Chapter 1

Heuristics

The official dogma on parametric estimation is: Good estimators converge to the right thing and have limiting normal distributions; moreover, the variance of the limiting distribution can't be smaller than a quantity defined by the Fisher information function; estimators that achieve the asymptotic lower bound are called efficient; maximum likelihood estimators are efficient.

The dogma is not quite correct, but much of it can be rescued in slighlty altered form. And therein hangs a tale. This Chapter starts the story by describing (non-rigorously) one method for building estimators that typically have good properties, by explaining when the estimators should be efficient, and by showing what can go wrong.

SECTION *1 establishes some notation that helps to clarify the role of parameters and models.*

SECTION *2 sketches the typical steps involved in establishing asymptotic behavior of an estimator, with asymptotic normality of M-estimators as a guiding example.*

SECTION *3 introduces the concept of efficiency. It presents a nonrigorous argument for why the limiting distributions for M-estimators should not have variance smaller than the information bound and why maximum likelihood should achieve that bound.*

SECTION *4 explains why the efficiency assertions from the previous Section are not true without some extra constraints on limiting behavior of estimators. It points to the rigorous treatments in later Chapters.*

1.1 Notation and truth

Heuristics::notation

There is a large body of statistical theory and literature regarding optimality and large sample approximation, some of it true, some of it almost true, and some of it a little bit wrong. However, there are grains of truth buried amongst the chaff of ideas that are not quite correct. Some ideas—such as efficiency and sufficiency—have survived mathematical indignities and counterexamples by evolving to retain their secure place at the foundations of statistics. And some myths have died out.

In this book I will deal rigorously with some parts of the theory that are both useful and mathematically correct. Initially, the emphasis will be on models smoothly indexed by a finite-dimensional parameter, but with a

gradual infusion of modern probabilistic techniques that simplify the rigorous discussion of more abstract and general models.

To appreciate the virtues of rigor, you must first understand some of the folklore.

Many problems in mathematical statistics boil down to the following question. Let $\mathcal{P} := \{\mathbb{P}_\theta : \theta \in \Theta\}$ be a statistical model—a family of probability measures all defined on the same sigma-field $\mathcal{F}$ on a set Ω. Let $T = T(\omega)$ be a random variable (or, more generally, a random vector, or even a random element of some wonderfully abstract space). What is the distribution of T under each $\mathbb{P}_\theta$ model?

Typically T is thought of as an estimator for some function $\tau(\theta)$ of the indexing parameter θ, or perhaps it represents a choice from a set of possible actions. From an exact knowledge of the distribution of T under each $\mathbb{P}_\theta$, one could in principle calculate all the various expectations used to evaluate the performance of T in the traditional decision theoretic senses. Unfortunately, it is seldom possible to calculate the distributions explicitly. Instead, one is often forced to invoke simplifying approximations, or, more formally, find limiting forms of distributions for sequences of estimators $\{T_n\}$ under a sequence of models $\mathcal{P}_n := \{\mathbb{P}_{\theta,n} : \theta \in \Theta_n\}$. The extra parameter n typically denotes a sample size. The approximations are called then ***large-sample*** (or ***asymptotic***) distributions.

It is also traditional to use statistical models to guide the construction of an estimator and then to evaluate hypothetical behavior only under those models. For example, concepts such as sufficiency and efficiency refer only to behavior under probability measures in some specific model. At times the dual role of the model can cause some confusion. In what follows, I will adopt notation that I believe helps to distinguish the two roles.

Let me start with a more concrete example to establish some finer points of notation.

mle.example <1>

Example. Let $\{f_\theta : \theta \in \Theta\}$ be a family of probability densities (with respect to some fixed dominating measure, such as Lebesgue measure on the real line or on some $\mathbb{R}^k$) for probability measures $\{P_\theta : \theta \in \Theta\}$ on a fixed sigma-field. Let $X_1, \ldots, X_n$ be random variables on Ω, and let $\mathbb{P}_{\theta,n}$ be a probability measure on Ω under which the $\{X_i\}$ are independent, each with distribution P_θ. The method of maximum likelihood defines $\widehat{\theta}_n = \widehat{\theta}_n(\omega)$ as the value of θ that maximizes

$$L_n(\omega, \theta) := \prod\nolimits_{i \le n} f(X_i(\omega), \theta).$$

That is, $\widehat{\theta}_n(\omega) := \operatorname{argmax}_{\theta \in \Theta} L_n(\theta, \omega)$. For the moment I will ignore all

questions of existence, uniqueness, or measurability of a maximizing value.

The parameter θ is now playing two roles: as a dummy variable, a placeholder that indicates a function to be maximized; and as a label that identifies models giving particular hypothetical distributions for the observations $\mathbf{X}(\omega) := (X_1(\omega), \ldots, X_n(\omega))$. To emphasize the first role, it helps to think of $\widehat{\theta}_n$ not as a function of ω but rather as a function on $\mathbb{R}^n$ (or $\mathcal{X}^n$, if the variables X_i were to take values in a set $\mathcal{X}$). That is, we can define an ***estimating function***

$$\widehat{\theta}_n(\mathbf{x}) := \operatorname{argmax}_{s\in\Theta} \prod_{i\le n} f(x_i, s) \qquad \text{where } \mathbf{x} := (x_1, \ldots, x_n),$$

then define the ***estimator*** $\widehat{\theta}_n(X_1(\omega), \ldots, X_n(\omega))$. In the definition of $\widehat{\theta}_n(\mathbf{x})$ I have even used a different letter for the dummy variable in order save θ for its second role.

This approach focuses attention on $\widehat{\theta}_n(\mathbf{x})$ as a function defined by the fitted model, without making any particular assumptions about how $\mathbf{x}$ is to be interpreted. Performance of $\widehat{\theta}_n(\mathbf{X}(\omega))$ under various probabilistic mechanisms for generation of the sample $\mathbf{X}(\omega)$, and not just for those mechanisms prescribed by the models, becomes a separate question. That is, the view of $\widehat{\theta}_n(\cdot)$ as a function of $\mathbf{x}$ disentangles the issue of definition via a model from the issue of behaviour of the estimator under those models.

Remark. Nearly always it is only the distribution of the random vector $X(\omega)$ under some probability measure $\mathbb{P}$ on Ω that matters. That is, the necessary calculations all involve only probability measures defined on $\mathcal{X}^n$. We can save on notation by taking Ω equal to $\mathcal{X}^n$ itself, regarding $\omega = \mathbf{x} = (x_1, \ldots, x_n)$ as the observed data and $X_i(\omega) = x_i$ as a single observation.

For asymptotic purposes, where behavior as n tends to infinity is of interest, we could also take Ω to be $\mathcal{X}^{\mathbb{N}}$, a countable product of the coordinate spaces. For the nth in a sequence of models, the observed data would be $(x_1, \ldots, x_n)$, an initial segment of $\omega = (x_1, x_2, \ldots)$. The $\mathbb{P}_{\theta,n}$'s could then be defined on the sub-sigma-field $\mathcal{F}_n$ generated by $x_1, \ldots, x_n$.

The downside of choosing Ω to be $\mathcal{X}^n$ or $\mathcal{X}^{\mathbb{N}}$ is revealed when statistical procedures involve an auxiliary randomization, perhaps generated using a random variable U distributed Unif$(0,1)$ independently of all the X_i's. It would be unnatural to require U to be defined as a (measurable) function on $\mathcal{X}^n$ or $\mathcal{X}^{\mathbb{N}}$. Instead, we could use a "richer probability space" $\mathcal{X}^{\mathbb{N}} \times (0,1)$ and replace $\mathbb{P}_{\theta,n}$ by a product measure $\mathbb{P}_{\theta,n} \otimes \text{Unif}(0,1)$

When we study the behaviour of $\widehat{\theta}_n$ under the model corresponding to a particular θ, that value of θ is sometimes referred to as the "true value",

or $\mathbb{P}_{\theta,n}$ as the "true model". Of course if we actually knew the truth we wouldn't need to estimate; the word "true" serves merely to distinguish one particular parameter value during the course of a calculation. A name like "test case" or "hypothetical model" might be less misleading. It sometimes helps to denote the temporarily true value by another symbol, such as θ_0, to avoid confusion with θ as a variable over which to optimize.

□

1.2 Limit theory

Heuristics::limit.theory

With those preliminaries about truth out of the way, let me turn to a general problem that illustrates a number of important asymptotic ideas. Suppose the observed data are given by random quantities $\mathbf{X} = (X_1, \ldots, X_n)$, with each X_i taking values in a set $\mathcal{X}$ (such as the real line). Suppose Θ is subset of the real line, perhaps interpreted as the indexing set of a model, or perhaps not. Suppose $\{g(\cdot, \theta) : \theta \in \Theta\}$ is a collection of real-valued functions on $\mathcal{X}$. Define an estimating function $\widehat{\theta}_n(\mathbf{x})$ as the value of θ that minimizes

$$G_n(\mathbf{x}, \theta) := n^{-1} \sum\nolimits_{i \le n} g(x_i, \theta).$$

That is, $\widehat{\theta}_n(\mathbf{x}) = \operatorname{argmin}_{\theta \in \Theta} G_n(\mathbf{x}, \theta)$. In the terminology of Huber (1964), the corresponding $\widehat{\theta}_n(\mathbf{X})$ is called an ***M-estimator***.

What can we say about the behaviour of $\widehat{\theta}_n(\mathbf{X})$ when the X_i are independent, each with marginal distribution P? Equivalently, how does $\widehat{\theta}_n(\mathbf{x})$ behave under the product measure P^n on $\mathcal{X}^n$?

For the purposes of an asymptotic answer to this question, we might regard the data as the initial segment of an infinite sequence of independent $\mathcal{X}$-valued random variables $X_1, X_2, \ldots$, all defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with each X_i having distribution P. Alternatively, we might treat the data as one row in a triangular array of random variables, defined on a probability space $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ that can change with n. The X_1 for the nth row of the array might be completely unrelated to the X_1 element in other rows. It might even be better to make this possibility explicit, by writing the data as $\mathbf{X}_n := (X_{n,1}, \ldots, X_{n,n})$. The distribution P could also be replaced by a P_n that changes with n, a generalization that will be needed when discussing behavior of estimators under sequences of alternatives. However, for the moment I will work with the simpler setting of a fixed underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a fixed distribution P.

To understand how $\widehat{\theta}_n$ behaves for large samples, we first have to understand what G_n is doing. The key idea is to approximate G_n by another

function, or even another random process, whose minimizing value is more easily analyzed. For a rigorous treatment we will have to determine the effect of the errors of approximation uniformly over various sets of θ values, to ensure that the minimizing values are close. For the moment, though, I will approximate with abandon.

The traditional analysis breaks into three stages.

1.2.1 Local concentration (consistency)

Asymptotic arguments often start with an attempt to show that $\widehat{\theta}_n$ concentrates with high probability near some fixed θ_0, which depends on the underlying distributions. If the $\{X_i\}$ are all defined on a fixed Ω and if the $\mathbb{P}$ does not change with n, it makes sense to ask about convergence at $\mathbb{P}$-almost all ω. However, if we allow the model to change with n, with each $\widehat{\theta}_n$ analyzed under a different $\mathbb{P}_n$ model, almost sure convergence is ill-defined. It is then better to enquire about convergence in $\mathbb{P}_n$-probability of $\widehat{\theta}_n$ to θ_0 as n tends to infinity, or even just about concentration around some (possibly random) value that might change with n:

$$\mathbb{P}_n\{|\widehat{\theta}_n - \theta_n| > \epsilon\} \to 0 \qquad \text{for each } \epsilon > 0.$$

Of course, we really want to be able to assert some form of concentration for a collection of probability measures on Ω, maybe $\{\mathbb{P}_\theta : \theta \in \Theta\}$ or $\{\mathbb{P}_{\theta,n} : \theta \in \Theta\}$. (The value around which the estimator concentrates should depend on θ.) A concentration result that held for only one data-generating mechanism would serve only as an illustration of one possible behaviour. If $\widehat{\theta}_n$ converges to θ almost surely under $\mathbb{P}_\theta$, for each $\theta \in \Theta$, then the estimator is said to be ***strongly consistent*** (for θ). If the convergence holds only in $\mathbb{P}_\theta$-probability (or $\mathbb{P}_{\theta,n}$-probability), the estimator is said to be ***weakly consistent***.

> **Remark.** I believe weak consistency is actually the more useful idea because, typically, consistency is just a prelude to a more detailed analysis of asymptotic behaviour. For me, strong consistency is often of interest only because it implies weak consistency.

For the M-estimation problem, at each fixed θ a law of large numbers (strong or weak?) implies that $G_n(\theta)$ should be close to its expected value $G(\theta) = P^x g(x, \theta)$. That is, as a first approximation, we should have $G_n(\theta) \approx G(\theta)$. We might then hope that $\operatorname{argmin}_\theta G_n(\theta) \approx \operatorname{argmin}_\theta G(\theta)$. That is, we might hope that $\widehat{\theta}_n$ lies close to the value $\theta_0 := \operatorname{argmin}_\theta G(\theta)$, the value that is defined to minimize the approximating G.

You will learn in Chapter 2 one rigorous way, essentially due to Wald (1949), to make this approximation idea more precise and establish consistency.

Remark. Sometimes consistency is the most challenging part of an asymptotic argument because it requires global approximations to G_n.

1.2.2 Rate of convergence

If we know that $\widehat{\theta}_n$ concentrates near some θ_0 (or some θ_n), then we can ask, How near is near? or, How rapidly does it converge?. Often a rate is easier to establish (once consistency holds) than the consistency itself because we are able to make use of local approximations (such as Taylor expansions) to G_n; it is usually easier to establish the required uniformity of approximation if we can concentrate on smaller neighborhoods of θ_0.

Again we have a choice of asking about rates at almost all ω or about rates in the sense of convergence in probability. Again I believe the in-probability assertion is generally the more relevant, in part because of its role as a necessary preliminary to the next stage in the analysis.

The rigorous treatment of rates of convergence will begin with Chapter 3, where local errors will be bounded by means of Taylor expansions. Chapter 4 will generalize the method.

For many M-estimation problems, good estimators converge in probability at a $1/\sqrt{n}$ rate, that is, $\widehat{\theta}_n = \theta_0 + O_p(1/\sqrt{n})$, a property sometimes referred to as ***root-n consistency***. *[See the discussion near the start of Chapter 3 if you are not familiar with the $O_p(\cdot)$ and $o_p(\cdot)$ notation.]*

Note that the rate of convergence, as just defined, is only an upper bound. With this terminology an estimator has many rates of convergence. For example, for a root-n consistent estimator we also have $\widehat{\theta}_n = \theta_0 + O_p(n^{-r})$ for each $r < 1/2$. It is more satifactory to have a rate $\widehat{\theta}_n - \theta_0 = O_p(\alpha_n)$ for which the analogous assertion fails if α_n is replaced by any β_n for which $\beta_n/\alpha_n \to 0$. Chapter 9, on minimax rates of convergence, will describe one way to formalize this idea.

1.2.3 Limiting distribution

Another way to remove the ambiguity about a rate of convergence is to demonstrate existence of a nontrivial limiting distribution for the standardized estimator. In many classical problems the limiting distribution is normal.

The method that will be presented in Chapters 3 and 4 for proving existence of limit distributons is essentially just a refinement of the rate calculations, but for concentration near a random θ_n. For classical parametric problems we typically have $O_p(1/\sqrt{n})$ concentration near θ_0 and $o_p(1/\sqrt{n})$ concentration near a θ_n defined by the argmin of a random process that approximates G_n, with a uniform $o_p(1/n)$ error over $O_p(1/\sqrt{n})$-neighborhoods of θ_0.

A slightly more specific assertion is sometimes possible. Instead of mere existence of a limiting distribution, we might have an asymptotic representation $\sqrt{n}(\widehat{\theta}_n - \theta_0) = W_n + o_p(1)$, where W_n has known limiting behaviour under various models. *[In Chapter 7 you will learn one reason why such a representation is so useful.]*

For M-estimation with a smoothly parametrized g, we can learn about the behaviour of G_n near θ_0 by means of Taylor expansion of $g(x,\cdot)$ about θ_0. To avoid later confusion with transposes of vectors, I will denote partial derivatives with respect to θ by dots:

$$\dot{g}(x,\theta) := \frac{\partial}{\partial\theta} g(x,\theta), \qquad \ddot{g}(x,\theta) := \frac{\partial^2}{\partial\theta^2} g(x,\theta),$$

and so on. The pointwise Taylor expansion

$$g(x,\theta) \approx g(x,\theta_0) + (\theta-\theta_0)\dot{g}(x,\theta_0) + \tfrac{1}{2}(\theta-\theta_0)^2\ddot{g}(x,\theta_0)$$

gives a quadratic approximation for G_n near θ_0:

$$\begin{aligned} G_n(\mathbf{X},\theta) &= \frac{1}{n}\sum\nolimits_{i\le n} g(X_i(\omega),\theta) \\ &\approx \frac{1}{n}\sum\nolimits_{i\le n} \left(g(X_i,\theta_0) + (\theta-\theta_0)\dot{g}(X_i,\theta_0) + \tfrac{1}{2}(\theta-\theta_0)^2\ddot{g}(X_i,\theta_0)\right). \end{aligned}$$

It also gives an approximation for G near θ_0:

$$\begin{aligned} G(\theta) &:= P^x g(x,\theta) \\ &\approx P^x g(x,\theta_0) + (\theta-\theta_0)P^x\dot{g}(x,\theta_0) + \tfrac{1}{2}(\theta-\theta_0)^2 P^x\ddot{g}(x,\theta_0) \\ &= G(\theta_0) + 0 + \tfrac{1}{2}(\theta-\theta_0)^2\ddot{G}(\theta_0). \end{aligned}$$

Notice that the linear term must vanish at a minimizing value, at least if θ_0 is an interior point of Θ. This property corresponds to "differentiation under the expectation",

$$0 = \dot{G}(\theta_0) = \frac{\partial}{\partial\theta} P^x g(x,\theta)\big|_{\theta=\theta_0} = P^x \frac{\partial}{\partial\theta} g(x,\theta)\big|_{\theta=\theta_0} = P^x\dot{g}(x,\theta_0),$$

an operation sometimes justified by explicit domination assumptions. Identification of $\ddot{G}(\theta_0)$ with $P^x\ddot{g}(x,\theta_0)$ is often justified in a similar way.

Remark. For some important examples, the function G is twice differentiable without $g(x,\cdot)$ being twice differentiable. The integral over P can sometimes provide extra smoothness.

The random variables $\dot{g}(X_i,\theta_0)$ have zero expected value. Assuming that they also have a finite variance, $\sigma^2 := P^x\dot{g}(x,\theta_0)^2$, we can then conclude that the standardized average

$$Z_n := \sum\nolimits_{i\le n} \dot{g}(X_i,\theta_0)/\sqrt{n}$$

should be approximately $N(0,\sigma^2)$ distributed.

Typically, the random variables $\ddot{g}(X_i,\theta_0)$ will have a strictly positive expected value $P^x\ddot{g}(x,\theta_0) = \ddot{G}(\theta_0)$, otherwise higher-order derivatives might be needed to guarantee even a local minimum at θ_0. The average $\sum_{i\le n}\ddot{g}(X_i,\theta_0)/n$ should be close to this expected value, which makes the random criterion function approximately a quadratic in $\theta-\theta_0$:

$$G_n(\theta) \approx G_n(\theta_0) + (\theta-\theta_0)Z_n/\sqrt{n} + \tfrac{1}{2}(\theta-\theta_0)^2 J \qquad \text{where } J := \ddot{G}(\theta_0).$$

The minimizing $\widehat{\theta}_n$ for G_n should be close to the value $\theta_0 - Z_n/(J\sqrt{n})$ that minimizes the quadratic. The standardized estimator $\sqrt{n}(\widehat{\theta}_n-\theta_0)$ should be close to $-Z_n/J$, which has an approximate $N(0,\sigma^2/J^2)$ distribution.

As you will se in later chapters, these heuristic arguments can often be made rigorous if we can gain some sort of uniform control over the errors in the approximations.

1.3 Efficiency heuristics

Heuristics::eff.heuristic

If the heuristics in the previous Section are to be believed, there is a wide class of estimators that have approximate normal distributions, with variances that decrease like $1/n$. It is natural to look for a g that gives the smallest possible multiple of $1/n$ for the approximate variance.

Actually, the task is slightly more complicated than choosing g to minimize the variance at a fixed $\theta_0 = \operatorname{argmin}_\theta P^x g(x,\theta)$. After all, we would not bother to estimate θ if we knew the distribution P exactly. The more challenging problem is to minimize the asymptotic variance simultaneously for a whole set of possible probability measures P.

It is traditional to consider models $\mathcal{P} = \mathbb{P}_{\theta,n}$ under which the X_i's are independent observations on P_θ, with $\{P_\theta : \theta\in\theta\}$ a prescribed set of probability measures. For that case, we need at least that the estimator $\widehat{\theta}_n$

converges in $\mathbb{P}_{\theta,n}$ probability to θ, for each θ in Θ (weak consistency). Then, considering the asymptotic variance as a function of θ, we need to find g to minimize that function at every θ.

Consider first the question of consistency. For independent observations from a fixed distribution P the heuristics suggested that $\widehat{\theta}_n$ converges in probability to $\operatorname{argmin}_s P^x g(x,s)$. For samples from P_θ the estimator should converge in $\mathbb{P}_{\theta,n}$-probability to the value of s that minimizes the function $s \mapsto P_\theta^x g(x,s)$. Temporarily call the minimizing value $s_0(\theta)$, with the θ to remind us that the expectation is calculated using the P_θ distribution. We should require $s_0(\theta) = \theta$ for every θ, that is,

$$\operatorname{argmin}_s P_\theta^x g(x,s) = \theta \qquad \text{for all } \theta \text{ in } \Theta.$$

That is,

$$P_\theta^x g(x,\theta) \le P_\theta^x g(x,s) \qquad \text{for all } s \text{ and all } \theta.$$

Some authors call this property ***Fisher consistency***.

Remark. Fisher consistency is a property of the function defined by g and the set of probability measures $\{P_\theta : \theta \in \theta\}$. It is not directly a large-sample property of an estimator.

The task becomes: given $\mathcal{P}$, find a Fisher consistent g to minimize $\sigma_g^2(\theta)/J_g(\theta)^2$ for all θ, where

$$\sigma_g^2(\theta) := P_\theta^x \dot{g}(x,\theta)^2 \qquad \text{and} \qquad J_g(\theta) := P_\theta^x \ddot{g}(x,\theta).$$

If P_θ has a density f_θ, and if differentiation under integral signs is justified, and if minima correspond to zeros of derivatives, then Fisher consistency implies

$$0 = \frac{\partial}{\partial s} P_\theta^x g(x,s)\big|_{s=\theta} = \frac{\partial}{\partial s} \int f_\theta(x) g(x,s)\big|_{s=\theta} = \int f_\theta(x) \dot{g}(x,\theta).$$

That is,

Fisher.consistent<2>

$$P_\theta \dot{g}(x,\theta) = 0 \qquad \text{for all } \theta.$$

Consequently $\sigma_g^2(\theta)$ is also the variance of $\dot{g}(x,\theta)$ under P_θ.

I assert that the constrained minimum of $\sigma_g^2(\theta)/J_u(\theta)^2$ is achieved for every θ by $g_0(x,\theta) := -\log f_\theta(x)$. That is, the lower bound for asymptotic variance is achieved by the g function that defines the maximum likelihood estimator. Jensen's inequality implies that g_0 is Fisher consistent:

$$P_\theta \left(g_0(x,\theta) - g_0(x,s)\right) = \int f_\theta(x) \log\left(\frac{f_s(x)}{f_\theta(x)}\right) \le \log \int f_\theta(x) \frac{f_s(x)}{f_\theta(x)} = 0.$$

The derivative

$$\ell_\theta(x) := -\dot{g}_0(x,\theta) = \frac{\partial}{\partial\theta}\log f_\theta(x) = \dot{f}_\theta(x)/f_\theta(x)$$

is called the ***score function*** for the model. By <2>,

$$P_\theta \ell_\theta(x) = 0 \qquad \text{for all } \theta.$$

The corresponding variance

$$\mathbb{I}(\theta) := \sigma^2_{g_0}(\theta) = \mathrm{var}_\theta(\ell_\theta) = P^x_\theta \ell_\theta(x)^2 = \int \dot{f}_\theta(x)^2/f_\theta(x)$$

is called the ***(Fisher) information function*** for the model. Moreover, from the assumed validity of yet another interchange of integration and differentiation,

$$\int \ddot{f}_\theta(x) = \frac{\partial^2}{\partial\theta^2}\int f_\theta(x) = \frac{\partial^2}{\partial\theta^2}1 = 0,$$

and the expression for the second derivative,

$$\ddot{g}_0(x,\theta) = -\dot{\ell}_\theta(x) = -\ddot{f}_\theta(x)/f_\theta(x) + \ell_\theta(x)^2,$$

it follows that

$$J_{g_0}(\theta) = -P^x_\theta \dot{\ell}_\theta(x) = -\int \ddot{f}_\theta(x) + \int \ell_\theta(x)^2 f_\theta(x) = \mathbb{I}(\theta)$$

Thus $\sigma^2_{g_0}(\theta)/J_{g_0}(\theta)^2 = 1/\mathbb{I}(\theta)$ and, according to the heuristics, the standardized maximum likelihood estimator $\sqrt{n}(\widehat{\theta}_n - \theta)$ has a limiting $N(0, 1/\mathbb{I}(\theta))$ distribution under $\mathbb{P}_{\theta,n}$.

To show that g_0 achieves the constrained lower bound, differentiate through the identity <2> to derive a weaker constraint,

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int f_\theta(x)\dot{g}(x,\theta) \\ &= \int f_\theta(x)\ddot{g}(x,\theta) + \int \dot{f}_\theta(x)\dot{g}(x,\theta) \\ &= J_g(\theta) + P_\theta \dot{g}(x,\theta)\ell_\theta(x). \end{aligned}$$

It follows, by the Cauchy-Schwarz inequality, that

$$J_g(\theta)^2 = \big(-P_\theta \dot{g}(x,\theta)\ell_\theta(x)\big)^2 \le P_\theta \dot{g}(x,\theta)^2 P_\theta \ell_\theta^2 = \sigma_g^2(\theta)\mathbb{I}(\theta)$$

or

$$\sigma^2(\theta)/J(\theta)^2 \ge 1/\mathbb{I}(\theta).$$

Remark. The preceding argument is close to the proof of the Cramér-Rao inequality (also known as the Information inequality) for finite samples. The differences are that Fisher consistency replaces the finite-sample unbiasedness assumption and the Cauchy-Schwarz inequality is applied to the limit distribution; the bound for the limit distribution is not just a limit of finite sample Cramér-Rao inequalities.

In summary: If we require the M-estimator $\widehat{\theta}_n$ to converge in probability to θ under independent sampling from P_θ with density f_θ, *for every* θ, then the asymptotic variance cannot be smaller than $1/\mathbb{I}(\theta)$, where

$$\mathbb{I}(\theta) := \operatorname{var}_\theta \left(\frac{\partial}{\partial \theta} \log f_\theta(X_1) \right) = -P_\theta \left(\frac{\partial^2}{\partial \theta^2} \log f_\theta(X_1) \right).$$

The asymptotic normal distribution for the maximum likelihood estimator has variance equal to the lower bound.

At least that is what the heuristics suggest.

I have not been rigorous about the conditions required for the arguments leading to "asymptotic optimality" of the maximum likelihood estimator amongst the class of M-estimators. For example, the argument is mostly nonsense when f_θ denotes the Uniform$(0, \theta)$ density, which is not everywhere differentiable.

As the next Section explains, optimality is a slippery concept even for models that seem unlikely candidates as troublemakers. A completely rigorous treatment can seem quite difficult—if one does not have the right tools. The development of the rigorous theory has been a major theme in modern theoretical statistics.

1.4 The concept of efficiency

Heuristics::efficiency.concept

If the heuristics are to be believed, in typical cases M-estimators cannot do better than mimic the limiting behaviour of the maximum likelihood estimator, which asymptotically achieves the information bound. No standardized estimator, $\sqrt{n}(\widehat{\theta}_n - \theta_0)$, should have an asymptotic normal distribution with a variance smaller than the reciprocal of the Fisher information. In some asymptotic sense, the maximum likelihood estimator might be called an efficient estimator.

In fact, it was long accepted in the statistics literature that the maximum likelihood estimator has similar optimality properties amongst an even wider class of estimators than described in the previous Section. As Fisher (1922,

page 277) put it, "The criterion of efficiency is satisfied by those statistics which, when derived from large samples, tend to a normal distribution with the least possible standard deviation." Unfortunately, the unqualified assertion about the limit distributions is not quite valid, even when the distributions have smooth densities, although it can be rescued. There exist estimators that beat the efficiency bound, as shown by a construction due to Hodges.

super.efficient <3>

Example. Suppose the X_i are independent with $N(\theta, 1)$ distribution under $\mathbb{P}_{\theta,n}$. The information function is identically 1. The maximum likelihood estimator $\widehat{\theta}_n$ is given by the sample mean, which has a $N(\theta, 1/n)$ distribution under $\mathbb{P}_{\theta,n}$. It achieves the efficiency lower bound, even for finite n.

Fix a θ_0 in Θ. Modify $\widehat{\theta}_n$ so that it performs superefficiently if θ_0 happens to be the true value, without disturbing its performance elsewhere. Let $\alpha_n = n^{-1/4}$, a sequence of positive real numbers that converges to zero more slowly than $1/\sqrt{n}$. Define

$$\theta_n^*(\mathbf{X}) := \widehat{\theta}_n(\mathbf{X})\{|\widehat{\theta}_n - \theta_0| > \alpha_n\} + \theta_0\{|\widehat{\theta}_n - \theta_0| \le \alpha_n\}.$$

Under the $\mathbb{P}_{\theta_0,n}$ model the modification takes effect with probability tending to one, $\mathbb{P}_{\theta_0,n}\{\theta_n^* = \theta_0\} \to 1$, which results in an estimator with obvious merits,

$$\sqrt{n}(\theta_n^* - \theta_0) \to 0 \qquad \text{in } \mathbb{P}_{\theta_0,n} \text{ probability.}$$

In particular, the efficiency bound is well beaten. Up to terms of order $o_p(n^{-1/2})$ the modified estimator behaves like the constant estimator, θ_0, when the true value is θ_0. Unlike the constant estimator, θ_n^* can adapt when the true value is not θ_0,

$$\mathbb{P}_{\theta,n}\{\theta_n^* = \widehat{\theta}\} \to 1 \qquad \text{if } \theta \ne \theta_0,$$

because $\widehat{\theta}_n$ then stays away from the shrinking neighborhood of θ_0. Under $\mathbb{P}_{\theta,n}$ for $\theta \ne \theta_0$, the estimator θ_n^* has the same asymptotic behaviour as $\widehat{\theta}$. The estimator θ_n^* achieves the efficiency bound at all points of Θ, except at θ_0, where it does much better than the naively formulated concept of efficiency would allow. It is superefficient.

The superefficient estimator does well at fixed points of Θ, but the modification has bad consequences at local alternatives $\{\theta_n\}$ that approach θ_0 at an $O(1/\sqrt{n}\,)$ rate through Θ. For $\mathbb{P}_{\theta_0,n}$, the slowly shrinking neighborhood $\{|\theta - \theta_0| \le \alpha_n\}$ was designed to capture $\widehat{\theta}_n$ and pull it to θ_0. The modification works because α_n decreases more slowly than the $O_p(1/\sqrt{n}\,)$ rate at

which $\widehat{\theta}_n$ converges to θ_0. Unfortunately, it has the same effect for $\mathbb{P}_{\theta_n,n}$, under which $\widehat{\theta}_n$ is distributed $N(\theta_n, 1/n)$, so that $\widehat{\theta}_n$ is still within $O_p(1/\sqrt{n})$ of θ_0. Thus $\mathbb{P}_{\theta_n,n}\{\theta_n^* = \theta_0\} \to 1$. In particular, if $\theta_n = \theta_0 + \delta_n/\sqrt{n}$ with $\delta_n \to \delta$, then

$$\sqrt{n}(\theta_n^* - \theta_n) \to -\delta \qquad \text{in } \mathbb{P}_{\theta_n,n} \text{ probability,}$$

which is not so good if $|\delta|$ is large. If we relax the definition of a local alternative, allowing δ_n that wander off to infinity more slowly than $\sqrt{n}\alpha_n$, we can even arrange

$$|\sqrt{n}(\theta_n^* - \theta_n)| \to \infty \qquad \text{in } \mathbb{P}_{\theta_n,n} \text{ probability.}$$

The estimator θ_n^* has achieved its superefficient status at the expense of poor behaviour under certain types of alternative.

□

A similar shrinkage construction can be applied in other parametric problems. In fact (Problem [1]), for rather general situations it is even possible to create a randomized estimator that is superefficient at a dense set of parameter values. Clearly the efficiency heuristics don't tell the whole story.

A requirement of good behavior under local sequences of alternatives was not part of Fisher's concept of efficiency. The requirement excludes the Hodges estimator and its ilk from the optimality competition. It plays a role in two modern approaches—the Convolution Theorem and the Local Asymptotic Minimax Theorem—to rescuing Fisher's idea. See Chapter 12.

The Convolution Theorem tightens up another assertion made by Fisher (1924) regarding efficient estimators. He described the effect of inefficient estimation as equivalent, asymptotically—a qualification that was seldom made explicit during the period when Fisher first contributed to the subject—to the addition of an independent source of error beyond what one should expect of an efficient estimator.

> Let A be the efficient statistic with variance σ^2/n, and B the inefficient statistic with variance σ^2/En; ... the correlation of A with $(B - A)$ is zero, so that the deviations of B from the population value may be regarded as made up of two parts: one, an error of random sampling, properly so called, is the deviation of A from the population value; the other, distributed independently of the first, is the error of estimation by which the inferior estimate, B, differs from the superior estimate, A.
>
> [Fisher 1924, page 446]

Fisher's assertion corresponds to an asymptotic comparison between a $\sqrt{n}$-consistent estimator $\widehat{\tau}_n$ and an efficient estimator $\widehat{\theta}_n$: under each $\mathbb{P}_{n,\theta_0}$,

$$\sqrt{n}(\widehat{\tau}_n - \theta_0) = \sqrt{n}(\widehat{\tau}_n - \widehat{\theta}_n) + \sqrt{n}(\widehat{\theta}_n - \theta_0),$$

where, in some sense, the two terms on the right-hand side should be asymptotically independent. If limiting distributions existed, we could interpret asymptotic independence to mean

Fisher.rep<4>
$$\left(\sqrt{n}(\widehat{\tau}_n - \widehat{\theta}_n), \sqrt{n}(\widehat{\theta}_n - \theta_0)\right) \rightsquigarrow (M, Z)$$

with Z distributed $N(0, 1/\mathbb{I}(\theta_0))$ independently of the random noise M. Consequently, we would have

Fisher.convolution<5>
$$\sqrt{n}(\widehat{\tau}_n - \theta_0) \rightsquigarrow M + Z.$$

This limit distribution is least dispersed when M is degenerate. For example, when variances are finite, as would be the case when M had a normal distribution, the equality

$$\mathbb{P}|M + Z|^2 = \mathbb{P}|M|^2 + \mathbb{P}|Z|^2$$

shows that the mean-squared error is a minimum if $M = 0$ almost surely. (An assumption of asymptotic normality was implicit in Fisher's concept of efficiency.)

More generally, if $\rho(\cdot)$ is nonnegative, symmetric, and convex, the symmetry of the distribution of Z gives

$$\mathbb{P}\rho(M + Z) = \tfrac{1}{2}\mathbb{P}\rho(M + Z) + \tfrac{1}{2}\mathbb{P}\rho(-M + Z) \geq \mathbb{P}\rho(Z),$$

with strict inequality if $\rho(\cdot)$ is strictly convex and if M is not degenerate at zero. Asymptotically efficient estimators are those for which $M \equiv 0$. The distribution of Z provides an asymptotic lower bound for the accuracy of estimation (Fisher's "error of estimation"). Only asymptotically efficient estimators $\widehat{\tau}_n$ can achieve that bound and then the difference $\sqrt{n}(\widehat{\tau}_n - \widehat{\theta}_n)$ converges in probability to zero; $\widehat{\tau}_n$ and $\widehat{\theta}_n$ are asymptotically equivalent up to terms of order $o_p(n^{-1/2})$.

Unfortunately, this second view of efficiency is also not quite valid, although it too can be rescued. The Convolution Theorem makes the rescue by requiring that $\sqrt{n}(\widehat{\theta}_n - \theta_n)$ have the same limiting distribution under $\mathbb{P}_{\theta_n,n}$ whenever $\sqrt{n}(\theta_n - \theta_0)$ has a finite limiting value.

The Asymptotic Minimax Theorem imposes a slightly different sort of local regularity, by consideration of the worst expected loss $\mathbb{P}_{\theta,n}\rho\left(\sqrt{n}(\widehat{\theta}_n - \theta)\right)$ over θ near θ_0.

In each of the rigorous approaches to asymptotic efficiency, one requires some form of local uniformity of good behavior, not just behavior at each fixed θ. Acceptable behaviour at local alternatives then rules out superlative behaviour at a single, fixed θ_0.

Problems

[1] [General version of Le Cam (1953, pages 286–289)] Index set $\Theta \subseteq \mathbb{R}^d$. Countable dense set $S = \{s_1, s_2, \dots\}$. Let $S_k = \{s_1, \dots, s_k\}$. For an estimator T_n with $(T_n - \theta)/\beta_n \rightsquigarrow Q_\theta$ under $\mathbb{P}_{\theta,n}$, with $\beta_n \to 0$, choose α_n with $\alpha_n/\beta_n \to \infty$. Construct $T^*_{n,k}$ by shrinkage towards S_k, then choose $T^*_n = T^*_{n,k}$ with probability c_k. Get mixture of Q_θ and δ_0 for limiting distribution of $(T^*_n - \theta)/\beta_n$. *dense.supereffici*

Notes

Hodges's shrinkage example was descibed by Le Cam (1953, page 280). Apparently the result was not published by Hodges himself. Check

References

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

Fisher, R. A. (1924). The conditions under which χ^2 measures the discrepency between observation and hypothesis. *Journal of the Royal Statistical Society 87*, 442–450.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics 35*, 73–101.

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

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics 20*, 595–601.