Stat 8153 Lecture Notes

Charles J. Geyer

Draft: June 22, 2001

Contents

Chapter 1

Asymptotics for Exponential Families

1.1 Convex Functions

The *extended real number* system consists of the real numbers and two additional "numbers" $-\infty$ and $+\infty$. As sets, the real numbers are denoted $\mathbb{R}$ and the extended real numbers are denoted $\overline{\mathbb{R}}$.

It turns out to be very useful to define convex functions taking values in $\overline{\mathbb{R}}$. These are called extended-real-valued convex functions. Of course we are mostly interested in their behavior where they are real-valued, but allowing the values $+\infty$ and $-\infty$ turns out to be a great convenience.

For any function $f : S \to \overline{\mathbb{R}}$, where S is any set,

$$\operatorname{dom} f = \{\, x \in S : f(x) < +\infty \,\}$$

is called the *effective domain* of f. Such a function is said to be *proper* if its effective domain is nonempty and it is real-valued on its effective domain, or, what is equivalent, f is proper if $-\infty < f(x)$ for all x and $f(x) < +\infty$ for at least one x.

Note that these two definitions (of "effective domain" and "proper") treat plus and minus infinity very differently. The reason is that the theory of convex functions finds most of its applications in minimization problems (that is, the object is to minimize a convex function) and they too treat plus and minus infinity very differently.

A subset S of $\mathbb{R}^d$ is *convex* if

$$sx + (1 - s)y \in S, \qquad x, y \in S \text{ and } 0 < s < 1.$$

Note that it would make no difference if the definition were changed by replacing $0 < s < 1$ with $0 \le s \le 1$.

An extended-real-valued function f on $\mathbb{R}^d$ is *convex* if

$$f\big(sx + (1-s)y\big) \le sf(x) + (1-s)f(y), \qquad x, y \in \operatorname{dom} f \text{ and } 0 < s < 1$$

The inequality in this formula is also known as the *convexity inequality*. Note that on the right hand side of the convexity inequality neither term can be $+\infty$ because of the requirement $x, y \in \operatorname{dom} f$. Thus there is no need to define what we mean by $\infty - \infty$ in order to use this definition. Similarly since we are requiring $0 < s < 1$ there is no need to define what we mean by $0 \cdot \infty$. All we need are the obvious rules of arithmetic $s \cdot (-\infty) = -\infty$ when $0 < s < \infty$ and $-\infty + x = -\infty$ when $x < +\infty$. Together they imply that the right hand side of the convexity inequality is $-\infty$ whenever either $f(x)$ or $f(y)$ is $-\infty$.

1.2 Concave Functions

An extended-real-valued function f is said to be *concave* if $-f$ is convex. For concave functions we change the definitions of "effective domain" and "proper." Instead of applying the original definitions to f, instead we say that the effective domain of a concave function f is the same as the effective domain of the convex function $-f$ and, similarly, that a concave function f is proper if and only if the convex function $-f$ is proper. The reason is that the theory of concave functions finds most of its applications in maximization problems (that is, the object is to maximize a concave function).

In fact, we only need one theory. If interested in maximizing a concave function, stand on your head and you are minimizing a convex function. The difference in the two situations is entirely trivial, a mere change in terminology and notation. Nothing of mathematical significance changes, which is why we want our definitions of "effective domain" and "proper" to match. Why take convex functions as the main notion and treat concave functions as the secondary notion that cannot be properly understood without reference to the other? Tradition and the way most optimization books are written.

1.3 Exponential Families

We use the notation $\langle \cdot, \cdot \rangle$ for the usual inner product on $\mathbb{R}^d$, that is, if $x = (x_1, \ldots, x_d)$ and $\theta = (\theta_1, \ldots, \theta_d)$, then

$$\langle x, \theta \rangle = \sum_{i=1}^{d} x_i \theta_i$$

or in the notation where x and θ are both considered "column vectors" ($d \times 1$ matrices) $\langle x, \theta \rangle = x'\theta$, where the prime indicates the transpose operation.

Let λ be a nonzero Borel measure on $\mathbb{R}^d$. The *Laplace transform* of λ is the extended-real-valued function c on $\mathbb{R}^d$ defined by

$$c(\theta) = \int e^{\langle x, \theta \rangle} \lambda(dx).$$

For any $\theta \in \operatorname{dom} c$, the real-valued function f_θ on $\mathbb{R}^d$ defined by

$$f_\theta(x) = \frac{1}{c(\theta)} e^{\langle x, \theta \rangle} \tag{1.1}$$

is a probability density with respect to λ. The family $\{ f_\theta : \theta \in \operatorname{dom} c \}$ is the *full standard exponential family* of densities *generated* by λ. Any subset of this family is called a *standard exponential family*. The measure λ is sometimes called the *base measure* of the family.

A *general* exponential family is the family of densities of a random variable X taking values in any space but having a statistic $t(X)$ that induces a standard exponential family, which means that $t(X)$ takes values in $\mathbb{R}^d$ for some d. Then $t(X)$ is called the *natural statistic* (also called *canonical statistic*) of the family. This generalization makes only a slight change in notation. The densities become

$$f_\theta(x) = \frac{1}{c(\theta)} e^{\langle t(x), \theta \rangle}, \tag{1.2}$$

where

$$c(\theta) = \int e^{\langle t(x), \theta \rangle} \lambda(dx) \tag{1.3}$$

and now λ is a measure on the space where X takes values. A change in descriptive terminology is now also required. The function c is no longer the Laplace transform of λ (technically it is the Laplace transform of the image measure $\lambda \circ t^{-1}$).

The parameterization used so far is called the *natural parameterization* of the family (also called *canonical parameterization*), and θ is called the *natural* (or *canonical*) parameter. If we introduce another parameterization, say $\theta = g(\varphi)$, where g is a one-to-one mapping from some other parameter space into $\mathbb{R}^d$, the formula for the densities becomes

$$f_\varphi(x) = \frac{1}{c\big(g(\varphi)\big)} e^{\langle t(x), g(\varphi) \rangle},$$

where c is still given by (1.3).

General exponential families clutter the theory of exponential families. Since the densities (1.2) only depend on x through $t(x)$, it follows by the factorization criterion that $t(X)$ is a sufficient statistic. Hence the sufficiency principle says we may use the standard exponential family induced by $t(X)$ for inference. Moreover, all the theory is much cleaner and clearer when stated in terms of standard families. Of course, you are usually given an exponential family problem in nonstandard form, and you must start by recognizing the natural statistic and parameter, but after you have made this recognition everything of interest about the problem can be derived from the theory of standard exponential families.

1.4 Identifiability

A family of probability densities $\{ f_\theta : \theta \in \Theta \}$ with respect to a measure λ is *identifiable* if there do not exist distinct densities corresponding to the same probability measure, that is, if there do not exist $\theta_1 \neq \theta_2$ such that f_{θ_1} and f_{θ_2} are equal almost everywhere with respect to λ.

A *hyperplane* in $\mathbb{R}^d$ is a subset of the form

$$H = \{ x \in \mathbb{R}^d : \langle x, \varphi \rangle = b \} \tag{1.4}$$

for some nonzero $\varphi \in \mathbb{R}^d$ and some $b \in \mathbb{R}$.

Lemma 1.1. *A full standard exponential family is identifiable if and only if its base measure is not concentrated on a hyperplane.*

Proof. Because the log function is one-to-one, the densities f_{θ_1} and f_{θ_2} are equal almost everywhere if and only if the log densities are equal almost everywhere, and the log densities are equal for x such that

$$\langle x, \theta_2 - \theta_1 \rangle = \log \frac{c(\theta_2)}{c(\theta_1)} \tag{1.5}$$

Since $\theta_1 \neq \theta_2$ the set of x such that (1.5) holds is a hyperplane. Hence if the base measure is not concentrated on a hyperplane, distinct densities cannot be equal almost everywhere.

Conversely, if the base measure is concentrated on the hyperplane (1.4) and we denote the base measure by λ and its Laplace transform by c as before and θ is any point in $\operatorname{dom} c$, then

$$\begin{aligned} c(\theta + \varphi) &= \int e^{\langle x, \theta + \varphi \rangle} \lambda(dx) \\ &= \int_H e^{\langle x, \theta + \varphi \rangle} \lambda(dx) \\ &= e^b \int_H e^{\langle x, \theta \rangle} \lambda(dx) \\ &= e^b c(\theta) \end{aligned}$$

because $\langle x, \varphi \rangle = b$ on H. Hence $\theta + \varphi \in \operatorname{dom} c$ and

$$\begin{aligned} f_{\theta+\varphi}(x) &= \frac{1}{c(\theta + \varphi)} e^{\langle x, \theta + \varphi \rangle} \\ &= \frac{1}{c(\theta)} e^{\langle x, \theta \rangle} \cdot \frac{1}{e^b} e^{\langle x, \varphi \rangle} \\ &= f_\theta(x) \cdot e^{\langle x, \varphi \rangle - b} \end{aligned}$$

and this equals $f_\theta(x)$ for $x \in H$. So the family is not identifiable. □

1.5 Affine Change of Parameter

A function $g : \mathbb{R}^p \to \mathbb{R}^d$ is *affine* if it has the form

$$g(\varphi) = A\varphi + b$$

for some linear transformation $A : \mathbb{R}^p \to \mathbb{R}^d$ and some $b \in \mathbb{R}^d$. We now want to consider such an affine change of parameter for a standard exponential family. Then writing $\theta = g(\varphi)$ we have

$$\langle x, \theta \rangle = \langle x, A\varphi \rangle + \langle x, b \rangle = \langle A^* x, \varphi \rangle + \langle x, b \rangle$$

where $A^* : \mathbb{R}^d \to \mathbb{R}^p$ is the *adjoint* of A, which is characterized by

$$\langle x, A\varphi \rangle = \langle A^* x, \varphi \rangle, \qquad x \in \mathbb{R}^d, \ \varphi \in \mathbb{R}^p.$$

If we write this in matrix notation it becomes $x'A\varphi = \varphi' Bx$, where B is the matrix representing the linear transformation A^*. Using the rule for the transpose of a product $x'A\varphi = \varphi' A' x$, we see that $B = A'$ so the adjoint operation on linear transformations corresponds to the transpose operation on matrices.

Thus in the new parameterization, the densities are

$$f_\varphi(x) = \frac{1}{c(A\varphi + b)} e^{\langle A^* x, \varphi \rangle} e^{\langle x, b \rangle}$$

From the functional form we see that we again have a standard exponential family with natural statistic A^*X, natural parameter φ, and a new base measure having density $e^{\langle x, b \rangle}$ with respect to the old base measure.

Thus an affine change of parameter just gives us a new standard exponential family. It does not take us into any new realm of theory. It is important in applications that we allowed the domain and range of the affine change of variable to have different dimensions. Most of linear regression and generalized linear model theory is a special case of this section. We observe data Y_1, ..., Y_n that are independent and all have distributions in the same one-parameter exponential family but are not identically distributed because they have different values of the natural parameter: Y_i having the distribution corresponding to natural parameter θ_i. The joint density is

$$f_\theta(y) = \frac{1}{\prod_i c(\theta_i)} e^{\sum_i y_i \theta_i}$$

which is clearly of the exponential family form when we consider y and θ as vectors, since $\sum_i y_i \theta_i = \langle y, \theta \rangle$. A regression model involves other data, called *covariates*, which may be either random or deterministic, but are treated as being deterministic. If the covariates are random, then the inference is conditional on the observed covariate values. The regression model specifies a new parameterization

$$\theta = A_X \varphi + b_X$$

where A_X is a linear transformation and b_X a vector, both considered deterministic because X is treated as deterministic. This is just an affine change of parameter as discussed above.

1.6 Independent, Identically Distributed Data

Another important application that does not take us out of standard exponential families is independent and identically distributed data. If X_1, ..., X_n are i. i. d. from the distribution having densities (1.1), then the joint density

$$f_\theta(x_1, \ldots, x_n) = \frac{1}{c(\theta)^n} e^{\langle \sum_i x_i, \theta \rangle}$$

(this is a density with respect to the product measure λ^n). If we do a one-to-one linear change of variables from $(x_1, \ldots, x_n)$ to $(u_1, \ldots, u_n)$ so that $u_n = \sum_i x_i$, then the joint density of the new variables will be

$$f_\theta(u_1, \ldots, u_n) = \frac{j}{c(\theta)^n} e^{\langle u_n, \theta \rangle}$$

where j is a constant (the jacobian). If we integrate out all the variables but u_n we get the marginal density of $U = \sum_i X_i$

$$f_\theta(u) = \frac{j_2}{c(\theta)^n} e^{\langle u, \theta \rangle} \tag{1.6}$$

where j_2 is another constant. It is now not so clear what measure this is a density with respect to, but a little thought reveals that it must be the n-fold convolution of the original base measure λ.

Thus, here too, we have not entered any new realm of theory. Independent and identically distributed sampling just gives another standard exponential family with the same natural parameter as the original family and natural statistic $\sum_i X_i$ if X_1, ..., X_n were the original natural statistics. Thus there is, for the most part, no need to explicitly mention i. i. d. sampling. It is just a special case of the general notion of a standard exponential family. If we understand the general case, i. i. d. sampling adds nothing new.

1.7 Moment Generating Functions

The *moment generating function* of the random variable X having density (1.1) is the function M_θ on $\mathbb{R}^d$ defined by

$$M_\theta(t) = E_\theta e^{\langle X, t \rangle} = \frac{c(\theta + t)}{c(\theta)}$$

if M_θ is finite on a neighborhood of zero, that is, if $\theta \in \operatorname{int}(\operatorname{dom} c)$. If θ is not in the interior of $\operatorname{dom} c$, then M_θ is not a moment generating function in the classical sense.

If $\theta \in \operatorname{int}(\operatorname{dom} c)$, then the first two absolute moments are given by the

derivatives of M_θ at zero.

$$E_\theta(X) = \nabla M_\theta(0) = \frac{\nabla c(\theta)}{c(\theta)}$$
$$E_\theta(XX') = \nabla^2 M_\theta(0) = \frac{\nabla^2 c(\theta)}{c(\theta)}$$

If we define the *cumulant function* $k = \log c$, we find that

$$\nabla k(\theta) = \frac{\nabla c(\theta)}{c(\theta)} = E_\theta(X)$$

and

$$\begin{aligned}\nabla^2 k(\theta) &= \frac{\nabla^2 c(\theta)}{c(\theta)} - \frac{\nabla c(\theta)[\nabla c(\theta)]'}{c(\theta)^2} \\ &= E_\theta(XX') - E_\theta(X)E_\theta(X)' \\ &= \mathrm{Var}_\theta(X)\end{aligned}$$

1.8 Maximum Likelihood

The log likelihood of the standard exponential family with densities (1.1) is

$$l_x(\theta) = \langle x, \theta\rangle - k(\theta)$$

where k is the cumulant function (log Laplace transform). It can be shown that k is a convex function (Problem 1-2) so the log likelihood is a concave function.

A point x is a *global minimizer* of an extended-real-valued function f on $\mathbb{R}^d$ if

$$f(y) \ge f(x), \qquad y \in \mathbb{R}^d,$$

and x is a *local minimizer* if there exists a neighborhood U of x such that

$$f(y) \ge f(x), \qquad y \in U.$$

Lemma 1.2. *For an extended-real-valued proper convex function, every local minimizer is a global minimizer.*

The proof is left as an exercise (Problem 1-3).

We won't even bother to state as a formal corollary the analogous fact for concave functions, in particular for log likelihoods of standard exponential families: every local maximizer is a global maximizer. This is just obvious from the "stand on your head" principle.

Lemma 1.3. *In an identifiable full standard exponential family, the maximum likelihood estimate is unique if it exists.*

Proof. If θ_1 and θ_2 are distinct maximizers of the log likelihood, say $l_x(\theta_1) = l_x(\theta_2) = m$, then concavity of the log likelihood implies

$$l_x(s\theta_1 + (1-s)\theta_2) \geq s l_x(\theta_1) + (1-s) l_x(\theta_2) = m \tag{1.7}$$

for $0 \leq s \leq 1$. But since the value at a global maximizer cannot be exceeded, we must have equality in (1.7) for $0 \leq s \leq 1$, that is,

$$k(s\theta_1 + (1-s)\theta_2) = \langle x, s\theta_1 + (1-s)\theta_2 \rangle - m.$$

Hence

$$\frac{d^2}{ds^s} k(s\theta_1 + (1-s)\theta_2) = 0, \qquad 0 < s < 1.$$

The map $s \mapsto s\theta_1 + (1-s)\theta_2$ is an affine change of parameter, hence by the theory of Section 1.5 it induces a one-parameter exponential family with natural parameter s, natural statistic $\langle X, \theta_1 - \theta_2 \rangle$, and cumulant function $s \mapsto k(s\theta_1 + (1-s)\theta_2)$. By the theory of moment generating functions, this cumulant function having second derivative zero means that the natural statistic $\langle X, \theta_1 - \theta_2 \rangle$ has variance zero, which means that X itself is concentrated on a hyperplane and the original family is not identifiable, contrary to hypothesis. Thus we have reached a contradiction and hence the assumption that distinct maximizers exist cannot be possible. □

If the log likelihood is maximized at an interior point of the effective domain, then its gradient is zero. That is, the MLE is the unique solution (unique by Lemma 1.3) of

$$\nabla l_x(\theta) = x - \nabla k(\theta) = 0,$$

that is, the θ such that

$$x = \tau(\theta) = \nabla k(\theta).$$

A function $f : U \to V$ where U and V are open sets in $\mathbb{R}^d$ is a C^1 *isomorphism* if it is a continuously differentiable function and has a continuously differentiable inverse.

Lemma 1.4. *If k is the cumulant function of an identifiable full standard exponential family then the function $\tau : \operatorname{int}(\operatorname{dom} c) \to \mathbb{R}^d$ defined by*

$$\tau(\theta) = \nabla k(\theta) = E_\theta(X)$$

is a C^1 isomorphism onto its range, which is an open set, and

$$\nabla \tau(\theta) = \nabla^2 k(\theta) = \operatorname{Var}_\theta(X) \tag{1.8a}$$

and if $x = \tau(\theta)$

$$\nabla \tau^{-1}(x) = \left(\nabla \tau(\theta)\right)^{-1}. \tag{1.8b}$$

Proof. Since τ is one-to-one by Lemma 1.3 and is onto by definition (every function maps onto its range), it is invertible. Also (1.8a) holds by the theory of moment generating functions, so τ is differentiable, and the derivative is nonsingular because X is not concentrated on a hyperplane. Thus by the inverse function theorem, if $x = \tau(\theta)$ then τ has an inverse defined on some neighborhood of x such that (1.8b) holds. Since x was any point of the range of τ, this implies the range of τ is open. □

Let W denote the range of τ. We would now like to define the MLE to be $\tau^{-1}(x)$, and that is fine for $x \in W$, but the MLE is undefined when $x \notin W$. Thus we extend the domain of definition by adding a point u (for "undefined") to the set of possible values, which we consider an isolated point of the range. One way to do this and still remain in a Euclidean space is to embed W in $\mathbb{R}^{d+1}$ as the set

$$\widetilde{W} = \{\, (x_1, \ldots, x_d, 0) : (x_1, \ldots, x_d) \in W \,\}$$

and defining

$$u = (0, \ldots, 0, 1)$$

so the distance between u and x is greater than or equal to one for any $x \in \widetilde{W}$. This is an artificial device, but any method of precisely defining u will have similar artifice. To simplify the notation, consider W and $\widetilde{W}$ the "same" set and drop the tilde. Now we define

$$\hat{\theta}(x) = \begin{cases} \tau^{-1}(x), & x \in \tau(\operatorname{int}(\operatorname{dom} k)) \\ u, & \text{otherwise} \end{cases}$$

Note that this function is measurable, because it is continuous on the open set $\operatorname{int}(\operatorname{dom} k)$ and constant (equal to u) on the complement of $\operatorname{int}(\operatorname{dom} k)$.

Now we consider i. i. d. data x_1, x_2, ... The log likelihood is

$$l_{x_1,\ldots,x_n}(\theta) = \sum_{i=1}^{n} \langle x_i, \theta \rangle - nk(\theta) = n\langle \bar{x}_n, \theta \rangle - nk(\theta) = n l_{\bar{x}_n}(\theta)$$

Thus we get the same likelihood problem as if we had observed data $\bar{x}_n$ as a sample of size one. Then the MLE is

$$\hat{\theta}(\bar{x}_n) = \tau^{-1}(\bar{x}_n).$$

Theorem 1.5. *For a standard exponential family with cumulant function k, if the true parameter value θ_0 is in the interior of the effective domain of k and the family is identifiable, then the maximum likelihood estimate $\hat{\theta}_n$ exists and is unique with probability converging to one and*

$$\sqrt{n}\big(\hat{\theta}_n - \theta_0\big) \xrightarrow{\mathcal{L}} \text{Normal}\big(0, I(\theta_0)^{-1}\big)$$

where $I(\theta) = \nabla^2 k(\theta)$.

Proof. Since θ_0 is in $\operatorname{int}(\operatorname{dom} c)$, moments of all orders exist, in particular,

$$\begin{aligned} \mu_0 &= E_{\theta_0}(X) = \tau(\theta) = \nabla k(\theta_0) \\ I(\theta_0) &= \operatorname{Var}_{\theta_0}(X) = \nabla \tau(\theta) = \nabla^2 k(\theta_0) \end{aligned}$$

where τ is the function defined above. Then the function $\hat{\theta}$ defined above is differentiable at μ_0 by Lemma 1.4 and has derivative

$$\nabla \hat{\theta}(\mu_0) = \nabla \tau^{-1}(\mu_0) = I(\theta_0)^{-1}.$$

By the multivariate central limit theorem

$$\sqrt{n} \left(\overline{X}_n - \mu_0 \right) \xrightarrow{\mathcal{L}} \text{Normal}\left(0, I(\theta_0)\right)$$

Thus the multivariate delta method implies

$$\sqrt{n} \left[\hat{\theta}(\overline{X}_n) - \hat{\theta}(\mu_0) \right] \xrightarrow{\mathcal{L}} I(\theta_0)^{-1} Z$$

where $Z \sim \mathcal{N}\left(0, I(\theta_0)\right)$. Since

$$\operatorname{Var}\{I(\theta_0)^{-1} Z\} = I(\theta_0)^{-1} I(\theta_0) I(\theta_0)^{-1} = I(\theta_0)^{-1}$$

were are done, the assertion about the probability of the MLE existing converging to one following from the portmanteau theorem. □

Problems

1-1. Show that the effective domain of a convex function is a convex set.

1-2. Show the following properties of Laplace transforms. If c is the Laplace transform of a nonzero Borel measure λ on $\mathbb{R}^d$, and $k = \log c$ is the corresponding cumulant function, then

(a) $c(\theta) > 0$ for all θ (hence k is a proper extended-real-valued function).

(b) c and k are both convex functions.

(c) c and k are both lower semicontinuous functions.

(d) if $\operatorname{dom} c$ is nonempty, then λ is a σ-finite measure.

For part (c) the following characterization of lower semicontinuity is useful. An extended real valued function f is *lower semicontinuous* at a point x if for any sequence $x_n \to x$

$$\liminf_{n \to \infty} f(x_n) \geq f(x).$$

1-3. Prove Lemma 1.2.

Chapter 2

Asymptotic Efficiency

2.1 Log Likelihood Derivatives

Let $\mathcal{F} = \{ f_\theta : \theta \in \Theta \}$ be a family of probability densities with respect to a measure λ. Suppose that the parameter space Θ is an open subset of $\mathbb{R}^d$. The log likelihood is $l(\theta) = \log f_\theta$. It is a well-known fact that, if the identity $\int f_\theta \, d\lambda = 1$ is twice differentiable with respect to θ and both derivatives can be passed under the integral sign, then

$$E_\theta \nabla l(\theta) = 0 \tag{2.1a}$$

and

$$\operatorname{Var}_\theta \nabla l(\theta) = -E_\theta \nabla^2 l(\theta) \tag{2.1b}$$

hold. Either side of (2.1b) is called the *Fisher information* for the parameter θ, denoted $I(\theta)$. These two identities and the higher-order identities obtained by repeated differentiation under the integral sign are called the *Bartlett identities*.

For an i. i. d. sample $X_1, \ldots, X_n$ we denote the log likelihood by

$$l_n(\theta) = \sum_{i=1}^{n} \log f_\theta(X_i) \tag{2.2}$$

Let $I_n(\theta)$ be the Fisher information for the sample of size n

$$I_n(\theta) = \operatorname{Var}_\theta \nabla l_n(\theta),$$

then $I_n(\theta) = nI_1(\theta)$, because the variance of a sum is the sum of the variance for independent random variables.

Since l_n and its derivatives are the sum of i. i. d. terms the weak law of large numbers and the central limit theorem give

$$-\frac{1}{n} \nabla^2 l_n(\theta_0) \xrightarrow{P} I(\theta_0) \tag{2.3a}$$

and

$$\frac{1}{\sqrt{n}} \nabla l_n(\theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, I(\theta_0)\big) \tag{2.3b}$$

where θ_0 is the true parameter value.

2.2 Some Handwaving

If we expand $\nabla l_n(\theta)$ in a Taylor series around θ_0 and only keep the first two terms, we get

$$\nabla l_n(\theta) \approx \nabla l_n(\theta_0) + \nabla^2 l_n(\theta_0)(\theta - \theta_0). \tag{2.4}$$

If the maximum likelihood estimate (MLE) $\hat{\theta}_n$ occurs at an interior point of the parameter space, the first derivative will be zero, so plugging $\hat{\theta}_n$ in for θ in (2.4) gives

$$0 \approx \nabla l_n(\theta_0) + \nabla^2 l_n(\theta_0)(\hat{\theta}_n - \theta_0),$$

or

$$\hat{\theta}_n - \theta_0 \approx - \big(\nabla^2 l_n(\theta_0)\big)^{-1} \nabla l_n(\theta_0), \tag{2.5}$$

if we assume $\nabla^2 l_n(\theta_0)$ is an invertible matrix.

The right hand side of (2.5) is not the right size to converge in distribution, but if we multiply both sides by $\sqrt{n}$, we get

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \approx - \left(\frac{1}{n} \nabla^2 l_n(\theta_0)\right)^{-1} \cdot \frac{1}{\sqrt{n}} \nabla l_n(\theta_0), \tag{2.6}$$

and the right hand side converges to $I_1(\theta_0)^{-1} Z$, where $Z \sim \mathcal{N}\big(0, I_1(\theta_0)\big)$, by Slutsky's theorem.

This gives us the "usual" asymptotics of maximum likelihood, because the random vector $I_1(\theta_0)^{-1} Z$ has mean zero and variance $I_1(\theta_0)^{-1}$, so

$$\sqrt{n}\big(\hat{\theta}_n - \theta_0\big) \approx \mathcal{N}\big(0, I_1(\theta_0)^{-1}\big). \tag{2.7}$$

As the section heading says, there is a lot of handwaving in this argument. The main issue, at least the one over which the main analytical sweat is expended in most textbooks on asymptotics, is under what conditions the representation (2.4) using only two terms of the Taylor series is valid. But there are a number of other issues that come before that. The assumption that the log likelihood is differentiable and the derivatives have moments satisfying the Bartlett identities is limiting. But even before that, the assumption that we are only interested in i. i. d. sampling is *very* limiting. There are many interesting problems involving time series, spatial statistics, and so forth in which there is no i. i. d. sequence of random variables. And last but not least, right at the outset we made an assumption that our statistical model has densities. That may not seem limiting, but it is.

Thus we now switch from what should be fairly familiar territory to a theory having vast generality. The very generality of the theory necessitates it being

highly abstract and thus difficult to understand, but it does make clear many facts about likelihood inference without superfluous regularity conditions. It is thus well worth understanding.

Before we start that theory, let us examine the behavior of the log likelihood itself (rather than the MLE) at the same level of handwaving as the rest of this section. If we expand the log likelihood itself, rather than its first derivative as we did in (2.4), and this time keep three terms of the Taylor series, we get

$$l_n(\theta) - l_n(\theta_0) \approx \nabla l_n(\theta_0)(\theta - \theta_0) + \tfrac{1}{2}(\theta - \theta_0)' \nabla^2 l_n(\theta_0)(\theta - \theta_0). \tag{2.8}$$

If we try to fix this up so that the right hand side converges to something, we immediately see that simply multiplying by a constant doesn't do the job because (2.3a) shows that we need to multiply the second term by $1/n$ to get convergence and (2.3b) we need to multiply the first term by $1/\sqrt{n}$ to get convergence. A different trick, however, does do the job.

From (2.7) it is clear that we are interested in parameter values that are on the order of $1/\sqrt{n}$ in distance from θ_0. That suggests we replace θ in (2.8) by $\theta_0 + n^{-1/2}\delta$ giving

$$l_n(\theta_0 + n^{-1/2}\delta) - l_n(\theta_0) \approx \frac{1}{\sqrt{n}} \nabla l_n(\theta_0)\delta + \frac{1}{2n} \delta' \nabla^2 l_n(\theta_0)\delta \tag{2.9}$$

which is the right size to get convergence. The right hand side converges in law to the random function q defined by

$$q(\delta) = \delta' Z - \frac{1}{2} \delta' I(\theta_0) \delta$$

where

$$Z \sim \mathcal{N}\big(0, I(\theta_0)\big).$$

This random quadratic function q having nonrandom Hessian matrix $-I(\theta_0)$ and random gradient vector Z is the limiting form of the log likelihood function in "nice" cases.

2.3 Contiguity

Eventually we want to consider statistical models to be families of probability *measures* rather than families of probability *densities* with respect to one fixed measure. This may seem like useless generality. When does one ever consider models like that in real applications? But it turns out that the extra generality comes at no extra cost. Assuming densities buys nothing. Most of the mathematical difficulty of this section would still remain.

Eventually we will consider sequences of statistical models

$$\mathcal{P}_n = \{\, P_{n,\theta} : \theta \in \Theta_n \,\} \tag{2.10}$$

But to start we consider just comparing two measures in each such model. Thus consider two sequences of probability measures $\{P_n\}$ and $\{Q_n\}$ such that P_n

and Q_n have the same sample space for each n (they are both in $\mathcal{P}_n$ in terms of the preceding discussion).

We want to consider the likelihood ratio for comparing these two measures. But what is that when they don't have densities? In fact they always do have densities with respect to the measure $\lambda_n = P_n + Q_n$ by the Radon-Nikodym theorem (there may be no measure that dominates the whole family $\mathcal{P}_n$ but λ_n always dominates the two measures P_n and Q_n). If p_n denotes the Radon-Nikodym derivative of P_n with respect to λ_n, then $q_n = 1 - p_n$ is the Radon-Nikodym derivative of Q_n with respect to λ_n, and

$$Z_n = \frac{q_n}{p_n} = \frac{1 - p_n}{p_n}$$

is the likelihood ratio comparing Q_n to P_n. We use the convention that $Z_n = +\infty$ when $p_n = 0$ so that Z_n is always well defined. We also use the conventions $\log(0) = -\infty$ and $\log(+\infty) = +\infty$ so that the log likelihood ratio $\log Z_n$ is always well defined.

If P_n and Q_n have densities f_n and g_n with respect to some other measure, then

$$p_n = \frac{f_n}{f_n + g_n} \quad \text{and} \quad q_n = \frac{g_n}{f_n + g_n}$$

and

$$Z_n = \frac{q_n}{p_n} = \frac{g_n}{f_n}$$

when these ratios are well defined, that is, when f_n and g_n are not both zero. Thus we see that our device of using densities with respect to λ_n solves two problems: (1) it handles the case of families that are not dominated by one measure (2) it handles the case of densities that are zero at some points. Otherwise, it just does what you expect.

For this section we define boundedness in probability as follows. A sequence of extended-real-valued random variables X_n is *bounded in probability* under P_n if

$$\lim_{M \to \infty} \limsup_{n \to \infty} P_n(|X_n| \geq M) = 0.$$

The usual definition would have sup instead of lim sup and would therefore require $P_n(|X_n| = \infty) = 0$ for all n, which makes it useless for variables that take infinite values with nonzero probability. For finite-valued random variables, the two concepts are equivalent (Problem 2-1). For random vectors, we continue to use the old definition (which is equivalent to tightness).

Theorem 2.1. *Using the notation defined in the preceding discussion, the following properties are equivalent.*

(a) *Every sequence of random variables that converges in probability to zero under P_n also converges in probability to zero under Q_n.*

(b) *Every sequence of random variables that is bounded in probability under P_n is also bounded in probability under Q_n.*

(c) *Z_n is bounded in probability under Q_n.*

(d) *Z_n is uniformly integrable under P_n and $Q_n(Z_n = \infty) \to 0$.*

(e) *Z_n is uniformly integrable under P_n and $E_{P_n}(Z_n) \to 1$.*

Note that if we assumed Q_n was always absolutely continuous with respect to P_n, which would be the case if they were measures in a dominated family of measures having densities that are everywhere strictly positive, this would only simplify the theorem by collapsing the last two conditions to "Z_n is uniformly integrable under P_n" since we would then have $Q_n(Z_n = \infty) = 0$ and $E_{P_n}(Z_n) = 1$ for all n. The theorem would still be a very powerful result. That is what we meant by "assuming densities buys nothing."

Proof. We will prove these implications in the following order. First we show (d) $\iff$ (e) and (c) $\iff$ (d). Then we finish by showing (a) $\implies$ (b) $\implies$ (c) $\implies$ (a). But before we start on any of these implications we establish the following identity. For any $M \geq 0$

$$\begin{aligned} E_{P_n}\left(Z_n 1_{\{Z_n \geq M\}}\right) &= \int_{p_n > 0} q_n 1_{\{Z_n \geq M\}} \, d\lambda_n \\ &= \int q_n 1_{\{M \leq Z_n < \infty\}} \, d\lambda_n \\ &= Q_n(M \leq Z_n < \infty). \end{aligned} \tag{2.11}$$

Now (d) $\iff$ (e). The case $M = 0$ in (2.11) gives

$$E_{P_n}(Z_n) = Q_n(Z_n < \infty) = 1 - Q_n(Z_n = \infty), \tag{2.12}$$

which proves (d) $\iff$ (e).

Next (c) $\iff$ (d). By (2.11)

$$Q_n(Z_n \geq M) = E_{P_n}\left(Z_n 1_{\{Z_n \geq M\}}\right) + Q_n(Z_n = \infty),$$

so (c) holds if and only if

$$\lim_{M \to \infty} \limsup_{n \to \infty} E_{P_n}\left(Z_n 1_{\{Z_n \geq M\}}\right) = 0 \tag{2.13}$$

and

$$\limsup_{n \to \infty} Q_n(Z_n = \infty) = 0.$$

Uniform integrability of Z_n under P_n is

$$\lim_{M \to \infty} \sup_{n \in \mathbb{N}} E_{P_n}\left(Z_n 1_{\{Z_n \geq M\}}\right) = 0. \tag{2.14}$$

We must show that (2.13) and (2.14) are equivalent. Clearly (2.14) implies (2.13) so we only have to show the converse. To show (2.14) we must show that for every $\epsilon > 0$ there is an M such that

$$E_{P_n}\left(Z_n 1_{\{Z_n \geq M\}}\right) \leq \epsilon, \tag{2.15}$$

for all n. Since (2.13) implies that for every $\epsilon > 0$ there are an M and N such that (2.15) holds for all $n \geq N$, we only need to show that by increasing M we can get (2.15) to hold for $n < N$ too. By (2.11) we see that (2.15) is finite for each n and M, hence by dominated convergence

$$\lim_{M \to \infty} E_{P_n}\left(Z_n 1_{\{Z_n \geq M\}}\right) = E_{P_n}\left(Z_n 1_{\{Z_n = \infty\}}\right) = 0.$$

Thus by increasing M we can get (2.15) to hold for all n, which proves (c) $\iff$ (d).

Next (a) $\implies$ (b). Assume (a), suppose X_n is bounded in probability under P_n, and suppose to get a contradiction that X_n is not bounded in probability under Q_n. Then there exists an $\epsilon > 0$ such that for all $M > 0$ we have $Q_n(|X_n| > M) > \epsilon$ infinitely often; in particular, there is a subsequence n_k such that

$$Q_{n_k}(|X_{n_k}| > k) > \epsilon, \qquad \text{for all } k. \tag{2.16}$$

Since X_n is bounded in probability under P_n, for every $\delta > 0$ there exists an M_δ such that $P_n(|X_n| > M_\delta) < \delta$ for all n. This implies $P_{n_k}(|X_{n_k}| > k) < \delta$ for all large enough k. Letting Y_k be the indicator variable of the event $|X_{n_k}| > k$, this implies Y_k converges in probability to zero under P_{n_k} so by (a) it also converges in probability to zero under Q_{n_k}. But this contradicts (2.16). Hence (a) $\implies$ (b).

Next (b) $\implies$ (c). Equation (2.12) implies $E_{P_n}(Z_n) \leq 1$, so by Markov's inequality $P_n(Z_n \geq M) \leq 1/M$, so Z_n is bounded in probability under P_n. Thus (b) implies Z_n is also bounded in probability under Q_n, so (b) $\implies$ (c).

Finally (c) $\implies$ (a). Suppose to get a contradiction that (c) holds but (a) does not. Then there is a sequence of random variables X_n that converges in probability to zero under P_n but not under Q_n, hence there is an $\epsilon > 0$ such that $P_n(|X_n| > \epsilon) \to 0$ but $Q_n(|X_n| > \epsilon) \not\to 0$. Letting A_n denote the event $|X_n| > \epsilon$, this means $P_n(A_n) \to 0$, but there is a $\delta > 0$ and a subsequence n_k such that $Q_{n_k}(A_{n_k}) \geq \delta$ for all k. Note that $Z_n = 1/p_n - 1$, so

$$\begin{aligned} Q_n(A_n) &= Q_n(A_n \text{ and } p_n \leq \eta) + Q_n(A_n \text{ and } p_n > \eta) \\ &\leq Q_n(p_n \leq \eta) + \int_{A_n \text{ and } p_n > \eta} Z_n p_n \, d\lambda_n \\ &\leq Q_n\left(Z_n \geq \frac{1}{\eta} - 1\right) + \frac{1}{\eta} P_n(A_n) \end{aligned}$$

Property (c) implies that we can choose $\eta > 0$ and N such that

$$Q_n\left(Z_n \geq \frac{1}{\eta} - 1\right) < \delta, \qquad n \geq N,$$

but this contradicts $Q_{n_k}(A_{n_k}) \geq \delta$. Hence (c) $\implies$ (a). □

If any one (and hence all) of the properties in the theorem holds, then we say that the sequence $\{Q_n\}$ is *contiguous to* the sequence $\{P_n\}$. If $\{P_n\}$ is

also contiguous to $\{Q_n\}$ we say the sequences are *contiguous*. We will only be interested in the symmetric case when the sequences are contiguous.

By $\mathcal{L}(X_n \mid P_n)$ we mean the law of the random variable X_n under P_n, that is, $P_n \circ X_n^{-1}$.

Theorem 2.2 (Le Cam's Third Lemma). *Using the notation defined in the preceding discussion, and letting X_n be a random element of a Polish space E defined on the common sample space of P_n and Q_n, the following properties are equivalent.*

(a) $\mathcal{L}((X_n, Z_n) \mid P_n) \to H$, *a probability measure on $E \times \mathbb{R}$ and Q_n is contiguous to P_n.*

(b) $\mathcal{L}((X_n, Z_n) \mid Q_n) \to H'$, *a probability measure on $E \times \mathbb{R}$.*

When (a) and (b) hold, $H'(dx, dz) = zH(dx, dz)$.

Proof. First note that under (b) Z_n is bounded in probability under Q_n (because convergence in law implies boundedness in probability) hence by part (c) of Theorem 2.1 Q_n is contiguous to P_n. Thus both parts (a) and (b) of the lemma assert the same contiguity statement, which we may thus assume.

Let g be a bounded continuous function on $E \times \overline{\mathbb{R}}$. Then

$$\begin{aligned} E_{Q_n}\{g(X_n, Z_n)\} &= \int_{p_n=0} g(X_n, Z_n) q_n \, d\lambda_n + \int_{p_n>0} g(X_n, Z_n) Z_n p_n \, d\lambda_n \\ &= E_{Q_n}\left\{g(X_n, Z_n) 1_{\{Z_n=\infty\}}\right\} + E_{P_n}\{g(X_n, Z_n) Z_n\} \end{aligned}$$

Under (a) the second term on the right converges to $\int g(x, z) z H(dx, dz)$ because Z_n is uniformly integrable by part (d) of Theorem 2.1. Under (b) the left hand side converges to $\int g(x, z) H'(dx, dz)$. Thus it only remains to be shown that the first term on the right hand side converges to zero under either (a) or (b). Since g is bounded, say by M,

$$\left|E_{Q_n}\left\{g(X_n, Z_n) 1_{\{Z_n=\infty\}}\right\}\right| \le M E_{Q_n}(Z_n = \infty).$$

Under (a) this converges to zero by part (d) of Theorem 2.1. Under (b) Z_n converges in law under Q_n to a random variable Z with law Q concentrated on $\mathbb{R}$. Thus the portmanteau theorem implies

$$\limsup_{n\to\infty} Q_n(Z_n = \infty) \to Q(Z = \infty) = 0.$$

□

Corollary 2.3. *Using the notation defined above, the following properties are equivalent.*

(a) $\mathcal{L}(Z_n \mid P_n) \to H$, *a probability measure on $\mathbb{R}$ and Q_n is contiguous to P_n.*

(b) $\mathcal{L}(Z_n \mid Q_n) \to H'$, *a probability measure on $\mathbb{R}$.*

When (a) and (b) hold, $H'(dz) = zH(dz)$.

Proof. In the theorem, take the case where E is a space with only one point. □

2.4 Local Asymptotic Normality

Consider a sequence of statistical models $\mathcal{P}_n = \{ P_{n,\theta} : \theta \in \Theta_n \}$ having parameter spaces Θ_n that are subsets of $\mathbb{R}^d$. In the i. i. d. case all of the models $\mathcal{P}_n$ will be n-fold products of one family $\mathcal{P} = \{ P_\theta : \theta \in \Theta \}$. That is, for each n we have $\Theta_n = \Theta$ and

$$P_{n,\theta} = \underbrace{P_\theta \times \cdots \times P_\theta}_{n \text{ terms}}$$

If the family $\mathcal{P}$ is dominated by a measure λ, so there are densities $f_\theta = dP_\theta/d\lambda$, then the log likelihood for the family $\mathcal{P}_n$ is given by (2.2), and looking back at (2.9) we see that our inference will depend on the asymptotic properties of the log likelihood ratio

$$l_n(\theta_0 + n^{-1/2}\delta) - l_n(\theta_0)$$

Thus in order to have the log likelihood converge to a nontrivial limit we have to change parameters from the original parameter θ to the new parameter δ, the two parameters being related by

$$\delta = \sqrt{n}(\theta - \theta_0) \qquad \text{or} \qquad \theta = \theta_0 + n^{-1/2}\delta \tag{2.17}$$

In general, however, we adopt no restrictions on the models $\mathcal{P}_n$ other than that they have parameter spaces of the same (finite) dimension. Thus even with i. i. d. sampling, we make no assumption that the same parameterization is used for all n, and hence there is no guarantee that (2.17) is the right reparameterization. When the sequence of models $\mathcal{P}_n$ does not arise from i. i. d. sampling, there is, in general, no simple change of parameters that does the job of making the log likelihood converge to a nontrivial limit. Thus there is no notation that can indicate the required reparameterization. We will simply have to assume that there is such a reparameterization and that it has been done.

We specify log likelihood ratios by the device used in the preceding section. For any two parameter values $\theta, \vartheta \in \Theta$, let $\lambda_{n,\theta,\vartheta} = P_{n,\theta} + P_{n,\vartheta}$ and let $p_{n,\theta}$ be the Radon-Nikodym derivative of $P_{n,\theta}$ with respect to $\lambda_{n,\theta,\vartheta}$, so $1 - p_{n,\theta}$ is the Radon-Nikodym derivative of $P_{n,\vartheta}$ with respect to $\lambda_{n,\theta,\vartheta}$. Then

$$l_n(\vartheta, \theta) = \log \frac{1 - p_{n,\theta}}{p_{n,\theta}}$$

is the log likelihood ratio for comparing the two distributions.

Definition 2.1 (Locally Asymptotically Normal). *A sequence of statistical models $\mathcal{P}_n$ with true parameter values θ_n is* locally asymptotically normal (LAN) *if the following three conditions hold*

(a) *For any $M > 0$, the parameter space Θ_n contains a ball of radius M centered at θ_n for all but finitely many n.*

(b) *For any bounded sequence δ_n in $\mathbb{R}^d$, the sequences P_{n,θ_n} and $P_{n,\theta_n+\delta_n}$ are contiguous.*

(c) *There exist sequences of random vectors S_n and random almost surely positive definite matrices K_n defined on the sample space of $\mathcal{P}_n$ such that K_n converges in probability to a nonrandom positive definite matrix K and for every bounded sequence δ_n in $\mathbb{R}^d$*

$$l_n(\theta_n + \delta_n, \theta_n) - \left[\delta_n' S_n - \tfrac{1}{2}\delta_n' K_n \delta_n\right] \xrightarrow{P} 0 \tag{2.18}$$

under P_{n,θ_n}.

Note that condition (a) guarantees that $\theta_n + \delta_n$ is in Θ_n for all sufficiently large n. Thus in the other two conditions the notation $P_{n,\theta_n+\delta_n}$ and $l_n(\theta_n + \delta_n, \theta_n)$ makes sense for sufficiently large n. The same S_n and K_n must work for all bounded sequences $\{\delta_n\}$. Although (c) only requires the left hand side of (2.18) to converge in probability to zero under P_{n,θ_n}, the contiguity assumed in (b) implies that it also converges in probability to zero under $P_{n,\theta_n+\delta_n}$.

Note that there is no mention of the normal distribution in the definition of LAN. The following theorem shows how the normal distribution comes in. In the following we change notation from $\mathcal{L}(X_n \mid P_{\theta_n})$ to $\mathcal{L}(X_n \mid \theta_n)$ to save one level of subscripts.

Theorem 2.4. *If the LAN condition holds for a sequence of models, and $\delta_n \to \delta$, then*

$$\mathcal{L}(S_n \mid \theta_n + \delta_n) \to \mathcal{N}(K\delta, K).$$

Proof. Conditions (b) and (c) of Theorem 2.1 taken together imply that $l_n(\theta_n + \delta_n, \theta_n)$ is bounded in probability under $P_{\theta_n+\delta_n}$ for any bounded sequence δ_n. Hence from the definition of LAN, the sequence S_n is also bounded in probability. Thus by Prohorov's theorem it has convergent subsequences $S_{n_k} \to S$. By the definition of LAN and Slutsky's theorem

$$Z_n = e^{l_n(\theta_n+\delta_n,\theta_n)}$$

is equal to

$$e^{\delta_n' S_n - \frac{1}{2}\delta_n' K_n \delta_n}$$

plus a term that converges in probability to zero. Hence under P_{θ_n}

$$Z_{n_k} \xrightarrow{\mathcal{L}} \exp\left\{\delta' S - \tfrac{1}{2}\delta' K \delta\right\}$$

and by Condition (e) of Theorem 2.1 it follows that the expectation of the right hand side is one. Thus

$$E \exp\left(\delta' S\right) = \exp\left\{\tfrac{1}{2}\delta' K \delta\right\}$$

Since this is true for all vectors δ, this shows that the moment generating function of S is that of the $\mathcal{N}(0, K)$ distribution. Thus we have shown

$$\mathcal{L}(S_{n_k} \mid \theta_{n_k}) \to \mathcal{N}(0, K).$$

Since the limit does not depend on the subsequence chosen, the whole sequence must converge to this limit.

By Le Cam's third lemma, (S_n, Z_n) must also converge under $P_{\theta_n+\delta_n}$. The only remaining task is to apply the lemma to find the distribution. Suppose

$$\begin{aligned} \mathcal{L}((S_n, Z_n) \mid \theta_n) &\to H \\ \mathcal{L}((S_n, Z_n) \mid \theta_n + \delta_n) &\to H' \end{aligned}$$

Then the lemma says $H'(ds, dz) = zH(ds, dz)$. The measure H is degenerate

$$Z = \exp\left\{\delta' S - \tfrac{1}{2}\delta' K \delta\right\} \tag{2.19}$$

holds with probability one and $S \sim \mathcal{N}(0, K)$. Thus the measure H' is also degenerate, (2.19) holding with probability one. Write the density of S with respect to Lebesgue measure under H as

$$f(s) = c \exp\left(-\tfrac{1}{2} s' K^{-1} s\right)$$

where c is a constant (its value doesn't matter). Then the density of S with respect to Lebesgue measure under H' is

$$\begin{aligned} f'(s) &= z f(s) \\ &= c \exp\left(\delta' s - \tfrac{1}{2}\delta' K \delta - \tfrac{1}{2} s' K^{-1} s\right) \\ &= c \exp\left\{-\tfrac{1}{2}(s - K\delta)' K^{-1} (s - K\delta)\right\} \end{aligned}$$

which is the $\mathcal{N}(K\delta, K)$ distribution asserted by the theorem. □

2.5 Asymptotic Equivalence

We say that two sequences of estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ of a parameter sequence θ_n are *asymptotically equivalent* if

$$\tilde{\theta}_n - \hat{\theta}_n = o_p(\hat{\theta}_n - \theta_n). \tag{2.20a}$$

The notation means, for any $\epsilon > 0$,

$$\Pr\bigl(\|\tilde{\theta}_n - \hat{\theta}_n\| \le \epsilon \|\hat{\theta}_n - \theta_n\|\bigr) \to 1, \qquad \text{as } n \to \infty, \tag{2.20b}$$

where $\|\cdot\|$ is any norm on $\mathbb{R}^d$.

We check that this is, as the name implies, an equivalence relation among estimating sequences. We must check three properties, that the relation is reflexive, symmetric, and transitive. To introduce a notation for the relation in question, write $\tilde{\theta}_n \preceq \hat{\theta}_n$ if (2.20a) holds.

Reflexivity is the requirement that $\hat{\theta}_n \preceq \hat{\theta}_n$ hold for any estimating sequence $\hat{\theta}_n$. This is trivial.

Symmetry is the requirement that $\tilde{\theta}_n \preceq \hat{\theta}_n$ implies $\hat{\theta}_n \preceq \tilde{\theta}_n$ for any estimating sequences $\hat{\theta}_n$ and $\tilde{\theta}_n$. This is proved as follows. By the triangle inequality

$$\|\tilde{\theta}_n - \hat{\theta}_n\| \le \epsilon \|\hat{\theta}_n - \theta_n\| \tag{2.21a}$$

implies

$$\|\tilde{\theta}_n - \hat{\theta}_n\| \le \epsilon \|\tilde{\theta}_n - \theta_n\| + \epsilon \|\tilde{\theta}_n - \hat{\theta}_n\|$$

which if $\epsilon < 1$ is the same as

$$\|\tilde{\theta}_n - \hat{\theta}_n\| \le \frac{\epsilon}{1 - \epsilon} \|\tilde{\theta}_n - \theta_n\|. \tag{2.21b}$$

Thus $\tilde{\theta}_n \preceq \hat{\theta}_n$ implies that (2.21a) can has probability converging to one whenever $0 < \epsilon < 1$, which in turn implies (2.21b) has probability converging to one for all such ϵ, and this in turn implies $\hat{\theta}_n \preceq \tilde{\theta}_n$.

Transitivity is the requirement that $\theta_n^* \preceq \tilde{\theta}_n$ and $\tilde{\theta}_n \preceq \hat{\theta}_n$ imply $\theta_n^* \preceq \hat{\theta}_n$ for any estimating sequences θ_n^*, $\hat{\theta}_n$, and $\tilde{\theta}_n$. This is proved as follows. Again by the triangle inequality

$$\|\theta_n^* - \tilde{\theta}_n\| \le \epsilon_1 \|\tilde{\theta}_n - \theta_n\| \tag{2.22a}$$

and

$$\|\tilde{\theta}_n - \hat{\theta}_n\| \le \epsilon_2 \|\hat{\theta}_n - \theta_n\| \tag{2.22b}$$

imply

$$\|\theta_n^* - \hat{\theta}_n\| \le \|\theta_n^* - \tilde{\theta}_n\| + \|\tilde{\theta}_n - \hat{\theta}_n\| \le \epsilon_1 \|\tilde{\theta}_n - \theta_n\| + \epsilon_2 \|\hat{\theta}_n - \theta_n\|$$

and if $\epsilon_1 < 1$ this implies, by the implication of (2.21b) by (2.21a),

$$\|\theta_n^* - \hat{\theta}_n\| \le \left(\frac{\epsilon_1}{1 - \epsilon_1} + \epsilon_2 \right) \|\hat{\theta}_n - \theta_n\|. \tag{2.22c}$$

Thus $\theta_n^* \preceq \tilde{\theta}_n$ and $\tilde{\theta}_n \preceq \hat{\theta}_n$ imply (2.22a) and (2.22b) have probability converging to one whenever $0 < \epsilon_1 < 1$ and $0 < \epsilon_2$, which in turn implies (2.22c) has probability converging to one for all such ϵ_1 and ϵ_2, and this in turn implies $\theta_n^* \preceq \hat{\theta}_n$.

2.6 Efficient Likelihood Estimators

Roughly speaking, the LAN condition says that $l_n(\theta_n + \delta, \theta_0)$ considered as a function of δ is well approximated by the random quadratic function $\delta' S_n - \frac{1}{2} \delta' K_n \delta$. The maximizer of this random quadratic function is $K_n^{-1} S_n$. The corresponding estimator of θ is

$$\hat{\theta}_n = \theta_n + K_n^{-1} S_n. \tag{2.23}$$

We say that a sequence of estimators $\tilde{\theta}_n$ is an *efficient likelihood estimator* (ELE) of the parameter θ_n of a model satisfying the LAN condition if it is

asymptotically equivalent to (2.23). It is clear from our proof that asymptotic equivalence is indeed an equivalence relation on estimating sequences that every ELE is asymptotically equivalent to every other ELE.

Note that the definition is not useful in calculating an ELE, because it refers to the true parameter value θ_n, which is unknown. Hence it is an open question to be answered by further investigation whether an ELE can be calculated. In "nice" models the MLE is an ELE, but we will have to do a fair amount of work to see that. Also we haven't yet defined efficiency or said what that has to do with ELEs.

We say that a sequence of estimators $\tilde{\theta}_n$ is a *regular estimator* of the parameter θ_n of a model satisfying the LAN condition if the limiting distribution of $\tilde{\theta}_n - (\theta_n + \delta_n)$ under $P_{\theta_n+\delta_n}$ is the same for any bounded sequence δ_n.

Corollary 2.5. *If the LAN condition holds for a sequence of models, $\delta_n \to \delta$, and $\tilde{\theta}_n$ is an ELE, then*

$$\mathcal{L}(\tilde{\theta}_n - \theta_n \mid \theta_n + \delta_n) \to \mathcal{N}(\delta, K^{-1}), \tag{2.24}$$

and $\tilde{\theta}_n$ is a regular estimator.

Proof. By definition of ELE $\tilde{\theta}_n - \theta_n = K_n^{-1} S_n + o_p(1)$ and hence converges in distribution under $P_{\theta_n+\delta_n}$ to $K^{-1}S$ where $S \sim \mathcal{N}(K\delta, K)$ by Theorem 2.4 and Slutsky's theorem. Clearly, $K^{-1}S$ is normal with mean δ and variance K^{-1}. This proves (2.24).

In the definition of regular estimator we do not assume $\delta_n \to \delta$, only that δ_n is bounded, but then there are convergent subsequences $\delta_{n_k} \to \delta$ and for such a subsequence, the first part of the proof implies

$$\mathcal{L}(\tilde{\theta}_{n_k} - (\theta_{n_k} + \delta_{n_k}) \mid \theta_{n_k} + \delta_{n_k}) \to \mathcal{N}(0, K^{-1}).$$

Since every such subsequence converges to the same limit, the whole sequence $\tilde{\theta}_n - (\theta_n + \delta_n)$ converges to this limit under $P_{\theta_n+\delta_n}$. Since the limit does not depend on the sequence δ_n, the estimator is regular. □

2.7 The Parametric Bootstrap

Readers who have heard of the bootstrap have most likely heard of the *nonparametric* bootstrap, which approximates sampling distributions by resampling the data with replacement. In the same paper in which Efron introduced the nonparametric bootstrap, he also pointed out the connections with the *parametric* bootstrap, which approximates sampling distributions by simulating new data from a fitted parametric model.

The nonparametric bootstrap was a completely new idea. Though vaguely related to the jackknife and to permutation tests, it was much more widely applicable and had a much more general theory. In contrast, the parametric bootstrap was not new at all. Efron had just given a new name to something

people had done since the beginnings of statistics. This was clear from Efron's examples of the parametric bootstrap, which were z and t confidence intervals.

Suppose X_1, ..., X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$ random variables, and suppose we want a confidence interval for μ. We base the confidence interval on the quantity

$$R(X_1, \ldots, X_n, \mu) = \overline{X}_n - \mu. \tag{2.25}$$

The confidence interval will be the set of all μ such that $|R(X_1, \ldots, X_n, \mu)| < c$, where c is chosen so that the the confidence interval has the right coverage probability. The problem is that we do not know the sampling distribution of (2.25). It is, of course, $\mathcal{N}(0, \sigma^2)$, but we do not know σ. The bootstrap principle tells us to simulate the sampling distribution of (2.25) using the fitted model, that is we simulate new data X_1^*, ..., X_n^* that are i. i. d. $\mathcal{N}(\hat{\mu}_n, \hat{\sigma}_n^2)$, where $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ are any consistent estimators of μ and σ. We calculate the critical value c from the sampling distribution of

$$R(X_1^*, \ldots, X_n^*, \hat{\mu}_n) = \overline{X}_n^* - \hat{\mu}_n. \tag{2.26}$$

where $\overline{X}_n^*$ is the sample mean of the X_i^*. The replacement of μ by $\hat{\mu}_n$ makes sense because $\hat{\mu}_n$ is the true mean of the X_i^*. Of course, in this simple example, we do not have to calculate the distribution of (2.26) by simulation. We know that it is $\mathcal{N}(0, \hat{\sigma}_n^2/n)$ and hence the appropriate value of c for a 95% confidence interval is $1.96\hat{\sigma}_n/\sqrt{n}$ and the 95% confidence interval is the usual z confidence interval

$$\overline{X}_n \pm 1.96 \cdot \frac{\hat{\sigma}_n}{\sqrt{n}}$$

The t confidence interval is produced by the same argument applied to a different "root". We replace (2.25) by the pivotal quantity

$$R(X_1, \ldots, X_n, \mu) = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}$$

where now we use a particular estimator of variance

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

Then (2.26) becomes

$$R(X_1^*, \ldots, X_n^*, \mu) = \frac{\overline{X}_n^* - \hat{\mu}_n}{S_n^*/\sqrt{n}}$$

and again we do not have to simulate to find the sampling distribution. It is exactly $t(n-1)$. Denoting the appropriate critical value for a $100(1-\alpha)\%$ confidence interval by $t_{\alpha/2}$, the confidence interval becomes

$$\overline{X}_n \pm t_{\alpha/2} \cdot \frac{S_n}{\sqrt{n}}$$

In these well-known examples, the "parametric bootstrap" argument seems just a way of making something simple seem complicated, but it gives us a method of great generality that can be applied to any parametric model. Suppose X had distribution P_θ, where θ is an unknown parameter and $R(X, \theta)$ is any scalar-valued function of the data and parameter. Suppose $\hat{\theta}(X)$ is an estimator of θ. The parametric bootstrap says to approximate the sampling distribution of $R(X, \theta)$ by simulating new data X^* from $P_{\hat{\theta}(X)}$ and calculating by simulation (or analytically if possible) the distribution of

$$R\big(X^*, \hat{\theta}(X)\big) \tag{2.27}$$

considering X fixed (that is, conditioning on the observed value of the real data). From the simulations, we find a c such that

$$\Pr\big\{ R\big(X^*, \hat{\theta}(X)\big) < c \;\big|\; X \big\} = 1 - \alpha$$

Then the parametric bootstrap principle says to take

$$\{\, \theta : R(X, \theta) < c \,\}$$

as an approximate $100(1 - \alpha)\%$ confidence region for θ.

Of course, in general there is no guarantee that the parametric bootstrap does the right thing, because $\hat{\theta}(X)$ is not the true parameter value, hence the sampling distribution of (2.27) is not exactly the same as the sampling distribution of $R(X, \theta)$ when θ is the true parameter value, which is the distribution we *should* use to determine the critical value. Even in our simple examples, it did exactly the right thing only in the t example. In the z example, it did approximately the right thing, producing the usual asymptotic confidence interval. In other situations, it may not do the right thing, even approximately.

Thus the parametric bootstrap is a heuristic that gives us a procedure that may or may not be valid. It leaves us the theoretical question of whether it provides a valid approximation to the sampling distribution in question. And let us stress again that, although the language may seem a bit strange, the parametric bootstrap includes what we do in *every* parametric statistical problem. We never know the true parameter value and must always rely on estimating it. And we always need some argument that explains why using an estimate instead of the truth does not destroy the validity of the procedure. The "parametric bootstrap" is just a general framework that encompasses all such situations.

With this explanation of the parametric bootstrap, we can now see the point of regular estimators. If the LAN condition holds for a sequence of models, and $\hat{\theta}_n$ is a regular estimator, then the parametric bootstrap gives a valid approximation of the sampling distribution of $\hat{\theta}_n$. More precisely, let

$$H(n, \delta) = \mathcal{L}(\hat{\theta}_n - (\theta_n + \delta) \mid \theta_n + \delta),$$

then the parametric bootstrap is the procedure of using $H(n, \hat{\theta}_n - \theta_n)$ as an approximation of $H(n, 0)$, the latter being the true sampling distribution of $\hat{\theta}_n - \theta_n$ and the former being its parametric bootstrap approximation.

By definition of "regular," for any bounded sequence δ_n

$$\mathcal{L}(\hat{\theta}_n - (\theta_n + \delta_n) \mid \theta_n + \delta_n) \to L \tag{2.28}$$

for some law L that does not depend on the sequence δ_n. In particular

$$\mathcal{L}(\hat{\theta}_n - \theta_n \mid \theta_n) \to H(n, 0),$$

so we see that $L = H(n, 0)$ and that another notation for (2.28) is

$$H(n, \delta_n) \to H(n, 0) \tag{2.29}$$

whenever δ_n is bounded.

By the Skorohod theorem we can find random variables

$$\begin{aligned} \hat{\theta}_n^* &\sim \mathcal{L}(\hat{\theta}_n \mid \theta_n) \\ Z &\sim H(n, 0) \end{aligned}$$

such that

$$\hat{\theta}_n^* - \theta_n \xrightarrow{\text{a. s.}} Z, \qquad \text{as } n \to \infty$$

Then for every ω not in the null set for which this convergence fails

$$\hat{\theta}_n^*(\omega) - \theta_n \longrightarrow Z(\omega), \qquad \text{as } n \to \infty \tag{2.30}$$

Let d be any metric that metrizes weak convergence. Fristedt and Gray (1997, Section 18.7) discuss one such metric, the *Prohorov metric*. But the details do not matter, only that there is such a metric. Thus (2.29) and (2.30) imply

$$d\big(H(n, \hat{\theta}_n^*(\omega) - \theta_n), H(n, 0)\big) \longrightarrow 0$$

or, dropping the ω's,

$$d\big(H(n, \hat{\theta}_n^* - \theta_n), H(n, 0)\big) \xrightarrow{\text{a. s.}} 0,$$

and, since almost sure convergence implies convergence in probability, we also have convergence in probability. But then we can drop the stars, since convergence in probability only depends on the laws of variables and the variables with and without stars have the same laws. Thus

$$d\big(H(n, \hat{\theta}_n - \theta_n), H(n, 0)\big) \xrightarrow{P} 0.$$

or

$$H(n, \hat{\theta}_n - \theta_n) \xrightarrow{P} H(n, 0).$$

This is a formal statement of the sense in which the parametric bootstrap "works." Note that the crucial assumption was that $\hat{\theta}_n$ is regular.

2.8 The Hájek Convolution Theorem

Theorem 2.6 (Hájek Convolution Theorem). *If the LAN condition holds for a sequence of models, and $\tilde{\theta}_n$ is a regular estimator, then*

$$\mathcal{L}(\tilde{\theta}_n - \theta_n \mid \theta_n) \longrightarrow \mathcal{L}(X + W)$$

where $X \sim \mathcal{N}(0, K^{-1})$ and W is independent of X. If $W = 0$ with probability one, then $\tilde{\theta}_n$ is an ELE.

Proof. Write $\hat{\theta}_n = \theta_n + K_n^{-1} S_n$. This is an ELE, so by Corollary 2.5

$$\mathcal{L}(\hat{\theta}_n - \theta_n \mid \theta_n) \longrightarrow \mathcal{L}(X).$$

We first prove the theorem under the additional assumption that $\tilde{\theta}_n$ and $\hat{\theta}_n$ converge jointly under P_{θ_n}

$$(\tilde{\theta}_n - \theta_n, \hat{\theta}_n - \theta_n) \xrightarrow{\mathcal{L}} H_0$$

(the theorem without this assumption will then follow from Prohorov's theorem and the subsequence principle). Then Le Cam's Third Lemma implies, if $\delta_n \to \delta$, then under $P_{\theta_n + \delta_n}$

$$(\tilde{\theta}_n - \theta_n, \hat{\theta}_n - \theta_n) \xrightarrow{\mathcal{L}} H_\delta$$

where

$$H_\delta(dy, dx) = \exp\{\delta' K x - \tfrac{1}{2}\delta' K \delta\} \cdot H_0(dy, dx). \tag{2.31}$$

We know from Corollary 2.5 that the marginal distribution of X under H_δ is $\mathcal{N}(\delta, K^{-1})$. We also know from the regularity assumption that the marginal distribution of $Y - \delta$ under H_δ does not depend on δ. What we need to show is that $Y = X + W$ with W independent of X.

We do this using a trick. Give δ a $\mathcal{N}(0, \Gamma^{-1})$ prior distribution. Then the joint distribution of the triple (Y, X, δ) has the measure

$$M(dy, dx, d\delta) = C \exp\{\delta' K x - \tfrac{1}{2}\delta'(K + \Gamma)\delta\} \cdot H_0(dy, dx)\, d\delta$$

where C is a constant (its value doesn't matter). Define $B = (K + \Gamma)^{-1} K$ and introduce the new variables

$$\begin{aligned} V &= BX \\ Z &= Y - V \\ \xi &= \delta - V \end{aligned}$$

Since this is a linear change of variables, the Jacobian is constant. In the new variables

$$\delta' K x - \tfrac{1}{2}\delta'(K + \Gamma)\delta = \tfrac{1}{2} V'(K + \Gamma) V - \tfrac{1}{2}\xi'(K + \Gamma)\xi$$

Note that fixing (X, Y) fixes (V, Z) and vice versa. Conditioning on either pair of variables, we see that the conditional distribution of ξ is $\mathcal{N}\bigl(0, (K + \Gamma)^{-1}\bigr)$. Since this distribution does not depend on (V, Z), we conclude that $\xi \perp (V, Z)$.

Now note that $Y - \delta = Z - \xi$, so both sides have the same distribution. Note that Z and ξ are independent and we know the distribution of ξ. By the regularity assumption $\mathcal{L}(Y - \delta \mid \delta)$ does not depend on δ. Hence this law is also the marginal of $Y - \delta$, which is equal to the marginal of $Z - \xi$. Thus we have established that $\mathcal{L}(Y - \delta \mid \delta)$ is the convolution of $L(Z)$, which is unknown, and the $\mathcal{N}\bigl(0, (K + \Gamma)^{-1}\bigr)$ distribution of ξ.

Now let $\Gamma \to 0$. This does not affect $\mathcal{L}(Y - \delta \mid \delta)$ which does not depend on Γ. However, the distribution of ξ converges to $\mathcal{N}(0, K^{-1})$, and the distribution of $Z = Y - BX$ converges to $\mathcal{L}(Y - X)$. Thus we see that $\mathcal{L}(Y - \delta \mid \delta)$ is the convolution of $\mathcal{N}(0, K^{-1})$ and some other distribution $\mathcal{L}(Y - X)$. And that proves the theorem under the additional assumption that $\tilde{\theta}_n$ and $\hat{\theta}_n$ converge jointly under P_{θ_n}.

To remove this assumption, note that under the conditions of the theorem $\tilde{\theta}_n$ and $\hat{\theta}_n$ both converge marginally under P_{θ_n}. Hence both are tight. Hence their joint distributions are tight. Thus every subsequence has a jointly convergent subsubsequence by Prohorov's theorem. By what we have proved above, every such sequence has the same limit. (If φ_Y, φ_X, and φ_W are the characteristic functions of Y, X, and $W = Y - X$, and X and W are independent, then $\varphi_Y = \varphi_X \varphi_W$, and since φ_X is never zero, $\varphi_W = \varphi_Y / \varphi_X$. Thus the distribution of W is determined by those of Y and X, which are determined by the hypotheses of the theorem.) Hence the whole sequence converges to the same limit, and that proves the theorem.

The last assertion of the theorem, the case when $W = 0$ almost surely, is now obvious: then $\tilde{\theta}_n - \hat{\theta}_n \xrightarrow{\mathcal{L}} 0$, hence these estimators are asymptotically equivalent, hence $\tilde{\theta}_n$ is an ELE too. □

What the theorem says is that no regular estimator can beat an ELE. The asymptotic distribution of any regular estimator has the form $X + W$ with X the asymptotic distribution of any ELE and W independent of X. Thus the regular estimator makes whatever error the ELE makes and then adds the error W on top of that. Hence it is worse unless W is concentrated at 0, in which case it does as well but no better than any ELE. This is easiest to see when W has a second moment. Then the asymptotic variance of the regular estimator is $K^{-1} + \operatorname{Var}(W)$, which is clearly no better than the asymptotic variance K^{-1} of an ELE.

Much more is said about contiguity theory in Le Cam and Yang (1990), although your humble author finds it hard to read. In fact the guide for our Theorems 2.1 and 2.2 was Jacod and Shiryaev (1987, Chapter V, Section 1a). Our guide for the Hájek convolution theorem was Le Cam and Yang (1990) who attribute their proof to van der Vaart.

2.9 Exponential Families Revisited

We return to exponential families to provide a concrete example to which all this theory applies and to get a much stronger theorem than Theorem 1.5 of the preceding chapter.

We now consider a sequence of standard exponential families with natural statistics (and natural parameters) of the same dimension, but for the present we make *no other assumptions*. Hence let λ_n be a sequence of measures on $\mathbb{R}^d$ having Laplace transforms c_n, cumulant functions k_n, densities with respect to λ_n

$$f_{n,\theta}(x) = \frac{1}{c_n(\theta)} e^{\langle x, \theta \rangle}$$

for $\theta \in \operatorname{dom} c_n = \operatorname{dom} k_n$. As in the rest of this chapter, we denote the true parameter value in the n-th model by θ_n.

Note that in the log likelihood for this sequence of families,

$$l_n(\theta) = \langle X_n, \theta \rangle - k_n(\theta),$$

where X_n is the data for the n-th family, the first term, which is linear in the parameter, is the only random term, the other term (the cumulant function), which has all the curvature, is nonrandom. Thus whether the log likelihood close to quadratic or not only depends on the cumulant function.

Definition 2.2 (LAQ Exponential Families). *A sequence k_n of cumulant functions on $\mathbb{R}^d$ (or the corresponding sequence of standard exponential families) is* locally asymptotically quadratic (LAQ) *if there exists a sequence of vectors b_n and a positive definite matrix K such that for every $\delta \in \mathbb{R}^d$*

$$k_n(\theta_n + \delta) - k_n(\theta_n) - \langle b_n, \delta \rangle \to \tfrac{1}{2} \langle \delta, K\delta \rangle, \qquad \text{as } n \to \infty.$$

Note that it is part of the definition that K is (strictly) positive definite, although the term "locally asymptotically quadratic" doesn't explicitly indicate this.

We interrupt our treatment of exponential families for a brief discussion of convergence of convex functions.

Theorem 2.7. *Let f_n be a sequence of extended-real-valued convex functions on $\mathbb{R}^d$ converging pointwise on a dense subset of $\mathbb{R}^d$ to a limit that is finite on some open subset of $\mathbb{R}^d$. Then there exists a unique extended-real-valued lower-semicontinuous convex function f on $\mathbb{R}^d$ such that f_n converges to f uniformly on every compact set that does not contain a boundary point of* $\operatorname{dom} f$.

The proof of this is rather complicated and uses a lot of convexity theory that we don't want to take time to develop here. So we will merely give a reference to the literature. This is part of Theorem 7.17 in Rockafellar and Wets (1998).

Corollary 2.8. *If the LAQ condition is satisfied, the convergence is actually uniform on compact sets, that is, for every compact set C in $\mathbb{R}^d$*

$$\sup_{\delta \in C} \left| k_n(\theta_n + \delta) - k_n(\theta_n) - \langle b_n, \delta \rangle - \tfrac{1}{2} \langle \delta, K\delta \rangle \right| \to 0, \qquad \text{as } n \to \infty.$$

This is a straightforward consequence of the theorem. Here the limit function $\delta \mapsto \frac{1}{2}\langle \delta, K\delta \rangle$ is everywhere finite, so its effective domain is all of $\mathbb{R}^d$, and we have uniform convergence on compact subsets of $\mathbb{R}^d$.

Lemma 2.9. *If X_n are random variables from a sequence of exponential families with cumulant functions k_n satisfying the LAQ condition and true parameter values θ_n, then*

$$\begin{aligned} \mu_n &= \nabla k_n(\theta_n) \\ K_n &= \nabla^2 k_n(\theta_n) \end{aligned} \tag{2.32}$$

exist for all but finitely many n, and

$$\begin{aligned} \mu_n - b_n &\to 0 \\ K_n &\to K \end{aligned}$$

where b_n and K are defined in the LAQ condition. Moreover,

$$X_n - b_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, K),$$

and for any bounded sequence δ_n the sequence

$$e^{\langle X_n - b_n, \delta_n \rangle}$$

is uniformly integrable.

Proof. By the theory of moment and cumulant generating functions, the derivatives in (2.32) exist as soon as k_n is finite on a neighborhood of θ_n. Corollary 2.8 implies this occurs for all but finitely many n.

Now, $X_n - b_n$ has moment generating function

$$\begin{aligned} M_n(t) &= E\left\{ e^{\langle X_n - b_n, t \rangle} \right\} \\ &= \int e^{\langle x - b_n, t \rangle} f_{n,\theta_n}(x) \lambda_n(dx) \\ &= \frac{1}{c_n(\theta_n)} \int e^{\langle x - b_n, t \rangle} e^{\langle x, \theta_n \rangle} \lambda_n(dx) \\ &= \frac{e^{-\langle b_n, t \rangle}}{c_n(\theta_n)} \int e^{\langle x, \theta_n + t \rangle} \lambda_n(dx) \\ &= \frac{c_n(\theta_n + t) e^{-\langle b_n, t \rangle}}{c_n(\theta_n)} \end{aligned}$$

and by the LAQ condition, this converges to the moment generating function $t \mapsto \exp(\frac{1}{2}\langle t, Kt \rangle)$ of a $\mathcal{N}(0, K)$ random vector. Since moment generating function convergence also implies convergence of moments, we also have

$$E(X_n - b_n) = \mu_n - b_n \to 0$$

and

$$\operatorname{Var}(X_n - b_n) = K_n \to K.$$

Moment generating function convergence also implies the random variables $e^{\langle X_n - b_n, t\rangle}$ are uniformly integrable for any fixed vector t. We can pick a finite set $T = \{t_1, \ldots, t_k\}$ such that the entire sequence δ_n is contained in the convex hull of T (the vertices of a sufficiently large cube, for example). This means every δ_n can be written as a convex combination

$$\delta_n = \sum_{i=1}^{k} s_{i,n} t_i$$

where the s_{in} are nonnegative and $s_{1n} + \cdots + s_{kn} = 1$. Then by the convexity inequality

$$e^{\langle X_n - b_n, \delta_n\rangle} \le \sum_{i=1}^{k} s_{i,n} e^{\langle X_n - b_n, t_i\rangle}$$

Hence this sequence is uniformly integrable. □

We now want to show that an LAQ sequence of exponential families is LAN, but in light of Corollary 2.8, it satisfies a much stronger property, which we now define.

Definition 2.3 (Uniformly Locally Asymptotically Normal). *A sequence of statistical models $\mathcal{P}_n$ with true parameter values θ_n is* uniformly locally asymptotically normal (ULAN) *if conditions (a) and (b) of the LAN definition (Definition 2.1) hold and condition (c) of that definition is strengthened to*

(c') *There exist sequences of random vectors S_n and random almost surely positive definite matrices K_n defined on the sample space of $\mathcal{P}_n$ such that K_n converges in probability to a nonrandom positive definite matrix K and for every compact set C in $\mathbb{R}^d$*

$$\sup_{\delta \in C} \left| l_n(\theta_n + \delta, \theta_n) - \left[\delta' S_n - \tfrac{1}{2}\delta' K_n \delta\right] \right| \xrightarrow{P} 0 \tag{2.33}$$

under P_{n,θ_n}.

The remarks that follow the LAN condition also apply here, since the ULAN condition is a strengthening of the LAN condition. There is an additional issue involved in the ULAN condition. It is not clear that the supremum in (2.33) is measurable. If is is not, then the left hand side is not even a random variable and hence it is meaningless to talk about it converging in probability. It turns out that in "nice" situations the supremum is measurable[1] so we won't worry about this issue.

[1]Here is what is true. We say a subset of a Polish space is *universally measurable* if it is in the completion of the Borel σ-field for every with respect to every Borel probability measure (Bertsekas and Shreve 1978, p. 167). The class of all universally measurable sets is a σ-field

Theorem 2.10. *An LAQ sequence of exponential families is ULAN. The MLE exists with probability converging to one and is an ELE.*

The proof depends on the following lemma about minimizers of convex functions.

Lemma 2.11. *Suppose f_n and f are as in Theorem 2.7. Suppose also that f is a proper convex function and has a unique local minimizer x that is an interior point of* $\operatorname{dom} f$. *Then*

$$\inf f_n \to f(x), \qquad \text{as } n \to \infty. \tag{2.34a}$$

In addition, suppose x_n is a sequence satisfying

$$f_n(x_n) - \inf f_n \to 0, \qquad \text{as } n \to \infty. \tag{2.34b}$$

Then $x_n \to x$ as $n \to \infty$.

Proof. The assumption that f is proper makes $f(x)$ finite. It is one of the assertions of Theorem 2.7 that f is a lower semicontinuous function. Define

$$B(x, \epsilon) = \{\, y : |y - x| \le \epsilon \,\}$$

and let $S(x, \epsilon)$ be the boundary of $B(x, \epsilon)$. Choose $\epsilon > 0$ is small enough so that $B(x, \epsilon) \subset \operatorname{dom} f$,

Note that since $S(x, \epsilon)$ is a compact set and f is lower semicontinuous, the infimum over this set is achieved and must be greater than $f(x)$, say $f(x) + \delta$ with $\delta > 0$, because x is assumed to be the unique minimizer.

Then the uniformity of convergence asserted by Theorem 2.7 guarantees that for any η satisfying $0 < \eta < \delta/4$ there is an integer N such that for all $n \ge N$ the following three inequalities all hold

$$\inf_{y \in S(x,\epsilon)} f_n(y) \ge f(x) + \delta - \eta \tag{2.35a}$$

$$\inf_{y \in B(x,\epsilon)} f_n(y) \ge f(x) - \eta \tag{2.35b}$$

$$f_n(x) \le f(x) + \eta \tag{2.35c}$$

Now consider any point $z \notin B(x, \epsilon)$. Then $\lambda = \epsilon / |z - x|$ is strictly between zero and one, and $w = (1 - \lambda)x + \lambda z$ lies in $S(x, \epsilon)$. Thus the convexity inequality implies

$$f_n(w) \le (1 - \lambda) f_n(x) + \lambda f_n(z)$$

called the *universal σ-field*. Suppose S and T are Polish spaces and $f : S \times T \to \overline{\mathbb{R}}$ is a jointly Borel measurable function, then for any Borel subset B of T the function $g_B : S \to \overline{\mathbb{R}}$ defined by

$$g_B(x) = \sup_{y \in B} f(x, y), \qquad x \in S,$$

is universally measurable (not necessarily Borel measurable) (Bertsekas and Shreve 1978, Corollary 7.42.1 and Proposition 7.47). Thus so long as our sample space and parameter spaces are both Polish spaces and our likelihood is a *jointly* Borel measurable function of data and parameters, the supremum in (2.33) will be measurable with respect to the *completion* of P_{n,θ_n}, and that is all we need.

or

$$\begin{aligned} f_n(z) &\geq \frac{f_n(w) - (1-\lambda) f_n(x)}{\lambda} \\ &\geq \frac{f(x) + \delta - \eta - (1-\lambda)[f(x) + \eta]}{\lambda} \\ &= \frac{\lambda f(x) + \delta - (2-\lambda)\eta}{\lambda} \\ &\geq f(x) + \frac{2\eta}{\lambda} \end{aligned}$$

Since $\lambda < 1$ for all such z we have

$$\inf_{y \notin B(x,\epsilon)} f_n(y) \geq f(x) + 2\eta \tag{2.36}$$

Thus for $n \geq N$

$$f(x) - \eta \leq \inf f_n \leq f_n(x) \leq f(x) + \eta,$$

and, since η can be chosen as small as we please, this proves (2.34a).

Also, still for $n \geq N$, combining (2.35c) and (2.36) gives

$$f_n(z) - \inf f_n \geq f_n(z) - f_n(x) \geq \eta, \qquad z \notin B(x, \epsilon).$$

comparing with (2.34b) we see that the sequence x_n must be eventually in $B(x, \epsilon)$. Since ϵ can be chosen as small as we please, this proves $x_n \to x$. □

Proof of the theorem. Let $X_n \sim P_{n,\theta_n}$ and define μ_n and K_n as in (2.32). Then, if we define $S_n = X_n - \mu_n$, we have

$$\begin{aligned} l_n(\theta_n + \delta, \theta_n) - \left[\langle S_n, \delta\rangle - \tfrac{1}{2}\langle \delta, K_n \delta\rangle\right] \\ = k_n(\theta_n) - k_n(\theta_n + \delta) - \langle \mu_n, \delta\rangle + \tfrac{1}{2}\langle \delta, K_n \delta\rangle \end{aligned}$$

which is not random and converges uniformly to zero on compact sets by Corollary 2.8 and Lemma 2.9. Thus the sequence of models satisfies condition (c) of the ULAN definition. Condition (b), contiguity, follows from the uniform integrability assertion in Lemma 2.9, which implies that, for any bounded sequence δ_n, the likelihood

$$\exp\{l_n(\theta_n + \delta_n, \theta_n)\} = \exp\{\langle X_n - \mu_n, \delta_n\rangle + k_n(\theta_n) - k_n(\theta_n + \delta) - \langle \mu_n, \delta\rangle\}$$

is uniformly integrable, thus the measures P_{θ_n} and $P_{\theta_n + \delta_n}$ are contiguous by condition (d) of Theorem 2.1.

Recall from the discussion in Section 1.8 that the function $\hat{\theta}_n : \mathbb{R}^d \to \mathbb{R}^d \cup \{u\}$ defined by

$$\hat{\theta}_n(x) = \begin{cases} \tau_n^{-1}(x), & x \in \operatorname{int}(\operatorname{dom} k_n) \\ u, & \text{otherwise} \end{cases}$$

where $\tau_n(\theta) = \nabla k_n(\theta)$, is measurable. We define $\hat{\theta}_n(X_n)$ to be the MLE.

By Lemma 2.9 $X_n - \mu_n \to \mathcal{N}(0, K)$, so by the Skorohod theorem there are random variables

$$\begin{aligned} X_n^* &\sim P_{n,\theta_n} \\ Z &\sim \mathcal{N}(0, K) \end{aligned}$$

all defined on the same probability space such that $X_n^* - \mu_n \xrightarrow{\text{a. s.}} Z$. Then

$$\begin{aligned} l_n^*(\theta_n + \delta, \theta_n) &= \langle X_n^*, \delta \rangle - k_n(\theta_n + \delta) + k_n(\theta_n) \\ &= \langle X_n^* - \mu_n, \delta \rangle - k_n(\theta_n + \delta) + k_n(\theta_n) + \langle \mu_n, \delta \rangle \\ &\xrightarrow{\text{a. s.}} \langle Z, \delta \rangle - \tfrac{1}{2} \langle \delta, K\delta \rangle \end{aligned}$$

Defining the right hand side to be a function

$$q_Z(\delta) = \langle Z, \delta \rangle - \tfrac{1}{2} \langle \delta, K\delta \rangle$$

we see that we have almost sure pointwise convergence of the log likelihood to the random function q_Z. Since log likelihoods are concave and since q_Z has all of $\mathbb{R}^d$ for its effective domain and has a unique maximizer $K^{-1}Z$, we see that for each ω for which the almost sure convergence holds $\hat{\theta}_n\big(X_n^*(\omega)\big)$ is defined (not u) for all but finitely many n and

$$\hat{\theta}_n\big(X_n^*(\omega)\big) - \theta_n \to K^{-1}Z(\omega)$$

by Lemma 2.11. Since $K_n^{-1}[X_n^*(\omega) - \mu_n]$ also converges to the same limit the MLE is an ELE for data X_n^*. But whether an estimator is an ELE or not only depends on laws not variables, hence the MLE is an ELE for any data $X_n \sim P_{n,\theta_n}$. □

Problems

2-1. Prove the assertion about the equivalence of the two concepts of boundedness in probability made in the text. If X_1, X_2, ... are real-valued random variables, then

$$\lim_{M \to \infty} \limsup_{n \to \infty} \Pr(|X_n| \geq M) = 0 \tag{2.37a}$$

holds if and only if

$$\lim_{M \to \infty} \sup_{n \in \mathbb{N}} \Pr(|X_n| \geq M) = 0. \tag{2.37b}$$

Chapter 3

Classical Likelihood Theory

3.1 Asymptotics of Maximum Likelihood

Following Ferguson (Chapter 18) we use the notation

$$\begin{aligned} \Psi(x,\theta) &= \nabla_\theta \log f_\theta(x) \\ \dot{\Psi}(x,\theta) &= \nabla_\theta^2 \log f_\theta(x) \end{aligned} \tag{3.1}$$

If derivatives with respect to θ can be passed under the integral sign in the identity $\int f_\theta \, d\nu = 1$, then the Bartlett identities

$$E_\theta \Psi(X,\theta) = 0 \tag{3.2a}$$

$$\operatorname{Var}_\theta \Psi(X,\theta) = -E_\theta \dot{\Psi}(X,\theta) \tag{3.2b}$$

hold, in which case either side of (3.2b) is the Fisher information $I(\theta)$. We also follow Ferguson in using the notation $\dot{l}_n$ and $\ddot{l}_n$ instead of ∇l_n and $\nabla^2 l_n$, respectively. Define the matrix norm

$$\|A\|_\infty = \max_{i,j} |a_{ij}|, \tag{3.3}$$

for any matrix A with components a_{ij}.

Theorem 3.1. *Let $\{ f_\theta : \theta \in \Theta \}$ be a family of densities with respect to a measure ν, and let X_1, X_2, ... be i. i. d. with density f_{θ_0}, for some $\theta_0 \in \Theta$. Suppose*

(1) *Θ is a subset of $\mathbb{R}^d$ and a neighborhood of θ_0 in $\mathbb{R}^d$,*

(2) *second partial derivatives of $f_\theta(x)$ with respect to θ exist and are continuous on the interior of Θ for all x, and the Bartlett identities (3.2a) and (3.2b) hold at $\theta = \theta_0$,*

(3) *there exists a function* $K(x)$ *such that* $E_{\theta_0} K(X) < \infty$ *and a* $\rho > 0$ *such that* $S_\rho = \{\, \theta : |\theta - \theta_0| \le \rho \,\}$ *is contained in* Θ *and*

$$\left\| \dot{\Psi}(x, \theta) \right\|_\infty \le K(x), \qquad \text{for all } x \text{ and all } \theta \in S_\rho,$$

(4) *the Fisher information* $I(\theta_0)$ *is positive definite,*

(5) $\hat{\theta}_n$ *is a measurable function of* X_1, ..., X_n *taking values in* Θ *such that*

$$\hat{\theta}_n \xrightarrow{P} \theta_0$$

(6) *and*

$$\frac{1}{\sqrt{n}} \dot{l}_n(\hat{\theta}_n) \xrightarrow{P} 0.$$

Then

$$\sqrt{n} \left(\hat{\theta}_n - \theta_0 \right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, I(\theta_0)^{-1}\right). \tag{3.4}$$

Conditions (1) through (4) of this theorem will be referred to as the "usual" or "Cramér style" regularity conditions for maximum likelihood, after the famous book (Cramér 1946), which had the first rigorous proof of something like this theorem. Cramér's conditions actually involved third derivatives. So Ferguson again makes an error of attribution in naming this theorem after Cramér—or obeys "Stigler's law of eponomy" (no scientific discovery is named after its inventor) if you prefer to think of it that way. I do not know who introduced the line of argument that involves only first and second derivatives. Of course, the "usual" here is also misleading. There are dozens of sets of "usual" conditions that have been used in the literature, all slightly different. There are also the "Le Cam style" conditions (LAN and so forth). What's "unusual" about them? All you really can expect when someone mentions the "usual" regularity conditions is that an argument involving Taylor series is coming. But in this course we will always use "usual" to refer to conditions (1) through (4) of this theorem.

We break up the proof into several lemmas.

Lemma 3.2. *Under conditions (1) through (3) of the theorem,*

$$\sup_{\theta \in S_\rho} \left\| \tfrac{1}{n} \ddot{l}_n(\theta) - \mu(\theta) \right\|_\infty \xrightarrow{a.\ s.} 0, \tag{3.5}$$

where

$$\mu(\theta) = E_{\theta_0} \dot{\Psi}(X, \theta),$$

and μ *is continuous on* S_ρ.

Proof. Theorem 16(a) in Ferguson says that if $\dot{\Psi}_{kl}$ are the components of $\dot{\Psi}$ and μ_{kl} those of μ

$$\max_{k,l} \sup_{\theta \in S_\rho} \left| \frac{1}{n} \sum_{i=1}^{n} \dot{\Psi}_{kl}(X_i, \theta) - \mu_{kl}(\theta) \right| \xrightarrow{\text{a. s.}} 0.$$

The order of the max and sup can be interchanged (either way it is the max over both), the sum is recognized as $\frac{1}{n}\ddot{l}_n(\theta)$, and this proves (3.5). The continuity of μ is established in the proof of Theorem 16(a). □

Now we apply the fundamental theorem of calculus to the function

$$s \mapsto \dot{l}_n\big(\theta_0 + s(\theta - \theta_0)\big)$$

obtaining

$$\dot{l}_n(\theta) - \dot{l}_n(\theta_0) = \int_0^1 \ddot{l}_n\big(\theta_0 + s(\theta - \theta_0)\big)(\theta - \theta_0)\, ds. \tag{3.6}$$

Defining

$$B_n(\theta) = -\int_0^1 \frac{1}{n}\ddot{l}_n\big(\theta_0 + s(\theta - \theta_0)\big)\, ds, \tag{3.7}$$

(3.6) can be rewritten as

$$\frac{1}{\sqrt{n}}\dot{l}_n(\theta) - \frac{1}{\sqrt{n}}\dot{l}_n(\theta_0) = -B_n(\theta)\sqrt{n}(\theta - \theta_0). \tag{3.8}$$

Lemma 3.3. *Under conditions (1), (2), (3), and (5) of the theorem,*

$$B_n(\hat{\theta}_n) \xrightarrow{P} I(\theta_0), \tag{3.9}$$

where $B_n(\theta)$ is defined by equation (3.7).

Proof. Note that with μ as defined in Lemma 3.2 $\mu(\theta_0) = -I(\theta_0)$. Hence

$$\begin{aligned}
\|B_n(\hat{\theta}_n) - I(\theta_0)\|_\infty &= \left\| \int_0^1 \left[\frac{1}{n}\ddot{l}_n\big(\theta_0 + s(\hat{\theta}_n - \theta_0)\big) - \mu(\theta_0) \right] ds \right\|_\infty \\
&\le \int_0^1 \left\| \frac{1}{n}\ddot{l}_n\big(\theta_0 + s(\hat{\theta}_n - \theta_0)\big) - \mu(\theta_0) \right\|_\infty ds \\
&\le \int_0^1 \left\| \frac{1}{n}\ddot{l}_n\big(\theta_0 + s(\hat{\theta}_n - \theta_0)\big) - \mu\big(\theta_0 + s(\hat{\theta}_n - \theta_0)\big) \right\|_\infty ds \\
&\quad + \sup_{0 \le s \le 1} \left\| \mu\big(\theta_0 + s(\hat{\theta}_n - \theta_0)\big) - \mu(\theta_0) \right\|_\infty
\end{aligned}$$

The term on the bottom line converges in probability to zero by the continuity of μ and the weak consistency of $\hat{\theta}_n$. The term on the next to bottom line also converges in probability to zero because for any $\epsilon > 0$

$$\begin{aligned}
&\Pr\left(\int_0^1 \left\| \frac{1}{n}\ddot{l}_n\big(\theta_0 + s(\hat{\theta}_n - \theta_0)\big) - \mu\big(\theta_0 + s(\hat{\theta}_n - \theta_0)\big) \right\|_\infty ds > \epsilon \right) \\
&\qquad \le \Pr\left(\hat{\theta}_n \notin S_\rho \right) + \Pr\left(\sup_{\theta \in S_\rho} \left\| \frac{1}{n}\ddot{l}_n(\theta) - \mu(\theta) \right\|_\infty > \epsilon \right)
\end{aligned}$$

the first term on the right going to zero by consistency of $\hat{\theta}_n$ and the second term on the right going to zero by Lemma 3.2. Hence we have proved (3.9). □

As was noted in a homework problem the map $w : A \mapsto A^{-1}$ is continuous, even differentiable on the space of nonsingular square matrices of a particular dimension, which is an open subset of the space of all square matrices of that dimension. We want to apply this function to arbitrary matrices so we define $w(A) = 0$ when A is not invertible. This map is still continuous and differentiable at every invertible A.

Proof of Theorem 3.1. From the Bartlett identities and the central limit theorem

$$\tfrac{1}{\sqrt{n}} \dot{l}_n(\theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, I(\theta_0)\right). \tag{3.10}$$

By the comment preceding the proof, the map w is continuous at $I(\theta_0)$ because this matrix is assumed to be nonsingular. Hence the continuous mapping theorem and Lemma 3.3 imply

$$w\big(B_n(\hat{\theta}_n)\big) \xrightarrow{P} I(\theta_0)^{-1}.$$

Condition (6) of the theorem and equation (3.8) imply

$$\frac{1}{\sqrt{n}} \dot{l}_n(\theta_0) = B_n(\hat{\theta}_n)\sqrt{n}(\hat{\theta}_n - \theta_0) + o_p(1) \tag{3.11}$$

and this implies

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = w\big(B_n(\hat{\theta}_n)\big) \frac{1}{\sqrt{n}} \dot{l}_n(\theta_0) + o_p(1) \tag{3.12}$$

because $w\big(B_n(\hat{\theta}_n)\big) B_n(\hat{\theta}_n)$ is equal to the identity with probability converging to one. Now Slutsky's theorem and (3.10) imply

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} I(\theta_0)^{-1} Z \tag{3.13}$$

where $Z \sim \mathcal{N}\left(0, I(\theta_0)\right)$, and this implies (3.4). □

For any vector v with components v_i define the norms $\|v\|_\infty = \max_i |v_i|$ and $\|v\|_1 = \sum_i |v_i|$. Then for any matrix A with components a_{ij}

$$\begin{aligned} \|Av\|_\infty &= \max_i \left| \textstyle\sum_j a_{ij} v_j \right| \\ &\le \max_i \textstyle\sum_j |a_{ij}| \cdot |v_j| \\ &\le \|A\|_\infty \|v\|_1 \end{aligned}$$

The special case where A is a row vector gives $\|\langle u, v\rangle\|_\infty \le \|u\|_\infty \|v\|_1$. Also

$$\begin{aligned} |\langle v, Av\rangle| &= \left| \textstyle\sum_i \sum_j a_{ij} v_i v_j \right| \\ &\le \textstyle\sum_i \sum_j |a_{ij}| \cdot |v_i| \cdot |v_j| \\ &\le \|A\|_\infty \textstyle\sum_i \sum_j |v_i| \cdot |v_j| \\ &= \|A\|_\infty \|v\|_1^2 \end{aligned}$$

Corollary 3.4 (Theorem 18 in Ferguson). *Under conditions (1) through (4) of the theorem and condition (5) (identifiability) of Theorem 17 in Ferguson, if $\hat{\theta}_n$ is the global maximizer of the log likelihood over S_ρ, then $\hat{\theta}_n$ is a strongly consistent estimator of θ_0 and (3.4) holds.*

Proof. To show strong consistency we need to apply Theorem 17 in Ferguson, all of the conditions of that theorem obviously hold except for (3), which we have to check. We expand $s \mapsto U\bigl(x, \theta_0 + s(\theta - \theta_0)\bigr)$ in a two-term Taylor series with the integral form of remainder

$$U(x,\theta) = \langle \Psi(x,\theta_0), (\theta - \theta_0) \rangle + \int_0^1 \langle (\theta - \theta_0), \dot{\Psi}\bigl(x, \theta_0 + s(\theta - \theta_0)\bigr)(\theta - \theta_0) \rangle (1 - s)\, ds$$

Hence

$$\sup_{\theta \in S_\rho} U(x,\theta) \le \lambda \|\Psi(x,\theta_0)\|_\infty + \frac{\lambda^2}{2} K(x) \tag{3.14}$$

where $\lambda = \sup_{\theta \in S_\rho} \|\theta - \theta_0\|_1 \le \sqrt{d}\rho$. Thus condition (3) of Theorem 17 also holds and we conclude that $\hat{\theta}_n \xrightarrow{\text{a. s.}} \theta_0$. This implies condition (5) of Theorem 3.1. Also for almost all ω we have $\hat{\theta}_n(\omega)$ in the interior of S_ρ for all large enough n and this implies that $\dot{l}_n(\hat{\theta}_n(\omega))$ is zero for all large enough n, which in turn implies condition (6) of Theorem 3.1. Hence we obtain the conclusion (3.4). □

3.2 Uniformly Locally Asymptotically Normal

What is the connection between the conditions of Theorem 3.1 and local asymptotic normality (LAN) studied earlier? The conditions of Theorem 3.1 are much stronger than the LAN condition, they in fact imply the model is ULAN (Definition 2.3) and more besides.

Lemma 3.5. *Under conditions (1) through (3) of Theorem 3.1 for any $\eta > 0$*

$$\sup_{\substack{|\delta| \le \eta \\ \theta_0 + n^{-1/2}\delta \in \Theta}} \Bigl| l_n(\theta_0 + n^{-1/2}\delta) - l_n(\theta_0) - \langle \delta, S_n \rangle + \tfrac{1}{2} \langle \delta, K\delta \rangle \Bigr| \xrightarrow{P} 0, \tag{3.15}$$

where

$$S_n = \frac{1}{\sqrt{n}} \dot{l}_n(\theta_0) \tag{3.16}$$

and $K = I(\theta_0)$. Also

$$\sup_{\substack{|\delta| \le \eta \\ \theta_0 + n^{-1/2}\delta \in \Theta}} \left\| \frac{1}{\sqrt{n}} \dot{l}_n(\theta_0 + n^{-1/2}\delta) - S_n + K\delta \right\|_\infty \xrightarrow{P} 0 \tag{3.17}$$

and

$$\sup_{\substack{|\delta| \le \eta \\ \theta_0 + n^{-1/2}\delta \in \Theta}} \left\| \frac{1}{n} \ddot{l}_n(\theta_0 + n^{-1/2}\delta) + K \right\|_\infty \xrightarrow{P} 0. \tag{3.18}$$

Moreover, $P_{n,\theta_0+n^{-1/2}\delta_n}$ is contiguous to P_{n,θ_0} for any bounded sequence δ_n, where $P_{n,\theta}$ is the joint distribution of X_1, ..., X_n for true parameter value θ.

The contiguity statement and (3.15) are the ULAN conditions. Equations (3.17) and (3.18) have no analogs in LAN theory, which doesn't assume anything about derivatives of the log likelihood. Hence the "usual regularity conditions" for maximum likelihood, the conditions assumed in Theorem 3.1, are much stronger than ULAN, which in turn is much stronger than LAN.

Proof. Equation (3.18) follows from Lemma 1. For n such that $n^{-1/2}\eta \le \rho$

$$\begin{aligned}\sup_{\lvert\delta\rvert\le\eta} \left\lVert \frac{1}{n}\ddot{l}_n(\theta_0 + n^{-1/2}\delta) + K \right\rVert_\infty & \\ \le \sup_{\theta\in S_\rho} \left\lVert \tfrac{1}{n}\ddot{l}_n(\theta) - \mu(\theta) \right\rVert_\infty & + \sup_{\lvert\delta\rvert\le\eta} \left\lVert \mu(\theta_0 + n^{-1/2}\delta) + K \right\rVert_\infty .\end{aligned}$$

The first term on the right converges almost surely to zero by Lemma 3.2, and the second term on the right converges to zero by continuity of μ, which is another conclusion of Lemma 3.2.

Equation (3.8) with $\theta = \theta_0 + n^{-1/2}\delta$ is

$$\frac{1}{\sqrt{n}}\dot{l}_n(\theta_0 + n^{-1/2}\delta) - S_n = -B_n(\theta_0 + n^{-1/2}\delta)\delta$$

Plugging in the definition of $B_n(\theta)$ in (3.7) we get

$$\frac{1}{\sqrt{n}}\dot{l}_n(\theta_0 + n^{-1/2}\delta) - S_n + K\delta = \int_0^1 \left[\frac{1}{n}\ddot{l}_n\bigl(\theta_0 + sn^{-1/2}\delta\bigr) + K\right]\delta\, ds.$$

So for n such that $n^{-1/2}\eta \le \rho$

$$\begin{aligned}\sup_{\lvert\delta\rvert\le\eta} \left\lVert \frac{1}{\sqrt{n}}\dot{l}_n(\theta_0 + n^{-1/2}\delta) - S_n + K\delta \right\rVert_\infty & \\ \le \sup_{\lvert\delta\rvert\le\eta} \left\lVert \frac{1}{n}\ddot{l}_n\bigl(\theta_0 + sn^{-1/2}\delta\bigr) + K \right\rVert_\infty & \lVert\delta\rVert_1\end{aligned}$$

and the the right hand side converges in probability to zero by (3.18), which we have already established. This proves (3.17).

To prove (3.15) we need a two-term Taylor series for the function $s \mapsto l_n\bigl(\theta_0 + s(\theta - \theta_0)\bigr)$ with the integral form of the remainder

$$\begin{aligned}l_n(\theta) = l_n(\theta_0) + \Bigl\langle \dot{l}_n(\theta_0), \theta - \theta_0 \Bigr\rangle & \\ + \int_0^1 \Bigl\langle \theta - \theta_0, \ddot{l}_n\bigl(\theta_0 + s(\theta - \theta_0)\bigr)(\theta - \theta_0) \Bigr\rangle & (1 - s)\, ds\end{aligned}$$

Plugging in $\theta = \theta_0 + n^{-1/2}\delta$ gives

$$l_n(\theta_0 + n^{-1/2}\delta) = l_n(\theta_0) + \langle \delta, S_n \rangle - \tfrac{1}{2}\langle \delta, K\delta \rangle + \int_0^1 \left\langle \delta, \left[\frac{1}{n} \ddot{l}_n(\theta_0 + sn^{-1/2}\delta) + K \right] \delta \right\rangle (1 - s)\, ds.$$

So for n such that $n^{-1/2}\eta \le \rho$

$$\sup_{|\delta| \le \eta} \left| l_n(\theta_0 + n^{-1/2}\delta) - l_n(\theta_0) - \langle \delta, S_n \rangle + \tfrac{1}{2}\langle \delta, K\delta \rangle \right| \le \frac{1}{2} \sup_{|\delta| \le \eta} \left\| \frac{1}{n} \ddot{l}_n(\theta_0 + n^{-1/2}\delta) + K \right\|_\infty \|\delta\|_1^2.$$

and again the the right hand side converges in probability to zero by (3.18) proving (3.15).

The statement about contiguity is proved as follows. Define

$$Z_n = \exp\Big(l_n(\theta_0 + n^{-1/2}\delta_n) - l_n(\theta_0) \Big)$$

where δ_n is a bounded sequence. One way to prove contiguity is to show that this sequence is uniformly integrable under P_{n,θ_0}.

First assume $\delta_n \to \delta$. Then (3.15) shows that

$$Z_n = \exp\big(\langle \delta_n, S_n \rangle - \tfrac{1}{2}\langle \delta_n, K\delta_n \rangle \big) + o_p(1)$$

so $Z_n \xrightarrow{\mathcal{L}} Z$, where

$$Z = \exp\big(\langle \delta, S \rangle - \tfrac{1}{2}\langle \delta, K\delta \rangle \big)$$

and $S \sim \mathcal{N}(0, K)$.

Now we use a fact about uniform integrability (Billingsley 1968, Theorem 5.4) if $Z_n \xrightarrow{\mathcal{L}} Z$ and $E(Z_n) \to E(Z)$ and these variables are nonnegative, then the sequence Z_n is uniformly integrable. Moreover, the "Fatou lemma" for convergence in distribution (Billingsley 1968, Theorem 5.3) says that $Z_n \xrightarrow{\mathcal{L}} Z$ implies $E(Z) \le \liminf_n E(Z_n)$ if the variables are nonnegative. Thus we only need to show that $\limsup_n E(Z_n) \le E(Z)$ in order to establish uniform integrability.

Returning to our definitions of Z_n and Z,

$$E e^{\langle \delta, S \rangle} = \exp\big(\tfrac{1}{2}\langle \delta, K\delta \rangle \big)$$

(the moment generating function of a multivariate normal). Hence $E(Z) = 1$. Now

$$Z_n = \prod_{i=1}^n \frac{f_{\theta_0 + n^{-1/2}\delta_n}(X_i)}{f_{\theta_0}(X_i)}$$

and each term has expectation less than or equal to one because

$$\begin{aligned} E_{\theta_0}\left\{\frac{f_{\theta_0+n^{-1/2}\delta_n}(X_i)}{f_{\theta_0}(X_i)}\right\} &= \int_{f_{\theta_0}(x)>0} \frac{f_{\theta_0+n^{-1/2}\delta_n}(x)}{f_{\theta_0}(x)} f_{\theta_0}(x)\,\nu(dx) \\ &= \int_{f_{\theta_0}(x)>0} f_{\theta_0+n^{-1/2}\delta_n}(x)\,\nu(dx) \\ &\leq 1 \end{aligned}$$

That proves uniform integrability under the additional assumption $\delta_n \to \delta$. To finish the proof we need a "subsequence principle" for uniform integrability. Uniform integrability of the sequence Z_n is the property that for every $\epsilon > 0$, there exists and $a \in \mathbb{R}$ such that

$$\int_{Z_n \geq a} Z_n \, dP_{n,\theta_0} \leq \epsilon.$$

Assume to get a contradiction that Z_n is not uniformly integrable. Then there exists an $\epsilon > 0$ and a subsequence n_k such that

$$\int_{Z_{n_k} \geq k} Z_{n_k} \, dP_{n_k,\theta_0} \geq \epsilon.$$

Clearly, it is not possible to pick out a uniformly integrable subsubsequence. But we can always pick a subsubsequence such that $\delta_{n_{k_j}} \to \delta$ and we proved above that $Z_{n_{k_j}}$ is uniformly integrable, which is a contradiction. Hence Z_n is uniformly integrable, and we are done. □

Corollary 3.6 (The MLE is an ELE). *Any estimator $\hat{\theta}_n$ satisfying the conditions of Theorem 3.1 is an efficient likelihood estimator.*

Proof. In the proof of Theorem 3.1 we find (3.12) that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = w\big(B_n(\hat{\theta}_n)\big)S_n + o_p(1) = K^{-1}S_n + o_p(1)$$

where S_n is given by (3.16) and $K = I(\theta_0)$. The ULAN condition (3.15) implies the model is LAN, in which case any estimator of the form $K^{-1}S_n + o_p(1)$ is an ELE. □

An alternative (deeper) proof would just cite the Hájek convolution theorem. This corollary has a sort of converse.

Corollary 3.7. *Under conditions (1) through (4) of Theorem 3.1, any ELE satisfies satisfies conditions (5) and (6) of that theorem.*

Hence we can also think of the theorem as characterizing the behavior of ELE's under the "usual regularity conditions."

Proof. Let $\hat{\theta}_n$ be any ELE. This means

$$\hat{\theta}_n = \theta_0 + I(\theta_0)^{-1} \frac{1}{n} \dot{l}_n(\theta_0) + o_p(1)$$

This is a consistent estimator, hence satisfies condition (5) of Theorem 3.1. Since Lemma 3.3 applies to any consistent estimator, that is, it does not require condition (6) of Theorem 3.1, this implies $B_n(\hat{\theta}_n) \xrightarrow{P} I(\theta_0)$. Plugging in $\hat{\theta}_n$ for θ in (3.8) gives

$$\begin{aligned} \frac{1}{\sqrt{n}} \dot{l}_n(\hat{\theta}_n) - \frac{1}{\sqrt{n}} \dot{l}_n(\theta_0) &= -B_n(\hat{\theta}_n) \sqrt{n}(\hat{\theta}_n - \theta_0) \\ &= -B_n(\hat{\theta}_n) I(\theta_0)^{-1} \dot{l}_n(\theta_0) + o_p(1) \end{aligned}$$

and this implies condition (6) of Theorem 3.1. □

3.3 One-Step Updates

Recall from Section 2.5 the definition of asymptotic equivalence of estimators: $\hat{\theta}_n$ and θ_n^* are asymptotically equivalent estimators of θ_0 if

$$\theta_n^* - \hat{\theta}_n = o_p(\hat{\theta}_n - \theta_0).$$

Lemma 3.8. *Suppose*

$$\sqrt{n}\big(\hat{\theta}_n - \theta_0\big) \xrightarrow{\mathcal{L}} Z, \tag{3.19}$$

where Z is any random variable satisfying $\Pr(Z = 0) = 0$, and suppose

$$\sqrt{n}\big(\hat{\theta}_n - \theta_n^*\big) \xrightarrow{P} 0. \tag{3.20}$$

Then θ_n^* and $\hat{\theta}_n$ are asymptotically equivalent.

Proof. For any $\epsilon > 0$ and $\delta > 0$

$$\Pr\big(\|\hat{\theta}_n - \theta_n^*\| \geq \epsilon \|\hat{\theta}_n - \theta_0\|\big) \leq \Pr\big(\sqrt{n}\|\hat{\theta}_n - \theta_n^*\| > \epsilon\delta\big) + \Pr\big(\sqrt{n}\|\hat{\theta}_n - \theta_0\| \leq \delta\big).$$

The first term on the right converges to zero by (3.20) and the second term has limit superior less than or equal to $\Pr(Z \leq \delta)$ by the portmanteau theorem, and $\Pr(Z \leq \delta) \to 0$ as $\delta \to 0$ by continuity of probability. Hence the second term can be made as small as desired for all sufficiently large n, and that proves what we want. □

Theorem 3.9. *Under the conditions (1) through (4) of Theorem 3.1, if $\tilde{\theta}_n$ is an auxiliary estimator such that $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ is bounded in probability, then the one-step Newton update estimator*

$$\bar{\theta}_n = \tilde{\theta}_n - w\left(\ddot{l}_n(\tilde{\theta}_n)\right) \dot{l}_n(\tilde{\theta}_n) \tag{3.21}$$

is an ELE, where the function $w(A)$ is defined (as in the proof of Theorem 3.1) to be A^{-1} if A is nonsingular and the zero matrix otherwise.

If the Fisher information $I(\theta)$ is continuous at θ_0, then the one-step scoring update estimator

$$\theta_n^* = \tilde{\theta}_n + \frac{1}{n} I(\tilde{\theta}_n)^{-1} \dot{l}_n(\tilde{\theta}_n) \tag{3.22}$$

is also an ELE.

Proof. Lemma 3.3 applies to any consistent estimator in place of $\hat{\theta}_n$; it does not require condition (6) of Theorem 3.1. Hence it implies $B_n(\tilde{\theta}_n) \xrightarrow{P} I(\theta_0)$ and also $B_n(\bar{\theta}_n) \xrightarrow{P} I(\theta_0)$. Let $\hat{\theta}_n$ be any ELE. By Corollary 3.7, we may assume it satisfies the conditions of Theorem 3.1. So all of the formulas in the proof of that theorem hold.

Plugging in $\tilde{\theta}_n$ for θ in (3.8) gives

$$\frac{1}{\sqrt{n}} \dot{l}_n(\tilde{\theta}_n) - \frac{1}{\sqrt{n}} \dot{l}_n(\theta_0) = -B_n(\tilde{\theta}_n) \sqrt{n}(\tilde{\theta}_n - \theta_0). \tag{3.23}$$

Subtracting (3.11) from (3.23) gives

$$\frac{1}{\sqrt{n}} \dot{l}_n(\tilde{\theta}_n) = -B_n(\tilde{\theta}_n) \sqrt{n}(\tilde{\theta}_n - \theta_0) + B_n(\hat{\theta}_n) \sqrt{n}(\hat{\theta}_n - \theta_0) + o_p(1).$$

Combining with (3.22) gives

$$\theta_n^* - \hat{\theta}_n = (\tilde{\theta}_n - \hat{\theta}_n) + I(\tilde{\theta}_n)^{-1} \left[B_n(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) - B_n(\tilde{\theta}_n)(\tilde{\theta}_n - \theta_0) + o_p(n^{-1/2}) \right]$$

$I(\theta)$ is continuous by assumption. Hence $I(\tilde{\theta}_n)^{-1} B_n(\hat{\theta}_n)$ converges in probability to the identity matrix, and so does $I(\tilde{\theta}_n)^{-1} B_n(\tilde{\theta}_n)$, and this implies

$$\theta_n^* - \hat{\theta}_n = o_p(1)(\hat{\theta}_n - \theta_0) + o_p(1)(\tilde{\theta}_n - \theta_0) + o_p(n^{-1/2}) = o_p(n^{-1/2})$$

since both $\hat{\theta}_n - \theta_0$ and $\tilde{\theta}_n - \theta_0$ are $O_p(n^{-1/2})$, the first by Theorem 3.1 and the second by the assumption of this theorem. Thus $\sqrt{n}(\theta_n^* - \hat{\theta}_n) = o_p(1)$ and θ_n^* and $\hat{\theta}_n$ are asymptotically equivalent by the preceding lemma.

The proof of the asymptotic equivalence of $\bar{\theta}_n$ and $\hat{\theta}_n$ is entirely analogous. The only point where a bit more argument is required is in showing that $\frac{1}{n} \ddot{l}_n(\tilde{\theta}_n) \xrightarrow{P} -I(\theta_0)$, but this follows from Lemma 3.2 by an argument similar to those in Lemma 3.3 or Lemma 3.5. □

What this theorem says is that it is not hard to find ELEs, provided one has any root-n-consistent estimator, where "root-n-consistent" is a shorthand that describes and estimator $\tilde{\theta}_n$ satisfying the condition of the theorem

$$\tilde{\theta}_n = \theta_0 + O_p(n^{-1/2}).$$

Such estimators are easy to find in some problems and impossible in others. Method of moments estimators are always root-n-consistent (even asymptotically normal) by the delta method, provided only that they are differentiable

functions of the sample moments used and those sample moments have variances (that is, population moments exist up to twice the order of any sample moments used). Method of moments estimators typically aren't *efficient* but they are root-n-consistent. Applying a one-step Newton update or Fisher scoring update to them gives ELEs.

3.4 Good Heuristics and Good Optimization

3.4.1 Asymptotics as Heuristics

What use is asymptotic theory? Why are we interested in what happens as n goes to infinity? Where is the data set having a sample size that is going anywhere, much less to infinity? Asymptotic theory describes what happens in a mythical land "asymptopia" that has no direct connection to the real world.

In the real world, a data set has the sample size it has, and n is not going anywhere. Asymptotic theory applied to such a problem gives an asymptotic approximation that may or may not be close enough to the exact sampling distribution in question to be of any use. No theorem in asymptotics comes with a "remainder term" that bounds the error of approximation. So theory cannot tell us how good an asymptotic approximation is. It only says the approximation would be good if the sample size were very large, perhaps much larger than the sample size of the data at hand. So, strictly speaking, asymptotics says nothing at all about any real-world problem.

The accuracy of asymptotic approximations can be checked by computer simulations. Statisticians have been doing such simulations for decades. Many simulations show asymptotics working well. Other simulations show it working badly. The result of all that simulation is inconclusive. Although a few rules of thumb have been devised for simple situations (like the $n > 30$ rule for the z-test of hypotheses about the mean and the "at least 5 expected in each cell" for chi-square tests in contingency tables), even these rules are known to be wrong in some situations, and in more complicated situations, there are no rules. You can't learn anything about your particular application from simulations in other applications. Each application requires its own simulation (a. k. a., parametric bootstrap).

In summary,

> *Asymptotics are only heuristics. They provide approximations that may or may not work.*
>
> *If you are worried about accuracy, you simulate. Theory is no help.*

There are no good or bad theorems. There are only theorems (true statements with proofs) and non-theorems (statements, true, false, or undecided, without proofs, including those having asserted proofs that are incorrect, i. e. non-proofs). There are, however, good and bad heuristics. Thus it makes sense

to ask whether a theorem about asymptotics provides a good or bad heuristic, although the answer may depend on exactly what heuristic we think the theorem provides.

The one-step update theorem of the preceding section provides a bad heuristic. It seems to say that one should actually do in practice what the theorem describes, a one-step Newton or Fisher scoring update of a root-n-consistent starting point. To see why this is a bad idea, we need to look at some optimization theory.

3.4.2 Good Optimization Algorithms

Asymptotics of Optimization

This section briefly sketches an entirely different kind of asymptotics from the kind studied in the rest of the course. So for one section, just forget probability and statistics. We are maximizing a function f, called the *objective function* in optimization theory, and we do so using an algorithm that produces a sequence x_n of iterates converging to a solution x of the problem. Generally we assume that x is a local maximum of f.

The algorithm is said to converge *linearly* if

$$x_{n+1} - x = O(|x_n - x|). \tag{3.24}$$

Note that this says almost nothing about the performance of the algorithm. By itself (3.24) doesn't even imply $x_n \to x$. This is the worst type of convergence an optimization algorithm can have. The algorithm is said to converge *superlinearly* if

$$x_{n+1} - x = o(|x_n - x|) \tag{3.25}$$

and *quadratically* if

$$x_{n+1} - x = O(|x_n - x|^2). \tag{3.26}$$

As we shall see, this is the best type of convergence an optimization algorithm can be expected to have on any wide class of optimization problems.

Newton

Newton's algorithm is more commonly called the *Newton-Raphson* algorithm by statisticians, but it is so important in optimization and has so many variants, quasi-Newton, safeguarded Newton, and so forth, that the longer eponym would be cumbersome. Newton's algorithm is a method of solving simultaneous nonlinear equations. Suppose $g : \mathbb{R}^n \to \mathbb{R}^n$ is a differentiable map and we are to solve the equation $g(x) = 0$. Write $J(x) = \nabla g(x)$. $J(x)$ is an $n \times n$ matrix, generally nonsymmetric, called the *Jacobian* of the map g at the point x. At any point x_n

$$g(x) = g(x_n) + J(x_n)(x - x_n) + o(|x - x_n|).$$

Setting this to zero and ignoring higher order terms, yields

$$x = x_n - J(x_n)^{-1} g(x_n)$$

if $J(x_n)$ is nonsingular.

If the one-term Taylor expansion is a perfect approximation, this is the solution. In general it is not, but we take it to be the next point in an iterative scheme. Let x_1 be any point, and generate a sequence x_2, x_3, ... by

$$x_{n+1} = x_n - J(x_n)^{-1} g(x_n).$$

One hopes the sequence converges to the solution, but in general this is only a hope. There is usually no guarantee of convergence.

In the context of unconstrained optimization, Newton's method tries to find a zero of the gradient of the objective function f. Write $g(x) = \nabla f(x)$ and $H(x) = \nabla^2 f(x)$ for the gradient and Hessian of the objective function, then $H(x)$ is the Jacobian of $g(x)$, and the *Newton update* becomes

$$x_{n+1} = x_n - H(x_n)^{-1} g(x_n).$$

Now the Hessian is a symmetric matrix (unlike a general Jacobian).

Another way to look at Newton's algorithm applied to optimization is that it replaces the objective function f with a quadratic model

$$w_n(x) = f(x_n) + \langle x - x_n, g(x_n) \rangle + \tfrac{1}{2} \langle x - x_n, H(x_n)(x - x_n) \rangle \tag{3.27}$$

The model function w_n has no maximum unless $H(x_n)$ is negative definite. It makes no sense to accept a Newton update unless the Hessian is negative definite.

What's Good About Newton

When it converges, Newton converges superlinearly, usually quadratically.

Theorem 3.10. *Suppose x_n is a sequence of Newton iterations for maximizing an objective function f converging to a local maximum x^*. Suppose $g(x) = \nabla f(x)$ and $H(x) = \nabla^2 f(x)$ are continuous in a neighborhood of x^* and $H(x^*)$ is strictly negative definite. Then Newton is superlinearly convergent.*

Proof. By the assumptions about $g(x)$ and $H(x)$,

$$g(y) = g(x) + H(x)(y - x) + o(|y - x|) \tag{3.28}$$

holds for all x and y in some neighborhood of x^*, and

$$H(y) = H(x^*) + o(1)$$

A characterization of the Newton update is

$$0 = g(x_n) + H(x_n)(x_{n+1} - x_n). \tag{3.29}$$

Plugging in Taylor expansions around x^* for $g(x_n)$ and $H(x_n)$ in (3.29) gives

$$\begin{aligned} 0 &= g(x^*) + H(x^*)(x_n - x^*) + o(|x_n - x^*|) + [H(x^*) + o(1)]\,(x_{n+1} - x_n) \\ &= H(x^*)(x_{n+1} - x^*) + o(|x_n - x^*|) + o(|x_{n+1} - x_n|) \end{aligned}$$

because $g(x^*) = 0$. Writing $\varepsilon_n = x_n - x^*$ gives

$$\begin{aligned} 0 &= H(x^*)\varepsilon_{n+1} + o(|\varepsilon_n|) + o(|\varepsilon_{n+1} - \varepsilon_n|) \\ &= H(x^*)\varepsilon_{n+1} + o(|\varepsilon_n|) + o(|\varepsilon_{n+1}|) \\ &= H(x^*)\varepsilon_{n+1} + o(|\varepsilon_n|), \end{aligned}$$

where we have used the triangle inequality and the fact that $o(|\varepsilon_{n+1}|)$ is negligible compared to $H(x^*)\varepsilon_{n+1}$. Since $H(x^*)$ is strictly negative definite, it is invertible. This proves

$$\varepsilon_{n+1} = o(|\varepsilon_n|),$$

which is superlinear convergence (3.25). □

Quadratic convergence requires a bit more than (3.28).

Theorem 3.11. *Suppose x_n is a sequence of Newton iterations for a function f converging to a local maximum x^*. Let $g(x) = \nabla f(x)$ and $H(x) = \nabla^2 f(x)$. Suppose $H(x^*)$ is strictly negative definite, and suppose*

$$g(y) = g(x) + H(x)(y - x) + O(|y - x|^2) \tag{3.30}$$

and

$$H(y) = H(x) + O(|y - x|) \tag{3.31}$$

for all x and y in some neighborhood of x^*. Then Newton converges quadratically.

Equation (3.31) is referred to as a *Lipschitz* condition. Equation (3.30) is similar, but not usually referred to by that terminology. Both would be implied by Taylor's theorem with remainder if third derivatives of f exist and are continuous in some neighborhood of x^*.

Proof. A characterization of the Newton update is

$$0 = g(x_n) + H(x_n)(x_{n+1} - x_n).$$

Using (3.30) and (3.31) to expand around x^* gives

$$\begin{aligned} 0 = g(x^*) &+ H(x^*)(x_n - x^*) + O(|x_n - x^*|^2) \\ &+ [H(x^*) + O(|x_n - x^*|)]\,(x_{n+1} - x_n) \end{aligned}$$

Since x^* is a local min, $g(x^*) = 0$. Thus, writing $\varepsilon_n = x_n - x^*$,

$$\begin{aligned} 0 &= H(x^*)\varepsilon_{n+1} + O\left(|\varepsilon_n|^2 + |\varepsilon_n|\,|\varepsilon_{n+1} - \varepsilon_n|\right) \\ &= H(x^*)\varepsilon_{n+1} + O\left(|\varepsilon_n|^2 + |\varepsilon_n|\,|\varepsilon_{n+1}|\right) \\ &= H(x^*)\varepsilon_{n+1} + O\left(|\varepsilon_n|^2\right), \end{aligned}$$

the last equality using Theorem 3.10. Since $H(x^*)$ is invertible, this proves

$$\varepsilon_{n+1} = O\left(|\varepsilon_n|^2\right),$$

which is quadratic convergence (3.26). □

Not only does Newton converge quadratically (under fairly weak regularity conditions). Any algorithm that converges superlinearly is asymptotically equivalent to Newton.

Theorem 3.12 (Dennis-Moré). *Suppose $x_n \to x^*$ is a sequence of iterations of an optimization algorithm converging to a local maximum of f. Let $g(x) = \nabla f(x)$ and $H(x) = \nabla^2 f(x)$, and suppose $H(x^*)$ is strictly negative definite. If the algorithm converges superlinearly, then it is asymptotically equivalent to Newton, in the sense that*

$$x_{n+1} - x_n = \Delta_n + o(|\Delta_n|), \tag{3.32}$$

where

$$\Delta_n = -H(x_n)^{-1} g(x_n)$$

is the Newton step at x_n.

Proof. Write $\delta_n = x_{n+1} - x_n$ for the steps taken by the algorithm, and write $\varepsilon_n = x_n - x^*$. So $\varepsilon_{n+1} = \delta_n + \varepsilon_n$. The hypothesis of superlinear convergence is that $\varepsilon_{n+1} = o(|\varepsilon_n|)$ or that $\delta_n = -\varepsilon_n + o(|\varepsilon_n|)$. Similarly the superlinear convergence of Newton asserted by Theorem 3.10 implies $\Delta_n = -\varepsilon_n + o(|\varepsilon_n|)$, and this implies $\delta_n = \Delta_n + o(|\varepsilon_n|)$ and $\delta_n = \Delta_n + o(|\Delta_n|)$. The latter is (3.32). □

Corollary 3.13. *All superlinearly convergent algorithms are asymptotically equivalent.*

Note again that all of the results in this section have nothing to do with statistical asymptotics. They describe the performance of an optimization algorithm on a fixed objective function.

Safeguarding

In the preceding section we found that Newton or asymptotically equivalent algorithms such as quasi-Newton (Fletcher 1987, Chapter 3) are the best of all possible algorithms, asymptotically speaking. When close to convergence, they are the best one can do. The theorems say nothing at all about their performance when not close to a solution.

Newton can be really terrible algorithm. It has no guarantee of convergence, and this is not a merely theoretical problem. There are many practical problems in which Newton does fail to converge unless started close to a solution. For this reason, optimization textbooks do not recommend Newton for any problem. They always recommend some modification of Newton that has better convergence properties.

The first property that one should ask of a maximization algorithm is that each iteration go uphill, that is,

$$f(x_{n+1}) > f(x_n)$$

where x_n is the sequence of iterates and f is the objective function. Any algorithm that doesn't have this property is exceedingly bad and should never be used. Newton doesn't have this property, hence should never be used.

But merely going uphill is a fairly weak property. It is also important that the algorithm make good progress toward a solution. Consider a problem with a twice continuously differentiable objective function and hence a quadratic approximation w_n given by (3.27). The Newton update and many similar update methods are based on this quadratic approximation. Recall that the quadratic approximation only makes sense when it is strictly concave, that is, when its Hessian $H(x_n)$ is (strictly) negative definite. Moreover, the approximation is only good in a neighborhood of x_n. When x is far from x_n, the approximation may be very bad, and hence should not be used. One way to test whether we are using the approximation outside the range of its validity is to compare the *predicted increase* for a step $w_n(x_{n+1}) - w_n(x_n)$ with the *actual increase* $f(x_{n+1}) - f(x_n)$. If w_n is a good approximation to f at the point x_{n+1}, then the predicted and actual increase will be very close, and if w_n is not a good approximation we shouldn't be using it.

This leads to the following idea. Choose a constant $0 < \alpha < 1$. The value is not important, $\alpha = 1/2$ works fine. Then we impose two requirements on each iterate. First, we require the quadratic approximation w_n to be strictly concave. Second, we require

$$f(x_{n+1}) - f(x_n) \geq \alpha[w_n(x_{n+1}) - w_n(x_n)] \tag{3.33}$$

The first assures that the quadratic model w_n has a maximum. The second assures a certain amount of uphill progress.

If the Newton step satisfies both of these conditions, then we can take it. Otherwise, we need to do something else in order to have a good algorithm. This is called "safeguarded" Newton.

What to do when the Newton step doesn't satisfy these conditions? There are many possibilities. Here we outline only one that is related to "restricted step" or Levenberg-Marquardt (Fletcher 1987, Chapter 5). The basic idea is that if the quadratic approximation w_n is only good in some ball $B(x_n, \epsilon)$, we should only use it in that ball. Thus we should only use the Newton step if it makes $|x_n - x| \leq \epsilon$. Otherwise we maximize w_n over $B(x_n, \epsilon)$. The solution to this restricted problem will occur on the boundary of the ball (if it occurred in the interior, it would be the Newton step). We find the solution using the method of Lagrange multipliers, we maximize

$$\begin{aligned} w_n(x) - \lambda\langle x - x_n, x - x_n\rangle = f(x_n) + \langle x - x_n, g(x_n)\rangle \\ + \tfrac{1}{2}\langle x - x_n, [H(x_n) - \lambda I](x - x_n)\rangle \end{aligned} \tag{3.34}$$

where λ is the Lagrange multiplier. This has gradient

$$g(x_n) + [H(x_n) - \lambda I](x - x_n),$$

and solution

$$x_{n+1} = x_n - [H(x_n) - \lambda I]^{-1} g(x_n). \tag{3.35}$$

The Lagrange multiplier λ is determined by the requirement $|x_{n+1} - x_n| = \epsilon$. Generally determining the correct Lagrange multiplier is a bit difficult (Fletcher 1987, pp. 103 ff.), so we won't go into details. We will just use (3.35) as a heuristic. Suppose we consider safeguarding steps of the form (3.35) where we do not bother to choose λ to obtain a fixed step length (we generally don't know how to fix the step length anyway). We will just regard λ as a parameter chosen to get an acceptable update step.

Note that (3.34) is also a quadratic model. Its maximum occurs at x_{n+1} given by (3.35). Certainly, the Hessian $H(x_n) - \lambda I$ is negative definite if we choose λ large enough. Also note that as $\lambda \to \infty$

$$\begin{aligned} x_{n+1} - x_n &= -[H(x_n) - \lambda I]^{-1} g(x_n) \\ &= \lambda^{-1} g(x_n) + o(\lambda^{-1}) \end{aligned}$$

and hence

$$f_n(x_{n+1}) - f_n(x_n) = \lambda^{-1} |g(x_n)|^2 + O(\lambda^{-2}).$$

Thus unless the gradient $g(x_n)$ is exactly zero, this type of step goes uphill for large enough λ. It will also satisfy the sufficient progress condition (3.33) for large enough λ because w_n is close to f when x_{n+1} is close to x_n.

Thus we see that unless the gradient $g(x_n)$ is exactly zero, there is a range of λ values of the form $\lambda_0 \le \lambda < \infty$ for which steps of the form (3.33) are good steps. If the gradient is exactly zero, then we are at a local maximum, in which case the algorithm may terminate because we have a solution, or we are at a local minimum or a saddle point, in which case we have a problem, but we expect this to occur so rarely that we will not deal with it.

Safeguarding Maximum Likelihood

We now return to finding ELEs by one-step updates. The question is what does safeguarding do to the one-step update theorem (Theorem 3.9). The answer is nothing. The one-step updates described by Theorem 3.9 violate the sufficient progress condition (3.33) with probability converging to zero. Hence whatever we do in the way of safeguarding, does not affect the asymptotic properties of the estimator.

Lemma 3.14. *Under the conditions of Theorem 3.9, for $0 < \alpha < 1$, the estimators $\bar{\theta}_n$ and $\tilde{\theta}_n$ defined in that theorem satisfy*

$$l_n(\bar{\theta}_n) - l_n(\tilde{\theta}_n) \ge \alpha[w_n(\bar{\theta}_n) - w_n(\tilde{\theta}_n)] \ge 0$$

with probability converging to one, where

$$w_n(\theta) = l_n(\tilde{\theta}_n) + \left\langle \dot{l}_n(\tilde{\theta}_n), \theta - \tilde{\theta}_n \right\rangle + \frac{1}{2} \left\langle \theta - \tilde{\theta}_n, \ddot{l}_n(\tilde{\theta}_n)(\theta - \tilde{\theta}_n) \right\rangle,$$

and the estimators θ_n^* and $\tilde{\theta}_n$ defined in that theorem satisfy the analogous equations with θ_n^* replacing $\bar{\theta}_n$ and $-nI(\tilde{\theta}_n)$ replacing $\ddot{l}_n(\tilde{\theta}_n)$.

Proof. As in the proof of Theorem 3.9 we will just do the Fisher scoring case, the Newton case being very similar. In this case the quadratic model is

$$w_n(\theta) = l_n(\tilde{\theta}_n) + \left\langle \dot{l}_n(\tilde{\theta}_n), \theta - \tilde{\theta}_n \right\rangle - \frac{n}{2} \left\langle \theta - \tilde{\theta}_n, I(\tilde{\theta}_n)(\theta - \tilde{\theta}_n) \right\rangle$$

and the predicted increase is

$$w_n(\theta_n^*) - w_n(\tilde{\theta}_n) = \frac{1}{2n} \left\langle \dot{l}_n(\tilde{\theta}_n), I(\tilde{\theta}_n)^{-1} \dot{l}_n(\tilde{\theta}_n) \right\rangle$$

this is nonnegative with probability converging to one because $I(\theta_0)$ is positive definite, $I(\theta)$ is assumed continuous at θ_0, and $\tilde{\theta}_n$ is consistent.

Using a two-term Taylor series with remainder expanded about $\tilde{\theta}_n$, we see that the actual increase is

$$\begin{aligned} l_n(\theta_n^*) - l_n(\tilde{\theta}_n) &= \left\langle \dot{l}_n(\tilde{\theta}_n), \theta_n^* - \tilde{\theta}_n \right\rangle \\ &\quad + \int_0^1 \left\langle \theta_n^* - \tilde{\theta}_n, \ddot{l}_n\bigl(\tilde{\theta}_n + s(\theta_n^* - \tilde{\theta}_n)\bigr)(\theta_n^* - \tilde{\theta}_n) \right\rangle (1 - s)\, ds \\ &= w_n(\theta_n^*) - w_n(\tilde{\theta}_n) \\ &\quad + \frac{1}{2n} \left\langle \dot{l}_n(\tilde{\theta}_n), I(\tilde{\theta}_n)^{-1} D_n(\tilde{\theta}_n) I(\tilde{\theta}_n)^{-1} \dot{l}_n(\tilde{\theta}_n) \right\rangle \end{aligned}$$

where

$$D_n(\tilde{\theta}_n) = I(\tilde{\theta}_n) + \frac{2}{n} \int_0^1 \ddot{l}_n\bigl(\tilde{\theta}_n + s(\theta_n^* - \tilde{\theta}_n)\bigr)(1 - s)\, ds$$

and hence

$$\begin{aligned} \left\| D_n(\tilde{\theta}_n) \right\|_\infty &\le \left\| I(\tilde{\theta}_n) - K \right\|_\infty \\ &\quad + 2 \int_0^1 \left\| K + \frac{1}{n} \ddot{l}_n\bigl(\tilde{\theta}_n + s(\theta_n^* - \tilde{\theta}_n)\bigr) \right\|_\infty (1 - s)\, ds \end{aligned}$$

where, as usual, $K = I(\theta_0)$. The first term on the right converges in probability to zero by the root-n-consistency of $\tilde{\theta}_n$. We claim the second term also converges in probability to zero because the norm term in the integrand converges in probability to zero. This is shown as follows. For any $\epsilon > 0$ there exists an $\eta > 0$ such that

$$\Pr\{\sqrt{n}|\tilde{\theta}_n - \theta_0| \le \eta \,\text{and}\, \sqrt{n}|\theta_n^* - \theta_0| \le \eta\} \ge 1 - \epsilon$$

by the assumed root-n-consistency of $\tilde{\theta}_n$ and θ_n^*. Hence for any $M > 0$

$$\Pr\left\{ \left\| K + \frac{1}{n}\ddot{l}_n\bigl(\tilde{\theta}_n + s(\theta_n^* - \tilde{\theta}_n)\bigr) \right\|_\infty > M \right\}$$
$$\leq \epsilon + \Pr\left\{ \sup_{\substack{|\delta| \leq \eta \\ \theta_0 + n^{-1/2}\delta \in \Theta}} \left\| \frac{1}{n}\ddot{l}_n(\theta_0 + n^{-1/2}\delta) + K \right\|_\infty > \frac{M}{2} \right\}$$

and the second term on the right converges in probability to zero by (3.18). Thus we see that

$$l_n(\theta_n^*) - l_n(\tilde{\theta}_n) = w_n(\theta_n^*) - w_n(\tilde{\theta}_n) + o_p\bigl(w_n(\theta_n^*) - w_n(\tilde{\theta}_n)\bigr)$$

which implies what was to be proved. □

Thus we see that there is no reason not to use safeguarding in producing an ELE from a root-n-consistent starting point. The safeguarding will actually be used with probability that goes to zero as n goes to infinity and hence does not affect the asymptotic properties of the estimator.

If it has no effect, why use it? It has no effect in *asymptopia*. It does have a very important effect in the real world, where n isn't going anywhere.

We also note that there is no reason to stop with one iteration of Newton or Fisher scoring. Applying the theorem twice shows that two-step updates of root-n-consistent starting points are ELEs. Further iteration shows that m-step updates for any fixed m are ELEs. Thus there is no reason not to do as many iterations as we please.

3.5 When the Model is Wrong

What happens to maximum likelihood when the model is wrong? Suppose our model is $\{ f_\theta : \theta \in \Theta \}$ and the true distribution of the data has density g that is not equal to any f_θ. Let's review the theory of Section 17 in Ferguson and see what changes. Let $U(x, \theta) = \log f_\theta(x) - \log g(x)$. Then Jensen's inequality still implies that

$$\lambda(\theta) = EU(X, \theta) = \int \log \frac{f_\theta(x)}{g(x)} g(x)\nu(dx) \leq 0,$$

but we no longer know that λ achieves its maximum at θ_0 (there is no θ_0!) We are forced to add the analogous property as an additional assumption. Then the whole rest of the theory goes through.

Theorem 3.15. *Let X_1, X_2, be i. i. d. with true density g. If*

1. *Θ is compact.*
2. *$f_\theta(x)$ is upper semicontinuous in θ for each x.*

3. *For every $\theta \in \Theta$ there exists a $\rho_0 > 0$ such that*

$$\varphi(x, \theta, \rho) = \sup_{\substack{\theta' \in \Theta \\ |\theta' - \theta| < \rho}} U(x, \theta')$$

is a measurable function of x for $0 < \rho \leq \rho_0$,

4. *$E\varphi(X, \theta, \rho_0) < \infty$,*

5. *there is a $\theta^* \in \Theta$ such that*

$$\lambda(\theta) < \lambda(\theta^*), \qquad \theta \in \Theta,\ \theta \neq \theta^*$$

then, for any sequence of maximum likelihood estimates $\hat{\theta}_n$

$$\hat{\theta}_n \xrightarrow{a.\ s.} \theta^*.$$

As was mentioned in class the densities f_θ may be defective. We require $\int f_\theta \, d\nu \leq 1$, but allow strict inequality. All that is required is $\int g \, d\nu = 1$. Then the Jensen's inequality argument goes through.

3.6 Estimating Equations

Doing maximum likelihood when the model is wrong is a special case of the notion of "estimating equations." Often when people claim to be doing maximum likelihood, they do not maximize anything but merely find a point θ such that $\nabla l_n(\theta) = 0$. In fact our Theorem 3.1 only required this.

Let us generalize this to the following. We are given a vector-valued function $\Psi(x, \theta)$. If the parameter space Θ is a subset of $\mathbb{R}^d$, then $\Psi(x, \theta)$ takes values in $\mathbb{R}^d$. Given i. i. d. data $X_1, \ldots, X_n$, we form

$$u_n(\theta) = \sum_{i=1}^{n} \Psi(X_i, \theta).$$

Then u_n is a random function from $\mathbb{R}^d$ to $\mathbb{R}^d$. Since u_n represents d equations in the d unknown parameters we may be able to solve $u_n(\theta) = 0$ under certain conditions. The solution will be our estimator $\hat{\theta}_n$. ("Estimator of *what*?" the alert reader should now ask.)

Since the functions $\Psi_i(x, \theta)$ now have no connection whatsoever with the probability densities in the model, it does no good to say that $\hat{\theta}_n$ is an estimator of θ_0. As in the preceding section, there is no θ_0. However, suppose there is a point $\theta^* \in \Theta$ such that

$$E\Psi(X_i, \theta^*) = 0 \tag{3.36}$$

The "E" here refers to the true state of nature. We don't say "E_{θ_0}" because the "parameter space" Θ may have nothing to do with the distributions in

the model. When (3.36) holds we say that $\Psi(X_i, \theta)$ are *unbiased estimating equations* for θ^*. We may expect $\hat{\theta}_n$ to converge to θ^*, because

$$\tfrac{1}{n} u_n(\hat{\theta}_n) = 0$$

is an "empirical" expectation that mimics (3.36).

This is in fact just what we have done in Theorem 3.1. With Ψ as defined there, $u_n = \dot{l}_n$. $E\Psi(X_i, \theta_0) = 0$ by the first Bartlett identity. So $\dot{l}_n$ are unbiased estimating equations for θ_0 and indeed under the "usual regularity conditions" we get the asymptotic normality result (3.4).

Now we want to drop the connection between $\Psi(x, \theta)$ and the likelihood function. We let Ψ be an arbitrary function, and impose only enough regularity conditions to let an argument analogous to the proof of Theorem 3.1 go through.

The first problem is that the Bartlett identities are gone. The first Bartlett identity (3.2a) is replaced by the requirement (3.36) that the estimating equations be unbiased. The second Bartlett identity has no analog. In fact $\dot{\Psi}(x, \theta)$ will no longer be a symmetric matrix, since it is no longer a matrix of second derivatives of a scalar function, but a matrix of first derivatives of a vector function. The analogs of the two sides of the second Bartlett identity (3.2b) are

$$\Sigma = \operatorname{Var} \Psi(X_i, \theta^*) \tag{3.37a}$$

$$J = E\dot{\Psi}(X_i, \theta^*) \tag{3.37b}$$

Being a covariance matrix, Σ is symmetric and positive semidefinite, but J need not even be symmetric. There is no possibility of equality between these two matrices. For this reason we have not put a minus sign in (3.37b) as occurs in (3.2b). There would be no point. We see now, that in the usual theory of maximum likelihood, the Fisher information $I(\theta_0)$ plays two roles, that of Σ and that of $-J$. We can carry through the argument as long as we are careful to separate these two roles.

Theorem 3.16. *Let X_1, X_2, ... be i. i. d., and let $\Psi(x, \theta)$ be estimating equations. Suppose*

(1) *Θ is a subset of $\mathbb{R}^d$ and a neighborhood of θ^* in $\mathbb{R}^d$,*

(2) *partial derivatives of $\Psi(x, \theta)$ with respect to θ exist and are continuous on the interior of Θ for all x,*

(3) *the estimating equations are unbiased (3.36), the expectations in (3.37a) and (3.37b) exist, Σ and J are nonsingular,*

(4) *there exists a function $K(x)$ such that $EK(X_i) < \infty$ and a $\rho > 0$ such that $S_\rho = \{\, \theta : |\theta - \theta^*| \le \rho \,\}$ is contained in Θ and*

$$\left\| \dot{\Psi}(x, \theta) \right\|_\infty \le K(x), \qquad \text{for all } x \text{ and all } \theta \in S_\rho,$$

(5) *$\hat{\theta}_n \xrightarrow{P} \theta^*$,*

(6) *and*

$$\frac{1}{\sqrt{n}} \dot{u}_n(\hat{\theta}_n) \xrightarrow{P} 0.$$

Then

$$\sqrt{n} \left(\hat{\theta}_n - \theta^* \right) \xrightarrow{\mathcal{L}} \mathcal{N} \left(0, J^{-1} \Sigma (J^T)^{-1} \right). \tag{3.38}$$

The proof is the same, word for word, as that of Theorem 3.1 with $\dot{l}_n$ and $\ddot{l}_n$ replaced by u_n and $\dot{u}_n$, respectively, θ_0 is replaced by θ^*, and $I(\theta_0)$ replaced by $-J$ everywhere except in (3.10), where it is replaced by Σ, all the way down to the last sentence, which is changed to: Now Slutsky's theorem and (3.10) [with Σ replacing $I(\theta_0)$] imply (3.13) [with $-J$ replacing $I(\theta_0)$], that is

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{\mathcal{L}} -J^{-1} Z$$

where $Z \sim \mathcal{N}(0, \Sigma)$, and this implies (3.38).

3.6.1 The Sandwich Estimator

If we can do the integrals in (3.37a) and (3.37b) we may use those integrals with $\hat{\theta}_n$ plugged in for θ^* to estimate Σ and J. Often we cannot do the integrals or do not wish to specify a "true" distribution of the data to be used in computing these expectations. Then we must estimate Σ and J. The natural estimate of Σ is

$$\widehat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^{n} \Psi(X_i, \hat{\theta}_n)^2$$

and the natural estimate of J is

$$\widehat{J}_n = \frac{1}{n} \sum_{i=1}^{n} \dot{\Psi}(X_i, \hat{\theta}_n)$$

and our natural estimate of $J^{-1}\Sigma(J^T)^{-1}$ is $\widehat{J}_n^{-1} \widehat{\Sigma}_n (\widehat{J}_n^T)^{-1}$. This is often referred to as the "sandwich estimator."

Use of the sandwich estimator results in a procedure that is partially nonparametric. The estimating equations are parametric, but the variance estimate is nonparametric. The only model assumptions in Theorem 3.16 are that $\Psi(X, \theta^*)$ has finite variance, $\dot{\Psi}(X, \theta^*)$ has finite expectation, and $K(X)$ has finite expectation. The class of distributions that satisfy these conditions is indeed nonparametric, too large to be continuously indexed by a subset of a finite-dimensional vector space.

Chapter 4

Likelihood Ratio and Related Tests

4.1 The Wilks Theorem

In this section we take a first pass at the theorem that twice the log likelihood ratio is asymptotically chi-squared. We will take the following as our basic "regularity condition."

Definition 4.1 (ULAN at rate α_n). *A sequence of statistical models*

$$\mathcal{P}_n = \{ P_{n,\theta} : \theta \in \Theta \}$$

is said to satisfy the ULAN conditions at rate α_n, where α_n is a sequence of positive numbers converging to zero, if

(a) *Θ is a neighborhood of θ_0 in $\mathbb{R}^d$.*

(b) *For any bounded sequence δ_n in $\mathbb{R}^d$, the sequences P_{n,θ_0} and $P_{n,\theta_0+\alpha_n\delta_n}$ are contiguous.*

(c) *There exist sequences of random vectors S_n and random almost surely positive definite matrices K_n defined on the sample space of $\mathcal{P}_n$ such that K_n converges in probability to a nonrandom positive definite matrix K and for every compact set C in $\mathbb{R}^d$*

$$\sup_{\substack{\delta \in C \\ \theta_0 + \alpha_n \delta \in \Theta}} \left| l_n(\theta_0 + \alpha_n\delta, \theta_0) - \left[\langle \delta, S_n \rangle - \tfrac{1}{2} \langle \delta, K_n \delta \rangle \right] \right| \xrightarrow{P} 0 \qquad (4.1)$$

under P_{n,θ_0}, where $l_n(\theta, \theta_0)$ is the log likelihood ratio comparing θ and θ_0.

We know from Lemma 3.5 that under the "usual regularity conditions" for the i. i. d. case the model is ULAN at rate $\alpha_n = n^{-1/2}$. The specific rate plays no role in what follows, so we gain a little bit of generality by allowing an arbitrary rate sequence α_n.

Theorem 4.1. *Suppose the conditions of Definition 4.1. Let $\hat{\theta}_n$ be an ELE for the full model and let θ_n^* be an ELE for the restricted model that fixes the first r components of θ, then*

$$L_n = 2\big(l_n(\hat{\theta}_n) - l_n(\theta_n^*)\big). \tag{4.2}$$

has an asymptotic $\text{chi}^2(r)$ *distribution, where $l_n(\theta)$ is any log likelihood for the problem.*

The last phrase of the theorem refers to the fact that we do not know θ_0, hence do not know $l_n(\theta, \theta_0)$. If we can write $l_n(\theta, \theta_0)$ as a difference $l_n(\theta) - l_n(\theta_0)$ then the $l_n(\theta_0)$ part will cancel in calculating the log likelihood ratio.

Of course the ELE for the large model is

$$\hat{\theta}_n = \theta_0 + \alpha_n K_n^{-1} S_n + o_p(1).$$

For the small model we have to first convince ourselves that this model is ULAN and determine what the ELE is in that model. It will help if we adopt the "partitioned" matrix and vector notation, writing vectors as

$$\delta = \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}$$

and

$$K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}$$

and so forth. Here we want the top part of the partition to have r rows and the bottom part to have $d - r$ rows (and similarly for columns of matrices). The small model fixes θ_1, the first part of the partition, so the corresponding part of δ, that is, δ_1 is zero. Thus we have

$$\begin{aligned} \langle S_n, \delta \rangle &= \langle S_{n,2}, \delta_2 \rangle \\ \langle \delta K_n, \delta \rangle &= \langle \delta_2, K_{n,22} \delta_2 \rangle \end{aligned}$$

Thus it is easy to see that the ULAN property transfers from the larger to the smaller model and that

$$\theta_n^* = \begin{pmatrix} 0 \\ \theta_{n,2}^* \end{pmatrix}$$

where

$$\theta_{n,2}^* = \theta_{0,2} + \alpha_n K_{n,22}^{-1} S_{n,2} + o_p(1).$$

Hence

$$\theta_n^* = \theta_0 + \alpha_n H_n S_n + o_p(1). \tag{4.3}$$

where

$$H_n = \begin{pmatrix} 0 & 0 \\ 0 & K_{n,22}^{-1} \end{pmatrix} \tag{4.4}$$

Proof. The ULAN condition and Corollary 2.5 imply that

$$\hat{\delta}_n = \alpha_n^{-1}\big(\hat{\theta}_n - \theta_0\big)$$

is bounded in probability, hence for any $\epsilon_1 > 0$ there is a compact set C such that $\Pr(\hat{\delta}_n \notin C) \le \epsilon_1$ for all n. By the ULAN condition for any $\epsilon_2 > 0$ and $\epsilon_3 > 0$ there is an N_2 such that

$$\Pr\left(\sup_{\substack{\delta\in C\\ \theta_0+\alpha_n\delta\in\Theta}} \big|l_n(\theta_0+\alpha_n\delta,\theta_0) - \big[\langle\delta, S_n\rangle - \tfrac{1}{2}\langle\delta, K_n\delta\rangle\big]\big| > \epsilon_3\right) \le \epsilon_2$$

for all $n \ge N_2$. Hence

$$\Pr\left(\Big|l_n(\hat{\theta}_n,\theta_0) - \Big[\langle\hat{\delta}_n, S_n\rangle - \tfrac{1}{2}\langle\hat{\delta}_n, K_n\hat{\delta}_n\rangle\Big]\Big| > \epsilon_3\right) \le \epsilon_1 + \epsilon_2$$

for all $n \ge N_2$, and this implies that

$$l_n(\hat{\theta}_n,\theta_0) - \Big(\langle\hat{\delta}_n, S_n\rangle - \tfrac{1}{2}\langle\hat{\delta}_n, K_n\hat{\delta}_n\rangle\Big)$$

converges in probability to zero, that is,

$$\begin{aligned} l_n(\hat{\theta}_n,\theta_0) &= \langle\hat{\delta}_n, S_n\rangle - \tfrac{1}{2}\langle\hat{\delta}_n, K_n\hat{\delta}_n\rangle + o_p(1)\\ &= \langle S_n, K_n^{-1}S_n\rangle - \tfrac{1}{2}\langle K_n^{-1}S_n, K_nK_n^{-1}S_n\rangle + o_p(1)\\ &= \tfrac{1}{2}\langle S_n, K_n^{-1}S_n\rangle + o_p(1) \end{aligned} \tag{4.5}$$

Similarly

$$l_n(\theta_n^*,\theta_0) = \langle S_n, H_nS_n\rangle - \tfrac{1}{2}\langle S_n, H_nK_nH_nS_n\rangle + o_p(1) \tag{4.6}$$

where H_n is given by (4.4). Subtracting (4.6) from (4.5) and multiplying by two gives

$$2\big(l_n(\hat{\theta}_n) - l_n(\theta_n^*)\big) = \langle S_n, A_nS_n\rangle + o_p(1) \tag{4.7}$$

where

$$A_n = K_n^{-1} - 2H_n + H_nK_nH_n \tag{4.8}$$

Thus by the assumptions in the ULAN condition and Corollary 2.5 (4.7) converges in law to $\langle S, AS\rangle$, where

$$S \sim \mathcal{N}(0,K)$$
$$A = K^{-1} - 2H + HKH$$

and

$$H = \begin{pmatrix} 0 & 0\\ 0 & K_{22}^{-1}\end{pmatrix}$$

It only remains to be shown that this has the chi-square distribution asserted by the theorem.

First note that

$$HK = \begin{pmatrix} 0 & 0 \\ 0 & K_{22}^{-1} \end{pmatrix} \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22}^{-1} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ K_{22}^{-1} K_{21} & I \end{pmatrix} \tag{4.9}$$

where I denotes the identity matrix, so

$$HKH = \begin{pmatrix} 0 & 0 \\ K_{22}^{-1} K_{21} & I \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & K_{22}^{-1} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & K_{22}^{-1} \end{pmatrix} = H \tag{4.10}$$

Hence $A = K^{-1} - H$. Now we want to use the lemma about the chi-square distribution that is Exercise 2 of Section 9 in Ferguson: $\langle X, PX \rangle$ is $\text{chi}^2(r)$ if $X \sim \mathcal{N}(0, I)$ and P is a projection of rank r. To put what we have in this form define $X = K^{-1/2} S$ and

$$P = K^{1/2} A K^{1/2} = I - K^{1/2} H K^{1/2} \tag{4.11}$$

so $\langle X, PX \rangle = \langle S, AS \rangle$ and X is as specified in the lemma. Thus it only remains to be proved that (4.11) is a projection of rank r. Since $HKH = H$ by (4.10),

$$\begin{aligned} P^2 &= I - 2K^{1/2} H K^{1/2} + K^{1/2} H K^{1/2} K^{1/2} H K^{1/2} \\ &= I - 2K^{1/2} H K^{1/2} + K^{1/2} H K H K^{1/2} \\ &= I - 2K^{1/2} H K^{1/2} + K^{1/2} H K^{1/2} \\ &= P \end{aligned}$$

and P is a projection. Hence all its eigenvalues are zero or one and the trace calculates the rank. Now

$$\begin{aligned} \text{tr}(P) &= \text{tr}(I) - \text{tr}(K^{1/2} H K^{1/2}) \\ &= \text{tr}(I) - \text{tr}(HK) \end{aligned}$$

and the trace of the identity is d and the matrix HK given in (4.9) is clearly the trace of the identity matrix in the 2-2 block of the partition, which is $d - r$. Thus the rank of P is $d - (d - r) = r$, and we are done. □

4.2 The Rao Test

Theorem 4.2. *Suppose conditions 1 through 4 of Theorem 3.1 (the "usual regularity conditions" for maximum likelihood) and suppose the Fisher information $I(\theta)$ is continuous at θ_0. Let $\hat{\theta}_n$ be an ELE for the full model and let θ_n^* be an ELE for the restricted model that fixes the first r components of θ, then*

$$R_n = \frac{1}{n} \langle \dot{l}_n(\theta_n^*), I(\theta_n^*)^{-1} \dot{l}_n(\theta_n^*) \rangle \tag{4.12}$$

is asymptotically equivalent to the likelihood ratio test statistic (4.2).

Proof. The "usual regularity conditions" imply those of Lemma 3.5 hence equation (3.17) of that lemma holds with S_n and K_n as in the lemma. This implies by an argument similar to the beginning of the proof of the Wilks theorem

$$\frac{1}{\sqrt{n}} \dot{l}_n(\theta_n^*) = S_n + K_n \sqrt{n}(\theta_n^* - \theta_0) + o_p(1)$$

and by consistency of θ_n^* and the assumed continuity of $I(\theta)$

$$I(\theta_n^*)^{-1} = K_n^{-1} + o_p(1)$$

Since the model is ULAN by Lemma 3.5 the analysis preceding the proof of the Wilks theorem holds with $\alpha_n = n^{-1/2}$ and from (4.3)

$$\sqrt{n}(\theta_n^* - \theta_0) = H_n S_n + o_p(1)$$

where H_n is given by (4.4). Hence

$$\frac{1}{\sqrt{n}} \dot{l}_n(\theta_n^*) = (I - K_n H_n) S_n + o_p(1)$$

and

$$R_n = \langle S_n, A_n S_n \rangle + o_p(1)$$

where A_n is given by (4.8). Comparison with (4.7) shows the two test statistics are asymptotically equivalent, so we are done. □

The test using the statistic R_n called the *Rao test* or the *score test* or the *Lagrange multiplier test*. An important point about the Rao test statistic is that, unlike the likelihood ratio test statistic, it only depends on the MLE for the null hypothesis θ_n^*. The MLE for the alternative hypothesis $\hat{\theta}_n$ does not have to be computed. This allows us to do a test asymptotically equivalent to the likelihood ratio test when only θ_n^* is computable.

4.3 The Wald Test

The Wald test is complementary to the Rao test. The Rao test depends only on the MLE for the null hypothesis; the Wald test depends only on the MLE for the alternative hypothesis $\hat{\theta}_n$.

Theorem 4.3. *Suppose the conditions of Definition 4.1 with $\alpha_n = n^{-1/2}$. Let $\hat{\theta}_n$ be an ELE for the full model. Let G be the constant $r \times d$ matrix of the form $G = \begin{pmatrix} I & 0 \end{pmatrix}$ where I is the $r \times r$ identity matrix. Then*

$$W_n = n \left\langle G(\hat{\theta}_n - \theta_0), (G K_n^{-1} G')^{-1} G(\hat{\theta}_n - \theta_0) \right\rangle \tag{4.13}$$

is asymptotically equivalent to the likelihood ratio test statistic (4.2) *or the Rao test statistic* (4.12).

The proof is left as an exercise. Here we will just explain a bit more fully what the test statistic is and give a simple proof that it is asymptotically $\text{chi}^2(r)$ distributed.

Although the theorem does not explicitly mention this, the intended null hypothesis is obviously the same as for the Wilks and Rao theorems, that is, it fixes the first r components of the parameter θ. Otherwise why use an asymptotically equivalent test statistic?

The matrix G "picks off" these components; $G\hat{\theta}_n$ is just the vector of length r that is the first r components of the MLE. These are fixed under the null hypothesis H_0, so their value in under the alternative is a sensible test statistic. Note also that although θ_0 is unknown even under the H_0 (because we have a compound null hypothesis), $G\theta_0$ is known, being the first r components, which are fixed under H_0. Thus despite first appearances $G(\hat{\theta}_n - \theta_0)$ is a statistic under H_0

We know from Corollary 2.5 that

$$\sqrt{n}\big(\hat{\theta}_n - \theta_0\big) \xrightarrow{\mathcal{L}} Z$$

where $Z \sim \mathcal{N}(0, K^{-1})$. Then by the delta method

$$\sqrt{n}G\big(\hat{\theta}_n - \theta_0\big) \xrightarrow{\mathcal{L}} GZ$$

Write $X = GZ$, then $X \sim \mathcal{N}(0, GK^{-1}G')$. The variance $GK^{-1}G'$ here is a nonsingular $r \times r$ matrix. Then from Lemma 1 of Section 9 in Ferguson

$$X'\big(GK^{-1}G'\big)^{-1}X = Z'G'\big(GK^{-1}G'\big)^{-1}GZ$$

has a $\text{chi}^2(r)$ distribution. And this is clearly the limiting distribution of the Wald test statistic W_n given by (4.13).

Chapter 5

Some Differential Geometry

5.1 Morphisms

A function $g : U \to \mathbb{R}^d$ where U is an open set in $\mathbb{R}^k$ is said to be C^p for some positive integer p if it is p times continuously differentiable on U. We also say that g is C^0 if it is merely continuous and C^∞ if it is infinitely differentiable. We also use the terms *map* and *morphism* as synonyms for "function." So "C^p morphism" means a function that has p continuous derivatives.

A map $g : U \to V$, where U and V are open subsets of $\mathbb{R}^d$, is called a C^p *isomorphism* if g is invertible and both g and g^{-1} are C^p morphisms. Such a map is called a *local C^p isomorphism* at a point x if U is a neighborhood of x.

An important tool in differential geometry (and many other areas of mathematics) is the inverse function theorem.

Theorem 5.1 (Inverse Functions). *If $g : U \to V$, where U and V are open subsets of $\mathbb{R}^d$, is a C^p morphism with $p \geq 1$ and $\nabla g(x)$ is invertible at some point $x \in U$, then there exists an open neighborhood W of x in U such that the restriction of g to W is a C^p isomorphism, and*

$$\nabla g^{-1}(y) = [\nabla g(x)]^{-1} \tag{5.1}$$

where $y = g(x)$.

No proof is given here. The statement we give here is taken from Lang (1993, Theorem 1.2 of Chapter XIV). Higher derivatives of g^{-1} are obtained by applying the chain rule and the product rule to the formula for the derivative of an matrix inverse.

Thus in order to check that g is a C^p isomorphism it is only necessary to check

- g has an inverse function,
- g is C^p, and

- $\nabla g(x)$ is nonsingular at each x.

Then the fact that g^{-1} is p times continuously differentiable follows from the inverse mapping theorem.

A C^p isomorphism provides a global change of coordinates for $\mathbb{R}^d$. A local C^p isomorphism provides a local change of coordinates.

5.2 Manifolds

We begin with the definition of "manifold" that we will use.[1]

Definition 5.1 (Manifold). *A subset M of $\mathbb{R}^d$ is said to be a* k-dimensional C^p manifold *if for every $x \in M$ there is a local C^p isomorphism $g : U \to V$ at x such that*

$$M \cap U = g^{-1}(S \cap V)$$

for some k-dimensional subspace S of $\mathbb{R}^d$.

The idea of the definition is that M may be a curved subset of $\mathbb{R}^d$ but at every point x there is a local change of coordinates that makes it flat, $V \cap S$ being an open subset of the subspace S.

It may not be possible to find a global change of coordinates that does the same job. The surface of a sphere in $\mathbb{R}^3$ is a 2-dimensional manifold. At every point of this manifold there is a local change of coordinates that flattens it out, but there can be no global change of coordinates that does the same job, because the sphere is a compact set with no boundary, and no compact set in $\mathbb{R}^2$ has that property.

A function $g : X \to Y$ between topological spaces is a *homeomorphism* (also called *topological isomorphism* or C^0 *isomorphism*) if it is invertible and both g and g^{-1} are continuous. By definition of continuity, this means that both g and g^{-1} map open sets to open sets.

If X is any topological space and Y is a subset of X, then the standard way to define a topology for Y is to declare that a set is open in Y if and only if it is of the form $Y \cap W$ for some open set W in X. Sets that are open in Y are typically not open in X. Sometimes the terminology "relatively open" is used to indicate "open in Y" as opposed to "open in X." When the topology of Y is defined this way we say Y is a *topological subspace* of X (as opposed to being just a subset of X) and the topology of Y is the *subspace topology*.

Using this definition we see that M is a topological subspace of $\mathbb{R}^d$, and a subset of M is (relatively) open in M if and only if it is of the form $M \cap W$ for

[1] This is less general than the abstract definition found in differential geometry books in two respects. First, what we are really definining is a k-dimensional submanifold of $\mathbb{R}^d$. The definition of k-dimensional manifold in differential geometry books defines it as an intrinsic object with no reference to an enclosing space $\mathbb{R}^d$. However this makes the description of the tangent space much more abstract and harder to visualize. Thus our approach. Second, what we describe is sometimes called a manifold "without boundary" in contrast to a more general notion of a manifold "with boundary" that models curved subsets of $\mathbb{R}^d$ with "edges."

some open set W in $\mathbb{R}^d$. Furthermore, if g is a local isomorphism having the properties asserted in the definition of "manifold" then g is a homeomorphism from U to V, and this implies that its restriction to $M \cap U$ is a homeomorphism from $M \cap U$ to $S \cap V$. To see this consider a relatively open set in $M \cap U$, which is necessarily of the form $M \cap U \cap W$ for some set W open in $\mathbb{R}^d$. This is mapped by g to $S \cap V \cap g(W)$, which is relatively open in $S \cap V$ because g maps open sets to open sets so $g(W)$ is open. The argument about g^{-1} is exactly the same.

It is always possible to choose the C^p isomorphism g in the definition of "manifold" so that the k-dimensional subspace S is $\mathbb{R}^k$. Then if we write

$$g(x) = \big(g_1(x), \ldots, g_d(x)\big),$$

then

$$x \longleftrightarrow \big(g_1(x), \ldots, g_k(x)\big)$$

is a one-to-one correspondence between the relatively open set $M \cap U$ of the manifold and the relatively open set $\mathbb{R}^k \cap V$ of $\mathbb{R}^k$, and this correspondence is a homeomorphism. Thus a manifold is topologically equivalent to $\mathbb{R}^k$ locally.

Lemma 5.2 (Change of Local Coordinates). *If M is a k-dimensional C^p manifold in $\mathbb{R}^d$ and $g : U \to V$, where U and V are open in $\mathbb{R}^d$, is a local C^p isomorphism at x that maps $M \cap U$ bijectively onto $\mathbb{R}^k \cap V$ and $h_1 : \mathbb{R}^k \cap V \to \mathbb{R}^k$ is another local C^p isomorphism, define a map $h : V \to \mathbb{R}^d$ by*

$$h(y_1, y_2) = \big(h_1(y_1), y_2\big),$$

then there exists an open set U_1 in $\mathbb{R}^d$ such that $x \in U_1 \subset U$ and the restriction of $h \circ g$ to U_1 is a local C^p isomorphism at x that maps $M \cap U_1$ bijectively onto $\mathbb{R}^k \cap W_1$, where $W_1 = h[g(M \cap U_1)]$.

Proof. Since

$$\nabla h(y_1, y_2) = \begin{pmatrix} \nabla h_1(y_1) & 0 \\ 0 & I \end{pmatrix}$$

it has full rank and the inverse function theorem says there is an open set V_1 in $\mathbb{R}^d$ such that the restriction of h to V_1 is a local C^p isomorphism. Now let $U_1 = g^{-1}(V_1)$ and $W_1 = h(V_1)$. Then the restriction of $h \circ g$ to U_1 is a C^p isomorphism because the composition of C^p isomorphisms is another C^p isomorphism by the chain rule, and it maps $M \cap U_1$ bijectively onto $\mathbb{R}^k \cap W_1$ by construction. □

Note that in the lemma g maps d-vectors to d-vectors and h_1 maps k-vectors to k-vectors. The lemma says that once we have found one set of local coordinates any C^p isomorphism h_1 of the local coordinates produces another set of local coordinates.

5.3 Constructing Manifolds

There are two ways in which manifolds naturally arise, either as the solution set of a constraint function or as the image of a mapping. To be a bit more precise, we say $h : U \to \mathbb{R}^{d-k}$ with U an open set in $\mathbb{R}^d$ and $0 < k < d$ is a "constraint function" if we use it to define the set

$$M = \{\, x \in U : h(x) = 0 \,\}$$

Under certain conditions, explained below, M is a k-dimensional manifold. The other method uses a function $h : U \to \mathbb{R}^d$ with U an open set in $\mathbb{R}^k$, then the set

$$M = \{\, h(x) : x \in U \,\}$$

is again a k-dimensional manifold if certain conditions are satisfied.

Lemma 5.3 (Construction of Manifolds I). *Suppose $h : U \to \mathbb{R}^{d-k}$, with U an open set in $\mathbb{R}^k$ and $0 < k < d$, is a C^p mapping, $p \geq 1$, such that $\nabla h(x)$ is surjective for every $x \in U$, then the set*

$$M = \{\, x \in U : h(x) = 0 \,\}$$

is either empty or a k-dimensional C^p manifold.

In matrix terminology, the hypothesis that $\nabla h(x)$ is surjective is the same as saying it has full rank.

Proof. Fix $x \in M$ such that $h(x) = 0$ (if there is no such x the theorem is trivially true).

Write $\mathbb{R}^d = V_1 + V_2$, where V_1 is the null space of $\nabla h(x)$ and V_2 is any complementary subspace, so that any $x \in \mathbb{R}^d$ has a unique representation $x = x_1 + x_2$ with $x_i \in V_i$. This correspondence $x \leftrightarrow (x_1, x_2)$ sets up a linear isomorphism between $\mathbb{R}^d$ and $V_1 \times V_2$. The dimension of V_1 is k and the dimension of V_2 is $d - k$. Consider the map

$$g : (x_1, x_2) \mapsto \big(x_1, h(x)\big)$$

The derivative of this map can be written as the partitioned matrix

$$\nabla g(x) = \begin{pmatrix} I & 0 \\ 0 & \nabla_2 h(x) \end{pmatrix}$$

where $\nabla_2 h(x)$ denotes the partial derivative of h with respect to x_2, the partial derivative with respect to x_1 being zero by definition of "null space." Since

$$\nabla h(x) = \begin{pmatrix} 0 & \nabla_2 h(x) \end{pmatrix}$$

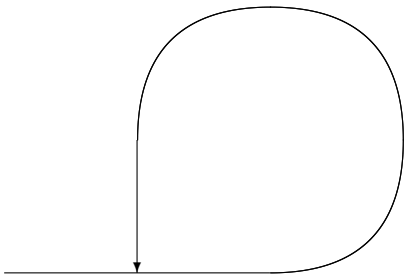
has rank k by the hypothesis of the theorem, $\nabla_2 h(x)$ has rank k and is invertible. Thus $\nabla g(x)$ is also invertible, hence g is a local C^p isomorphism by the inverse mapping theorem. By definition, $(x_1, x_2) \in M$ if and only $g(x_1, x_2) = (x_1, 0)$, thus g provides a one-to-one mapping of a neighborhood of x in M to a neighborhood of 0 in V_1. Hence M is a $d - k$ dimensional C^p manifold. □

Lemma 5.4 (Construction of Manifolds II). *Suppose $h : U \to \mathbb{R}^d$, with U a nonempty open set in $\mathbb{R}^k$ and $0 < k < d$, is a C^p mapping, $p \geq 1$, such that $\nabla h(x)$ is injective for every $x \in U$, then the set*

$$M = \{\, h(x) : x \in U \,\}$$

is a k-dimensional C^p manifold provided that h is a homeomorphism from U to M.

The additional condition that h is a homeomorphism rules out the kind of nonsense shown in the figure. The arrow means the vertical part of the curve



extends to the intersection but does not contain the point of intersection. This curve behaves like a 1-dimensional manifold locally but not globally.

Proof. Let V_1 be the range of $\nabla h(x)$ and V_2 any complementary subspace, so that any $y \in \mathbb{R}^d$ has a unique representation $y = y_1 + y_2$ with $y_i \in V_i$. This correspondence $y \leftrightarrow (y_1, y_2)$ sets up a linear isomorphism between $\mathbb{R}^d$ and $V_1 \times V_2$. The dimension of V_1 is k and the dimension of V_2 is $d - k$. Consider the map

$$q : U \times V_2 \to V_1 \times V_2$$

defined by

$$q(x, y_2) \mapsto \big(h_1(x), h_2(x)\big) + (0, y_2).$$

Fix $y_0 \in M$. Then $y_0 = h(x_0)$ for some $x_0 \in U$. Note that $q(x_0, 0) = y_0$ and

$$\nabla q(x_0, 0) = \begin{pmatrix} \nabla h_1(x_0) & 0 \\ \nabla h_2(x_0) & I \end{pmatrix}$$

Since

$$\nabla h(x_0) = \begin{pmatrix} \nabla h_1(x_0) \\ \nabla h_2(x_0) \end{pmatrix}$$

maps onto the first component by the definition of V_1, it must be that $\nabla h_2(x) = 0$ and hence $\nabla h_1(x)$ has full rank, from which it follows that $\nabla q(x, 0)$ also has full rank. Hence by the inverse function theorem there exists an open set W in $U \times V_2$ containing $(x_0, 0)$ such that q restricted to W is a local C^p isomorphism at $(x_0, 0)$. Let $\bar{q}$ denote this restriction. Then $\bar{q}^{-1}$ is a C^p isomorphism from

$q(W)$ to W. So $\bar{q}$ is a local C^p isomorphism at $(x_0, 0)$ and $\bar{q}^{-1}$ is a local C^p isomorphism at y_0.

Note that it is $\bar{q}^{-1}$ that is analogous to the g in the definition of "manifold," from which we see that the only thing that remains to be shown is that

$$M \cap q(W) = \bar{q}(U \cap W) = h(U \cap W)$$

or, if that does not hold, that we can make it hold by shrinking the set W, that is, there exists a open set W_1 in $\mathbb{R}^d$ that is a neighborhood of $(x_0, 0)$ such that

$$M \cap q(W_1) = h(U \cap W_1). \tag{5.2}$$

Now we use the condition that $h^{-1} : M \to U$ is a topological isomorphism, this means that there is an open subset M_1 in M that is a neighborhood of y_0 such that $x \in U \cap W$ whenever $y \in M_1$. By definition of "subspace topology" M_1 is necessarily of the form $M \cap Z_1$ for some open set Z_1 in $\mathbb{R}^d$, that is,

$$M \cap Z_1 \subset h(U \cap W)$$

and this implies

$$M \cap Z_1 = h\big(U \cap q^{-1}(Z_1)\big)$$

and since $q\big(q^{-1}(Z_1)\big) = Z_1$ we have (5.2) with $W_1 = q^{-1}(Z_1)$. □

5.4 Tangent Spaces

If M is a k-dimensional C^p manifold ($p \geq 1$) in $\mathbb{R}^d$ and $x \in M$, the *tangent space* at x, denoted $T_M(x)$ is the vector subspace of $\mathbb{R}^d$ consisting of all vectors v of the following form: there exists a sequence of points $x_n \in M$ and a sequence of scalars τ_n decreasing to zero such that

$$\frac{x_n - x}{\tau_n} \to v.$$

The definition invites us to think of the tangent space as the set of directions along which a sequence in M can converge to x.

If we look at the local coordinates provided by a local C^p isomorphism $g : U \to V$ that maps $M \cap U$ onto $S \cap V$ where S is a k-dimensional vector subspace of $\mathbb{R}^d$, we see that

$$\frac{g(x_n) - g(x)}{\tau_n} \to w,$$

where $w \in S$. By the rule for the product of limits

$$\frac{\|g(x_n) - g(x)\|}{\|x_n - x\|} \cdot \frac{\|x_n - x\|}{\tau_n} \to \|w\|,$$

and the first term on the left converges because g is differentiable; hence the second term on the left also converges, say to c. Then

$$\frac{g(x_n) - g(x)}{\|x_n - x\|} \to \nabla g(x) v$$

by the definition of differentiability. Thus

$$w = c \nabla g(x) v,$$

and, if $c > 0$ and we write $y = g(x)$,

$$v = c^{-1} \nabla g^{-1}(y) w.$$

Since g maps into a full k-dimensional neighborhood $S \cap V$ and c can be any nonnegative real number, it is clear that the set of vectors w that can arise in this fashion is the entire subspace S. Hence $T_M(x)$ is the image of the k-dimensional vector space S under the linear transformation $\nabla g^{-1}(y)$, that is

$$T_M(x) = \{\, g^{-1}(y) w : w \in S \,\}.$$

This shows us that the tangent space is a k-dimensional vector space and even gives us a formula for calculating it once we have found a local C^p isomorphism g. But the original definition is a better characterization to use as a definition because it doesn't depend on the particular local coordinates chosen and hence gives an intrinsic characterization of the tangent space.

5.5 Manifolds as Parameter Spaces

Now we want to allow C^p manifolds as parameter spaces. How can we do that? Up to now we have only allowed subsets of $\mathbb{R}^d$ that are neighborhoods of the true parameter value θ_0 as parameter sets. A manifold is (as we have defined it) a subset of $\mathbb{R}^d$, but in order to be a neighborhood in $\mathbb{R}^d$ of one of its points it would have to be a d-dimensional manifold, which is trivial (just a complicated way of describing an open set in $\mathbb{R}^d$). So allowing k-dimensional manifolds in $\mathbb{R}^d$ with $k < d$ is something new.

In general, there is no reason why a parameter space has to satisfy the condition we have up to now imposed. At first, a parameter space is just an index set Θ in a description of a statistical model $\mathcal{F} = \{\, f_\theta : \theta \in \Theta \,\}$. In order to serve as an index set it doesn't have to have any properties other than being a set. In fact, we can dispense with the notion of a parameter space entirely and just say our model $\mathcal{F}$ is a family of densities with no further description, or alternatively we could make $\mathcal{F}$ itself the index set with the trivial formula $\mathcal{F} = \{\, f : f \in \mathcal{F} \,\}$.

It is only when we start discussing convergence that we need at least a topology and preferably a metric on the parameter space. But there is still no reason why we want subspaces of $\mathbb{R}^d$ until we get to asymptotic normality.

Normal distributions live on a finite-dimensional vector spaces. If they are "nondegenerate," then they live on the full vector space under consideration rather than a proper subspace. It was our desire to have the space where the asymptotic distribution lives be the same vector space $\mathbb{R}^d$ as the vector space containing the parameter set and to have a nondegenerate limit that dictated our condition that the parameter space be a neighborhood of the true parameter value in $\mathbb{R}^d$.

Now we drop that condition and replace it with the following

Condition M. *The parameter space Θ is a k-dimensional manifold in $\mathbb{R}^d$.*

This is a bit less general than our previous condition in the case $k = d$ because it would require M to be a open set in $\mathbb{R}^d$, whereas before we allowed any neighborhood of the true parameter value, open or not. But Ferguson always required an open neighborhood of the true parameter value, so we are now being as restrictive as he is.[2]

What happens to all the theory we have developed so far when we allow this generalization? In one sense, nothing much because we can always choose local coordinates that map some relatively open neighborhood $\Theta \cap U$ of the true parameter value θ_0, where U is open in $\mathbb{R}^d$, to an open subset of $\mathbb{R}^k$ by a C^p isomorphism g so that

$$\theta \longleftrightarrow \big(g_1(\theta), \ldots, g_k(\theta)\big)$$

is a homeomorphism between $\Theta \cap U$ and its image $g(\Theta \cap U)$. If we write $\Phi = g(\Theta \cap U)$ and denote the restriction of g^{-1} to Φ by h, then $h : \Phi \to \Theta \cap U$ gives us a parameterization that satisfies the old condition, that is, Φ is a neighborhood in $\mathbb{R}^k$ of the true parameter value $\varphi_0 = g(\theta_0)$. Hence we can apply all the asymptotic theory developed so far in the "φ coordinates" (assuming the other regularity conditions are met in these coordinates).

In another sense, something profound happens to the theory, since we are not really interested in the "φ coordinates" but in the original parameterization "θ coordinates." What happens there? Suppose we have asymptotic normality in the "φ coordinates"

$$\sqrt{n}(\hat{\varphi}_n - \varphi_0) \xrightarrow{\mathcal{L}} Z,$$

where $Z \sim \mathcal{N}(0, K^{-1})$ for some invertible $k \times k$ matrix K. Well $\theta = h(\varphi)$, so we apply the delta method. This requires that h and hence g^{-1} be differentiable, so we need $p \geq 1$. Then applying the delta method gives

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \nabla h(\varphi_0) Z. \tag{5.3}$$

This formula tells us many things. First, $\hat{\theta}_n$ is also asymptotically normal. Second, since Z is k-dimensional, the right hand side of (5.3) is a normal random vector of dimension at most k, and hence is degenerate considered as a random element of $\mathbb{R}^d$. Third, Z lives on $\mathbb{R}^k$ and $\nabla h(\varphi_0) = \nabla g^{-1}(\varphi_0, 0)$ maps $\mathbb{R}^k$ onto

[2] If we wanted to introduce "manifolds with boundaries" we could keep the generality we had before in the $k = d$ case, but it does not seem worth the trouble.

the tangent space $T_\Theta(\theta_0)$, so the right hand side of (5.3) lives on the tangent space. Fourth, since $\nabla g^{-1}(\varphi_0, 0)$ is invertible (because g is a C^p isomorphism with $p \geq 1$), it maps k-dimensional subspaces to k-dimensional subspaces, so the right hand side of (5.3) is a nondegenerate normal random vector on the subspace $T_\Theta(\theta_0)$ of $\mathbb{R}^d$.

What degree of smoothness p should we require? Obviously, we want to impose the weakest conditions that will do the job, but what are they? It is clear from the analysis above that we need at least $p \geq 1$ in order to apply the delta method.

The fact that C^1 isomorphisms do transfer asymptotic properties via the delta method suggests that C^1 manifolds are enough and there is no reason to consider C^2 manifolds or manifolds with higher degrees of smoothness. But this is not quite all there is to be said on the subject, since our theorems about maximum likelihood involve regularity conditions. How do we apply them to parameter spaces that are manifolds?

As we did before, we can apply all the previously developed theory by transforming to local coordinates, called "φ coordinates" above, and applying the regularity conditions in those coordinates. To recapulate that analysis, we always can consider that the model has been given in terms of an open subset Φ of $\mathbb{R}^k$ containing the unknown parameter value φ_0 and a C^p map $h : \varphi \to \mathbb{R}^d$ satisfying the conditions of Lemma 5.4 so that $h(\Phi)$ is a C^p manifold in $\mathbb{R}^d$ containing the true parameter value $\theta_0 = h(\varphi_0)$. We may not have been given the model in this form, but there always exists a C^p isomorphism that puts it in this form. If we can verify the regularity conditions in the "φ coordinates," then any asymptotic results can be transferred to the "θ coordinates" by the delta method.

However, it would be very unsatisfactory if the regularity conditions only held in some "magic" local coordinates. What if you couldn't figure out the "magic" parameterization? Then there would be an asymptotic result, the model stated in terms of the manifold Θ would satisfy the regularity conditions, but you couldn't *prove* that it satisfied the conditions. This would be a very unsatisfactory state of affairs. Fortunately, it can't happen. Whether a model satisfies the regularity conditions, at least those we have studied, does not depend on the choice of local coordinates. By Lemma 5.2 all local coordinates are related by local C^p isomorphisms. Hence the following theorem does exactly what we need.

Theorem 5.5. *If a model is ULAN at rate α_n, then it is also ULAN at rate α_n after a reparameterization by a C^1 isomorphism.*

Before starting the proof, we need an observation and a lemma. Every linear operator on a finite-dimensional vector space is bounded, that is, if $\|\cdot\|$ is any norm for $\mathbb{R}^d$, the corresponding operator norm for a linear operator $A : \mathbb{R}^d \to \mathbb{R}^d$ defined by

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

is finite. This implies $\|Ax\| \le \|A\| \cdot \|x\|$ for all operators A and vectors x.

Lemma 5.6. *If $g : \Phi \to \Psi$ is a C^1 isomorphism, where Φ and Ψ are open subsets of $\mathbb{R}^k$, $\varphi \in \Phi$, $\psi_0 = g(\varphi_0)$, and α_n is any sequence converging to zero, and if $B(x, \rho)$ denotes the closed ball in $\mathbb{R}^k$ centered at x with radius ρ, then for any $r > 0$, there exists a $\rho > 0$ and an integer N depending on ρ such that*

$$g\big(B(\varphi_0, r\alpha_n) \cap \Phi\big) \subset B(\psi_0, \rho\alpha_n), \qquad \text{whenever } n \ge N. \tag{5.4}$$

In fact we can choose any $\rho > \|\nabla g(\varphi_0)\|$.

This lemma doesn't seem very simple, but the idea is simple. As a shorthand we summarize this lemma as the assertion that a C^1 isomorphic reparameterization eventually (meaning for $n \ge N$) maps bounded balls to bounded balls "on the α_n scale."

Proof. A point $\varphi \in B(\varphi_0, r\alpha_n) \cap \Theta$ has the form

$$\varphi = \varphi_0 + \alpha_n \delta$$

for some δ with $\|\delta\| \le r$. This φ maps to

$$\psi = \psi_0 + \alpha_n \eta$$

say, where

$$\begin{aligned} \eta &= \frac{\psi - \psi_0}{\alpha_n} \\ &= \nabla g(\varphi_0)\delta + \|\delta\| \cdot w(\alpha_n \delta) \end{aligned}$$

where w is some function continuous at zero with $w(0) = 0$. Still assuming $\|\delta\| \le r$, this implies for any $\epsilon > 0$ there is an integer N such that $\|w(\alpha_n\delta)\| \le \epsilon$ whenever $n \ge N$. Thus

$$\|\eta\| \le r\big(\|\nabla g(\varphi_0)\| + \epsilon\big), \qquad n \ge N.$$

That proves (5.4) and the assertion about the choice of ρ. □

Proof of the theorem. By definition a C^1 isomorphism g maps an open neighborhood of φ_0 to an open neighborhood of ψ_0.

By the lemma, if

$$\psi_0 + \alpha_n \eta_n = g(\varphi_0 + \alpha_n \delta_n),$$

then η_n is a bounded sequence if and only if δ_n is a bounded sequence. Thus contiguity trivially transfers from one parameterization to the other.

Since the true parameter value $\psi_0 = g(\varphi_0)$ is fixed throughout, let l_n denote the log likelihood in the "ψ coordinates" and let us change notation from $l_n(\psi, \psi_0)$ to $l_n(\psi)$, keeping the same meaning. This enables us to write the log likelihood for the parameter $\varphi = g^{-1}(\psi)$ as $\tilde{l}_n = l_n \circ g$. Note that $l_n(\psi_0) = \tilde{l}_n(\varphi_0) = 0$.

Now we assume the "ULAN at rate α_n" condition holds in the "φ coordinates" so for any $\rho > 0$

$$\sup_{\substack{\delta \in B(0,\rho) \\ \varphi_0 + \alpha_n \delta \in \Phi}} \left| \tilde{l}_n(\varphi_0 + \alpha_n \delta) - \left[\langle \delta, S_n \rangle - \tfrac{1}{2} \langle \delta, K_n \delta \rangle \right] \right|$$

converges in probability to zero for some random variables vectors S_n and matrices K_n with K_n converging in probability to a constant matrix K. We must show that a similar condition then also holds in the "ψ coordinates." Now write

$$\psi_0 + \alpha_n \eta = g(\varphi_0 + \alpha_n \delta),$$

so

$$\begin{aligned} \eta &= \frac{g(\varphi_0 + \alpha_n \delta) - g(\varphi_0)}{\alpha_n} \\ &= A\delta + o(1) \end{aligned}$$

where $A = \nabla g(\varphi_0)$. Introduce the random variables

$$\begin{aligned} R_n &= A^{-1} S_n \\ J_n &= A^{-1} K_n A^{-1} \end{aligned}$$

Then, again using the lemma, for any $r > 0$ there is an $\rho > 0$ such that

$$\begin{aligned} &\sup_{\substack{\eta \in B(0,r) \\ \psi_0 + \alpha_n \eta \in \Psi}} \left| l_n(\psi_0 + \alpha_n \eta) - \left[\langle \eta, R_n \rangle - \tfrac{1}{2} \langle \eta, J_n \eta \rangle \right] \right| \\ &\qquad \leq \sup_{\substack{\delta \in B(0,\rho) \\ \varphi_0 + \alpha_n \delta \in \Phi}} \left| (l_n \circ g)(\varphi_0 + \alpha_n \delta) - \left[\langle \delta, S_n \rangle - \tfrac{1}{2} \langle \delta, K_n \delta \rangle \right] \right| + o_p(1) \end{aligned}$$

because S_n and K_n are bounded in probability. □

5.6 Submanifolds

If M_1 is a k_1-dimensional C^p manifold in $\mathbb{R}^d$, we say that $M_0 \subset M_1$ is a k_0-dimensional C^p *submanifold* of M_1 if there for each $x \in M_0$ a local C^p isomorphism at x, denote it $g : U \to V$ that maps $M_1 \cap U$ to $S_1 \cap V$ and $M_0 \cap U$ to $S_0 \cap V$, where S_1 and S_0 are subspaces of dimension k_1 and k_0, respectively. Obviously, this requires $S_0 \subset S_1$.

We will not go into detail, but merely remark that the two construction methods in Lemmas 5.3 and 5.4 also work for submanifolds (this is almost completely obvious) after transformation to local coordinates.

5.7 Tests Revisited

5.7.1 The Likelihood Ratio Test

Theorem 5.7. *Suppose a model having a k_1-dimensional C^1 manifold as a parameter space satisfies the ULAN conditions at rate α_n. Let $\hat{\theta}_n$ be an ELE for the full model and let θ_n^* be an ELE for the restricted model that constrains θ to lie in a k_0-dimensional C^1 submanifold of the parameter space of the full model ($k_0 < k_1$), then*

$$L_n = 2\bigl(l_n(\hat{\theta}_n) - l_n(\theta_n^*)\bigr). \tag{5.5}$$

has an asymptotic $\text{chi}^2(k_1 - k_0)$ *distribution, where $l_n(\theta)$ is any log likelihood for the problem.*

Proof. Transform to local coordinates and apply Theorem 4.1 □

The point is that although the proof (at least our method of proof) requires transformation to local coordinates, the calculation of $\hat{\theta}_n$ and θ_n^* do not. The method of Lagrange multipliers allows one to find maxima of functions defined on manifolds defined by constraint equations (as in Lemma 5.3) without using transformation to local coordinates.

5.7.2 The Rao Test

The Rao test does not really have a similar form that is independent of local coordinates. The trouble is that we don't really know what $\nabla l_n(\theta)$ is supposed to mean when θ takes values in a manifold. Essentially, the calculation requires transformation to local coordinates in order to calculate this derivative.

There is one special case, however, this is worth some study. If we let the full model have a parameter space that is an open subset of $\mathbb{R}^d$ then $\nabla l_n(\theta)$ means just what it always has.

Theorem 5.8. *Suppose conditions 1 through 4 of Theorem 3.1 (the "usual regularity conditions" for maximum likelihood) and suppose the Fisher information $I(\theta)$ is continuous at θ_0. Let $\hat{\theta}_n$ be an ELE for the full model having an open subset of $\mathbb{R}^d$ for its parameter space, and let θ_n^* be an ELE for the restricted model that constrains θ to lie in a k-dimensional C^1 submanifold of $\mathbb{R}^d$ with $k < d$,*

$$R_n = \frac{1}{n}\langle \nabla l_n(\theta_n^*), I(\theta_n^*)^{-1} \nabla l_n(\theta_n^*) \rangle \tag{5.6}$$

is asymptotically equivalent to the likelihood ratio test statistic (5.5).

Proof. Transform to local coordinates and apply Theorem 4.2. The only issue is to show that the Rao statistic (5.6) is actually invariant under such a transformation. Suppose $\varphi = g(\theta)$ is the coordinate transformation, and write $h = g^{-1}$ so $\theta = h(\varphi)$. Also write $\varphi_n^* = g(\theta_n^*)$ and $A_n(\varphi) = \nabla h(\varphi)$. The log likelihood in the transformed coordinates is

$$\tilde{l}_n(\varphi) = (l_n \circ h)(\theta)$$

so the score vector is

$$\tilde{s}_n(\varphi_n^*) = \nabla \tilde{l}_n(\varphi_n^*) = \nabla l_n(\theta_n^*) A_n(\varphi_n^*) \tag{5.7}$$

by the chain rule. This formula makes sense when we think of the derivation via the chain rule. Both objects on the right are derivatives so $\nabla l_n(\theta_n^*)$ is represented as a $1 \times n$ matrix and $A_n(\varphi_n^*)$ as a $n \times n$ matrix. But if we want to think of the $\nabla l_n(\theta_n^*)$ and $\nabla \tilde{l}_n(\varphi_n^*)$ in "the usual way" as a "column vectors" we have to have to write the equation the other way around as

$$\tilde{s}_n(\varphi_n^*) = A_n(\varphi_n^*) s_n(\theta_n^*).$$

We worked out in a homework exercise that the transformation rule for Fisher information is

$$\widetilde{I}(\varphi) = A_n(\varphi) I(\theta) A_n(\varphi)$$

where $\theta = h(\varphi)$ so the Rao statistic in "φ coordinates" is

$$\begin{aligned} \tilde{s}_n(\varphi_n^*)' \widetilde{I}(\varphi)^{-1} \tilde{s}_n(\varphi_n^*) &= s_n(\theta_n^*)' A_n(\varphi_n^*) \bigl[A_n(\varphi_n^*) I(\theta_n^*) A_n(\varphi_n^*) \bigr]^{-1} A_n(\varphi_n^*) s_n(\theta_n^*) \\ &= s_n(\theta_n^*)' I(\theta_n^*)^{-1} s_n(\theta_n^*) \end{aligned}$$

hence the same as in "θ coordinates." □

5.7.3 The Wald Test

The Wald test also does not have a general form that can be stated without transformation to local coordinates. There is, as with the Rao test, a special case worth some study. Suppose, as we did with the Rao test, that the full model has a parameter space Θ_1 that is an open subset of $\mathbb{R}^d$ and the null hypothesis is given by a constraint function

$$\Theta_0 = \{\, \theta \in \Theta_1 : h(\theta) = 0 \,\}$$

where $h : \Theta_1 \to \mathbb{R}^{d-k}$ satisfies the conditions of Lemma 5.3. Now write $H(\theta) = \nabla h(\theta)$.

Theorem 5.9. *Suppose the conditions of Definition 4.1. Let $\hat{\theta}_n$ be an ELE for the full model. Let h and H be as described above, and write $h_n = h(\hat{\theta}_n)$ and $H_n = H(\hat{\theta}_n)$. Then*

$$W_n = n h_n' H_n (H_n K_n^{-1} H_n')^{-1} H_n h_n \tag{5.8}$$

is asymptotically equivalent to the likelihood ratio test statistic (5.5) *or the Rao test statistic* (5.6)*.*

The proof is the same as the proof for the Rao statistic. We derive this theorem from Theorem 4.3 by showing that the statistic (5.8) is invariant under coordinate transformation and that in the special case of a local coordinate

transformation that maps Θ_0 bijectively onto a relatively open subset of $\mathbb{R}^k$ we get (4.13).

The second point is easy to see. The constraint that sets the first k coordinates of θ to a specified value θ_0 is of the form

$$h(\theta) = G(\theta - \theta_0) = 0$$

where $G = \begin{pmatrix} I & 0 \end{pmatrix}$. The gradient of a linear function is the same function so $H(\theta) = G$. Plugging these in to (5.8) gives (4.13). The invariance part of the proof is much the same as with the Rao statistic. Use the chain rule and note that the "Jacobian matrices" cancel.

Appendix A

Odds and Ends

A.1 Big Oh Pee and Little Oh Pee

A sequence of random variables X_n is said to be $o_p(1)$ if

$$X_n \xrightarrow{P} 0.$$

This is just a convenient notational variant. A sequence of random variables X_n is said to be $O_p(1)$ if it is bounded in probability, which is the condition for Prohorov's theorem (that is, $O_p(1)$ implies that there exists a subsequence X_{n_k} converging in law to some random variable). This too is just a convenient notational variant.

The notations gain power when we consider pairs of sequences. Suppose X_n and Y_n are random sequences taking values in any normed vector space, then

$$X_n = O_p(Y_n)$$

means $X_n/\|Y_n\|$ is bounded in probability and

$$X_n = o_p(Y_n)$$

means

$$\frac{X_n}{\|Y_n\|} \xrightarrow{P} 0.$$

These notations are often used when the sequence Y_n is deterministic, for example $X_n = O_p(n^{-1/2})$. But they are also used when both are deterministic, for example, we say two sequences X_n and Y_n are *asymptotically equivalent* if

$$X_n - Y_n = o_p(Y_n).$$

Problems

A-1. Prove the following

(a) $O_p(X_n)O_p(Y_n) = O_p(X_nY_n)$.

(b) $O_p(X_n)o_p(Y_n) = o_p(X_nY_n)$.

(c) $o_p(X_n)o_p(Y_n) = o_p(X_nY_n)$.

(d) $o\big(O_p(X_n)\big) = o_p(X_n)$.

Bibliography

Bertsekas, D. P. and S. E. Shreve (1978). *Stochastic Optimal Control: The Discrete Time Case*. New York: Academic Press.

Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley.

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

Fletcher, R. (1987). *Practical Methods of Optimization* (Second ed.). Chichester and New York: Wiley.

Fristedt, B. and L. Gray (1997). *A Modern Approach to Probability Theory*. Boston: Birkhäuser.

Jacod, J. and A. N. Shiryaev (1987). *Limit Theorems for Stochastic Processes*. Berlin: Springer-Verlag.

Lang, S. (1993). *Real and Functional Analysis* (Third ed.). New York: Springer-Verlag.

Le Cam, L. and G. L. Yang (1990). *Asymptotics in Statistics: Some Basic Concepts*. New York: Springer-Verlag.

Rockafellar, R. T. and R. J.-B. Wets (1998). *Variational Analysis*. Berlin: Springer-Verlag.