, X_{-1}, X_0, X_1, \ldots$ to the autoregressive equation if and only if $|\theta| \neq 1$. To identify the parameter it is usually assumed that $|\theta| < 1$. If the density of the noise variables Z_j has finite Fisher information for location, then the sequence of models corresponding to observing $X_1, \ldots, X_n$ with parameter set $(-1, 1)$ is locally asymptotically normal at θ with norming matrices $r_n = \sqrt{n} I$.

The observations in this model form a stationary Markov chain. The result extends to general ergodic Markov chains with smooth transition densities (see [130]). □

7.17 ***Example (Gaussian time series).*** This example requires some knowledge of time-series models. Suppose that at time n the observations are a stretch $X_1, \ldots, X_n$ from a stationary, Gaussian time series $\{X_t : t \in \mathbb{Z}\}$ with mean zero. The covariance matrix of n

consecutive variables is given by the (Toeplitz) matrix

$$T_n(f_\theta) = \left(\int_{-\pi}^{\pi} e^{i(t-s)\lambda} \, f_\theta(\lambda) \, d\lambda \right)_{s,t=1,\ldots,n}.$$

The function f_θ is the *spectral density* of the series. It is convenient to let the parameter enter the model through the spectral density, rather than directly through the density of the observations.

Let $P_{n,\theta}$ be the distribution (on $\mathbb{R}^n$) of the vector $(X_1, \ldots, X_n)$, a normal distribution with mean zero and covariance matrix $T_n(f_\theta)$. The *periodogram* of the observations is the function

$$I_n(\lambda) = \frac{1}{2\pi n} \left| \sum_{t=1}^{n} X_t e^{it\lambda} \right|^2.$$

Suppose that f_θ is bounded away from zero and infinity, and that there exists a vector-valued function $\dot{\ell}_\theta : \mathbb{R} \mapsto \mathbb{R}^d$ such that, as $h \to 0$,

$$\int [f_{\theta+h} - f_\theta - h^T \dot{\ell}_\theta \, f_\theta]^2 \, d\lambda = o\big(\|h\|^2\big).$$

Then the sequence of experiments $(P_{n,\theta} : \theta \in \Theta)$ is locally asymptotically normal at θ with

$$r_n = \sqrt{n}, \qquad \Delta_{n,\theta} = \frac{\sqrt{n}}{4\pi} \int (I_n - \mathrm{E}_\theta I_n) \frac{\dot{\ell}_\theta}{f_\theta} \, d\lambda, \qquad I_\theta = \frac{1}{4\pi} \int \dot{\ell}_\theta \dot{\ell}_\theta^T \, d\lambda.$$

The proof is elementary, but involved, because it has to deal with the quadratic forms in the n-variate normal density, which involve vectors whose dimension converges to infinity (see [30]). □

7.18 ***Example (Almost regular densities).*** Consider estimating a location parameter θ based on a sample of size n from the density $f(x - \theta)$. If f is smooth, then this model is differentiable in quadratic mean and hence locally asymptotically normal by Example 7.8. If f possesses points of discontinuity, or other strong irregularities, then a locally asymptotically normal approximation is impossible.† Examples of densities that are on the boundary between these "extremes" are the triangular density $f(x) = \big(1 - |x|\big)^+$ and the gamma density $f(x) = xe^{-x}1\{x > 0\}$. These yield models that are locally asymptotically normal, but with norming rate $\sqrt{n \log n}$ rather than $\sqrt{n}$. The existence of singularities in the density makes the estimation of the parameter θ easier, and hence a faster rescaling rate is necessary. (For the triangular density, the true singularities are the points -1 and 1, the singularity at 0 is statistically unimportant, as in the case of the Laplace density.)

For a more general result, consider densities f that are absolutely continuous except possibly in small neighborhoods $U_1, \ldots, U_k$ of finitely many fixed points $c_1, \ldots, c_k$. Suppose that $f'/\sqrt{f}$ is square-integrable on the complement of $\cup_j U_j$, that $f(c_j) = 0$ for every j, and that, for fixed constants $a_1, \ldots, a_k$ and $b_1, \ldots, b_k$, each of the functions

$$x \mapsto f(x) - \big(a_j 1\{x < c_j\} + b_j 1\{x > c_j\}\big)|x - c_j|, \qquad x \in U_j,$$

† See Chapter 9 for some examples.

is twice continuously differentiable. If $\sum(a_j + b_j) > 0$, then the model is locally asymptotically normal at $\theta = 0$ with, for V_n equal to the interval $\left(n^{-1/2}(\log n)^{-1/4}, (\log n)^{-1}\right)$ around zero,†

$$r_n = \sqrt{n \log n}, \qquad I_0 = \sum_j (a_j + b_j),$$

$$\Delta_{n,0} = \frac{1}{\sqrt{n \log n}} \sum_{i=1}^{n} \sum_{j=1}^{k} \left(\frac{1\{X_i - c_j \in V_n\}}{X_i - c_j} - \int_{V_n} \frac{1}{x} f(x + c_j)\, dx \right).$$

The sequence $\Delta_{n,0}$ may be thought of as "asymptotically sufficient" for the local parameter h. Its definition of $\Delta_{n,0}$ shows that, asymptotically, all the "information" about the parameter is contained in the observations falling into the neighborhoods $V_n + c_j$. Thus, asymptotically, the problem is determined by the points of irregularity.

The remarkable rescaling rate $\sqrt{n \log n}$ can be explained by computing the Hellinger distance between the densities $f(x - \theta)$ and $f(x)$ (see section 14.5). □

Notes

Local asymptotic normality was introduced by Le Cam [92], apparently motivated by the study and construction of asymptotically similar tests. In this paper Le Cam defines two sequences of models $(P_{n,\theta} : \theta \in \Theta)$ and $(Q_{n,\theta} : \theta \in \Theta)$ to be *differentially equivalent* if

$$\sup_{h \in K} \| P_{n,\theta + h/\sqrt{n}} - Q_{n,\theta + h/\sqrt{n}} \| \to 0,$$

for every bounded set K and every θ. He next shows that a sequence of statistics T_n in a given *asymptotically differentiable* sequence of experiments (roughly LAN) that is asymptotically equivalent to the centering sequence $\Delta_{n,\theta}$ is asymptotically sufficient, in the sense that the original experiments and the experiments consisting of observing the T_n are differentially equivalent. After some interpretation this gives roughly the same message as Theorem 7.10. The latter is a concrete example of an abstract result in [95], with a different (direct) proof.

PROBLEMS

1. Show that the Poisson distribution with mean θ satisfies the conditions of Lemma 7.6. Find the information.
2. Find the Fisher information for location for the normal, logistic, and Laplace distributions.
3. Find the Fisher information for location for the Cauchy distributions.
4. Let f be a density that is symmetric about zero. Show that the Fisher information matrix (if it exists) of the location scale family $f\big((x - \mu)/\sigma\big)/\sigma$ is diagonal.
5. Find an explicit expression for the $o_{P_\theta}(1)$-term in Theorem 7.2 in the case that p_θ is the density of the $N(\theta, 1)$-distribution.
6. Show that the Laplace location family is differentiable in quadratic mean.

† See, for example, [80, pp. 133–139] for a proof, and also a discussion of other almost regular situations. For instance, singularities of the form $f(x) \sim f(c_j) + |x - c_j|^{1/2}$ at points c_j with $f(c_j) > 0$.

7. Find the form of the score function for a location-scale family $f\big((x-\mu)/\sigma\big)/\sigma$ with parameter $\theta = (\mu, \sigma)$ and apply Lemma 7.6 to find a sufficient condition for differentiability in quadratic mean.

8. Investigate for which parameters k the location family $f(x-\theta)$ for f the gamma$(k, 1)$ density is differentiable in quadratic mean.

9. Let $P_{n,\theta}$ be the distribution of the vector $(X_1, \ldots, X_n)$ if $\{X_t : t \in \mathbb{Z}\}$ is a stationary Gaussian time series satisfying $X_t = \theta X_{t-1} + Z_t$ for a given number $|\theta| < 1$ and independent standard normal variables Z_t. Show that the model is locally asymptotically normal.

10. Investigate whether the log normal family of distributions with density

$$\frac{1}{\sigma\sqrt{2\pi}(x-\xi)} e^{-\frac{1}{2\sigma^2}(\log(x-\xi)-\mu)^2} 1\{x > \xi\}$$

is differentiable in quadratic mean with respect to $\theta = (\xi, \mu, \sigma)$.