18.466 review session May 14, 2003: excerpts of sections 3.3-4.1. It's important to know all definitions given in this document. Several theorems are stated in simplified forms without listing all their assumptions. In place of such assumptions it's sufficient to say "under technical assumptions."

March 20, 2003

3.3 M-estimators and their consistency. A sequence of estimators T_n, one for each sample size n, possibly only defined for n large enough, is called *consistent* if for $X_1, X_2, \ldots$, i.i.d. (P_θ), $T_n = T_n(X_1, \ldots, X_n)$ converges in probability as $n \to \infty$ to a function $g(\theta)$ being estimated. This section will treat consistency of estimators which are more general than maximum likelihood estimators in two ways, first that the function being maximized may not be a likelihood, and second that it only needs to be approximately maximized.

It will be assumed that the parameter space Θ is a locally compact separable metric space with a metric d, such as an open or closed subset of a Euclidean space. $(X, \mathcal{A}, P)$ will be any probability space, and $h = h(\theta, x)$ is a measurable function on $\Theta \times X$ with values in the extended real number system $[-\infty, \infty]$. One example will be the negative of the log likelihood function, $h(\theta, x) \equiv -\log f(\theta, x)$. This will be called the *log likelihood case.* Let $X_1, X_2, \ldots$ be independent random variables with values in X and distribution P. A statistic $T_n = T_n(X_1, ..., X_n)$ with values in Θ will be called an *M-estimator* if

$$\tfrac{1}{n} \textstyle\sum_{i=1}^n h(T_n, X_i) \;=\; \inf_{\theta \in \Theta} \tfrac{1}{n} \textstyle\sum_{i=1}^n h(\theta, X_i).$$

Thus, in the log likelihood case, an M-estimator is a maximum likelihood estimator.

......

Statistics $T_n = T_n(X_1, ..., X_n)$ with values in Θ will be called a sequence of *approximate M-estimators* if as $n \to \infty$,

$$\tfrac{1}{n} \textstyle\sum_{i=1}^n h(T_n, X_i) - \inf_{\theta \in \Theta} \tfrac{1}{n} \textstyle\sum_{i=1}^n h(\theta, X_i) \to 0 \tag{3.3.1}$$

almost surely.

......

For any real function f, as usual let $f^+ := \max(f, 0)$ and $f^- := -\min(f, 0)$. A function $h(\cdot, \cdot)$ of x and θ will be called *adjusted* for P if

$$Eh(\theta, x)^- < \infty \quad \text{for all} \quad \theta \in \Theta, \quad \text{and} \tag{3.3.2}$$

$$Eh(\theta, x)^+ < \infty \quad \text{for some} \quad \theta \in \Theta. \tag{3.3.3}$$

To say that h is adjusted is equivalent to saying that $Eh(\theta, \cdot)$ is well-defined (possibly $+\infty$) and not $-\infty$ for all θ, and for some θ, also $Eh(\theta, \cdot) < +\infty$, so it is some finite real number.

If $a(\cdot)$ is a measurable real-valued function on X such that $h(\theta, x) - a(x)$ is adjusted for P, then $h(\cdot, \cdot)$ will be called *adjustable* for P and $a(\cdot)$ will be called an *adjustment function* for h and P. The next assumption is:

(A-3) $h(\cdot,\cdot)$ is adjustable for P.

From here on, if $h(\theta, x)$ is adjustable but not adjusted, let $\gamma(\theta) := \gamma_a(\theta) := E[h(\theta, x) - a(x)]$ for a suitable adjustment function $a(\cdot)$. As an example, let $h(\theta, x) := |x - \theta|$ for $\theta, x \in \mathbb{R}$. If P is a law on $\mathbb{R}$, such as the Cauchy distribution with density $(\pi(1 + x^2))^{-1}$, with $\int |x| dP(x) = +\infty$, then h itself is not adjusted and an adjustment function is needed. Let $a(x) := |x|$ in this case. Then for each θ, $|x - \theta| - |x|$ is bounded in absolute value (by $|\theta|$), so $\gamma(\theta)$ is defined and finite for all θ. Thus $|x|$ is in fact an adjustment function for any P.

......

Another assumption is:

(A-4) There is a $\theta_0 \in \Theta$ such that $\gamma(\theta) > \gamma(\theta_0)$ for all $\theta \neq \theta_0$.

θ_0 is called the *pseudo-true* value of θ.

.....

3.3.13 Theorem. Let $\{T_n\}$ be a sequence of approximate M-estimators. Assume (A-1) through (A-5) hold and T_n are measurable statistics. Then $T_n \to \theta_0$ almost surely.

......

To apply Theorem 3.3.13 to the case of maximum likelihood estimation the following will help. Let P and Q be two laws on a sample space $(X, \mathcal{B})$. Let

$$I(P, Q) := \int \log(R_{P/Q}) dP = -\int \log(R_{Q/P}) dP,$$

called the *Kullback-Leibler* information of P with respect to Q. Here we have $R_{P/Q} \equiv 1/R_{Q/P}$ with $1/0 := +\infty$ and $1/+\infty := 0$.

3.3.15 Theorem. Let $(X, \mathcal{B})$ be a sample space and P, Q any two laws on it. Then $I(P, Q) \geq 0$ and $I(P, Q) = 0$ if and only if $P = Q$.

.....

Consistency of approximate maximum likelihood estimators, under suitable conditions, does follow from Theorem 3.3.13, and assumption (A-3), and (A-4) for the true θ_0, will follow from Theorem 3.3.15 rather than having to be assumed:

3.3.16 Theorem. Assume (A-1) holds in the log likelihood case, so that $h(\theta, x) := -\log f(\theta, x)$. Also suppose $P = P_{\theta_0}$ for some $\theta_0 \in \Theta$ and $P_{\theta_0} \neq P_\theta$ for any $\theta \neq \theta_0$. Then h is always adjustable, with $a(x) = -\log f(\theta_0, x)$. Assume T_n are approximate maximum likelihood estimators, i.e. approximate M-estimators in this case. If (A-2) and (A-5) also hold, then the T_n are consistent.

.....

It turns out apparently to be simpler to treat exponential families directly rather than apply the above general theorems to them:

3.3.17 Theorem. Let $\{P_\theta,\ \theta \in \Theta\}$ be an exponential family in a minimal representation, where Θ is the interior of the natural parameter space, and $P = P_{\theta_0}$ for some $\theta_0 \in \Theta$. Then maximum likelihood estimators exist eventually a.s. and are consistent.

April 2, 2003

3.4 M-estimates and robust location estimates. M-estimators, as defined in Sec. 3.3, are sometimes called M-estimators *of ρ type* where the function $h(\theta, x)$ may also be called $\rho(\theta, x)$. Such estimators include maximum likelihood estimators as noted there. Another class of M-estimators is a class of estimators of location in $\mathbb{R}$ which are robust, meaning that they are not sensitive to contamination of the data by a few erroneous values, as will be seen.

Let $\psi(\theta, x)$ be a jointly measurable function, of θ in a parameter space Θ and x in a sample space X, where ψ has values in a Euclidean space $\mathbb{R}^k$. A statistic $T_n := T_n(X_1, \ldots, X_n)$ will be called an *M-estimator of ψ type* if $\sum_{i=1}^n \psi(T_n, X_i) \equiv 0$. Such an estimator is not necessarily an M-estimator of ρ type, but it is related in that if $\rho(\theta, x)$ has continuous first partial derivatives with respect to θ, then a necessary condition for T_n to be an M-estimator (of ρ type) is that it be one of ψ type with ψ = gradient of ρ. Under some rather special conditions, as for exponential families in Theorem 3.1.2, an M-estimator of ψ type where ψ is the gradient of ρ must also be one of ρ type. M-estimators of ψ type, where ψ is not necessarily a gradient, and the definition need only hold approximately as $n \to \infty$, will be further treated in the next two sections.

.....

3.4.1 Proposition. (a) For any law P on $\mathbb{R}$, m is a median of P if and only if $\int |x - \theta| - |x| dP(x)$ is minimized for $\theta = m$. (b) If $\int |x| dP(x) < \infty$, then m is a median of P if and only if $\int |x - \theta| dP(x)$ is minimized for $\theta = m$.

......

There is a class of M-estimators of location having some properties like those of the median, including robustness, but which are more often, or always, unique. A non-constant function $\rho(\theta, x)$ for x and θ real will be called a *wide-sense Huber function* if $\rho(\theta, x) \equiv \rho(|x - \theta|)$ where $\rho(x) \equiv \rho(-x)$, ρ is convex, and $\rho(x)/|x|$ is bounded as $|x| \to \infty$. The convexity and symmetry properties imply that ρ attains its absolute minimum at 0 (and perhaps elsewhere). Examples of wide-sense Huber functions include

(a) $\rho(x) := |x|$,

(b) $\rho(x) := (c^2 + x^2)^{1/2}$ for any real c, and

(c) $\rho(x) := x^2$ for $|x| \leq b$ and $\rho(x) := c|x| - d$ for $|x| > b$ where $b > 0$ and the other constants are chosen to make ρ continuously differentiable, so that $cb - d = b^2$ and $2b = c$, so $d = b^2$ and for $|x| > b$, $\rho(x) = b(2|x| - b)$.

Since Huber especially studied functions defined by (c), they might be called "narrow-sense Huber functions."

April 23, 2003

3.44 Robustness, breakdown points, and 1-dimensional location M-estimators.

.....

Let $X = (X_1, ..., X_n)$ and $Y = (Y_1, ..., Y_n)$ be samples of real numbers. For $j = 1, ..., n$ let $X =_j Y$ mean that $X_i = Y_i$ except for at most j values of i. More specifically, for

$y = (y_1, ..., y_j)$ let $X \ =_{j,y} Y$ mean that for some integers i_r with $1 \le i_1 < i_2 < ... < i_j \le n$, $Y_{i_r} = y_r$ for $r = 1, ..., j$ and $Y_i = X_i$ if $i \ne i_r$ for $r = 1, ..., j$. The idea is that in the latter case, X_i are i.i.d. from a nice distribution like a normal and y_r are errors or "bad" data. So the sample X contains $n - j$ good data points and j errors. A robust statistical procedure will be one that doesn't behave too badly if j is not too large compared to n.

"Breakdown point" is one of the main ideas in robustness. Let $T = T(X_1, ..., X_n)$ be a statistic taking values in some locally compact metric space Θ such as a Euclidean space. The closure of a set $A \subset \Theta$ will be denoted $\overline{A}$. If Θ is a Euclidean space then a set $A \subset \Theta$ has compact closure if and only if $\sup\{|x| : \ x \in A\} < \infty$. The *breakdown point* of T at X, or more specifically the *finite-sample breakdown point*, is defined as

$$\varepsilon^*(T, X) \ = \ \varepsilon^*(T; X_1, ..., X_n) = \frac{1}{n} \max\{j : \ \overline{\{T(Y) : \ Y \ =_j \ X\}} \ \text{ is compact}\}.$$

In other words $\varepsilon^*(T, X) = j/n$ for the largest j for which there is some compact set $K \subset \Theta$ such that $T(Y) \in K$ whenever $Y \ =_j \ X$. If $\varepsilon^*(T, X)$ doesn't depend on X, which is often the case, then let $\varepsilon^*(T) \ := \ \varepsilon^*(T, X)$ for all X. If Θ is a Euclidean space $\mathbb{R}^k$, then the compactness condition in the definition of ε^* is equivalent to

$$\sup\{|T(Y)| : \ Y \ =_j X\} \ < \ +\infty.$$

If a fraction of the data less than or equal to the breakdown point is bad (subject to arbitrarily large errors), the statistic doesn't change too much (it remains in a compact set), otherwise it can escape from all compact sets (in a Euclidean space, or by definition in other locally compact spaces, it can go to infinity). There are a number of definitions of breakdown point. The definition of finite-sample breakdown point as above is given in Hampel et al., 1986, p. 98, for a real-valued statistic.

Since j in the definition is an integer, the possible values of the breakdown point for samples of size n are $0, 1/n, 2/n, ..., 1$. A statistic with a breakdown point of 0 is (by definition) not robust. Larger values of the breakdown point indicate more robustness, up to just less than $1/2$. Finite-sample breakdown points $\ge 1/2$ are unattainable in some situations, e.g. Theorem 3.44.2 below.

Recall the definition of order statistics: for a sample $X_1, \ldots, X_n$ of real numbers, let $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ be the numbers arranged in order.

Examples. (i) For the sample mean $T = \bar{X} = (X_1 + ... + X_n)/n$, the breakdown point is 0 since for $j = 1$, if we let $y_1 \to \infty$ then $\bar{X} \to \infty$ (for n fixed).
(ii) Let $T = X_{(1)}$, the smallest number in the sample. Then the breakdown point of T is again 0 since for $j = 1$, as $y_1 \to -\infty$ we have $X_{(1)} \to -\infty$. Likewise the maximum $X_{(n)}$ of the sample has breakdown point 0.

So the statistics $\bar{X}$, $X_{(1)}$, $X_{(n)}$ are not robust. Other order statistics have some robustness (for fixed finite n):

Theorem 3.44.1. For sample size n, and each $j = 1, ..., n$, the order statistic $T = X_{(j)}$ has breakdown point $\varepsilon^*(T) = \frac{1}{n} \min(j - 1, n - j)$.

.....

For real-valued observations $X_1, \ldots, X_n$, a real-valued statistic $T = T(X_1, ..., X_n)$ will be called *equivariant for location* if for all real θ, and letting $X = (X_1, \ldots, X_n)$ and $X + \theta = (X_1 + \theta, ..., X_n + \theta)$,

$$T(X + \theta) \;=\; T(X) + \theta$$

for all n-vectors X of real numbers and all real θ.

For example, the order statistics $X_{(j)}$ and the sample mean $\bar{X}$ are clearly equivariant for location.

Theorem 3.44.2. For any real-valued statistic T equivariant for location, the breakdown point is $< 1/2$ at any $X = (X_1, ..., X_n)$.

.....

Now, we'll consider breakdown points of 1-dimensional location M-estimators. Let ψ be a real-valued function of a real variable which is odd (meaning $\psi(-x) \equiv -\psi(x)$), nondecreasing, nonconstant, and bounded. Then $\psi(-t) \le 0 = \psi(0) \le \psi(t)$ for all $t \ge 0$ and $\psi(-t) < 0 < \psi(t)$ for some $t > 0$ since ψ is nonconstant. We will have $\psi(t) \to A$ as $t \to +\infty$ for some $A > 0$. Examples of such functions ψ include the derivatives $\rho'(x)$ of wide-sense Huber functions (as defined in §3.4), where such derivatives are defined, with suitable choices where they are not defined, specifically, $\psi(0) = 0$ in all cases, $\psi(x) := \rho'(x+) := \lim_{h \downarrow 0}(\rho(x+h) - \rho(x))/h$ and $\psi(-x) := -\psi(x)$ for $x > 0$. Then for location, the psi function of two variables is defined by $\psi(\theta, x) := \psi(x - \theta)$, which is nonincreasing in θ. Given a sample $(X_1, \ldots, X_n)$, let

$$\theta^* \;:=\; \theta^*(X_1, ..., X_n) \;:=\; \sup\left\{\theta :\; \sum_{i=1}^{n} \psi(X_i - \theta) > 0\right\}.$$

This is finite since the sum is ≤ 0 for $\theta \ge X_{(n)}$ and also < 0 when $\theta \ge X_{(n)} + t$ for some t such that $\psi(t) > 0$. Analogously, define

$$\theta^{**} \;:=\; \theta^{**}(X_1, \ldots, X_n) := \inf\left\{\theta :\; \sum_{i=1}^{n} \psi(X_i - \theta) < 0\right\},$$

which is also finite since the sum is ≥ 0 for $\theta \le X_{(1)}$. We have $\theta^* \le \theta^{**}$ because of the monotonicity of ψ. Then a statistic $T_n = T_n(X_1, \ldots, X_n)$ will be an M-estimator of extended ψ type if and only if $\theta^* \le T_n \le \theta^{**}$. In order to have a unique estimator, *the* M-estimator defined by ψ and the sample will be defined, as for the sample median, by

$$\hat{\theta} \;:=\; \hat{\theta}((x_1, \ldots, x_n)\;) \;:=\; \frac{1}{2}(\theta^* + \theta^{**})((x_1, \ldots, x_n)\;).$$

As will be seen, such estimators have the same (finite sample) breakdown points as the median, converging to 1/2 as $n \to \infty$ and as large as possible. Consider also scale-adjusted M-estimators, where instead of $\sum_{i=1}^{n} \psi(X_i - \theta)$ we have $\sum_{i=1}^{n} \psi((X_i - \theta)/S)$ and S is a scale estimator, with nonnegative values. The resulting θ^*, θ^{**}, and $\hat{\theta}$ will be called θ^*_S, θ^{**}_S, and $\hat{\theta}_S := (\theta^* + \theta^{**})/2$ respectively. If $S = 0$, then by definition set

$$\psi((X_i - \theta)/S) := \begin{cases} A, & X_i > \theta \\ 0, & X_i = \theta \\ -A, & X_i < \theta. \end{cases}$$

It's easily seen that if $S = 0$ then the M-estimator $\hat{\theta}_S$ based on the above definitions is exactly the median.

To get a particular choice of S, let M be the median of the sample, defined as $X_{(k+1)}$ if $n = 2k + 1$ is odd, and $(X_{(k)} + X_{(k+1)})/2$ if $n = 2k$ is even. Let MAD denote the median absolute deviation, namely the median of $|X_i - M|$, and $S = \text{MAD}/.6745$, where the constant 0.6745 is (to the given accuracy) the median of $|Z|$ for a standard normal variable Z, and thus, S estimates the standard deviation σ for normally distributed data.

Theorem 3.44.5. Let ψ be a function from $\mathbb{R}$ into $\mathbb{R}$, which is odd, nondecreasing, nonconstant, and bounded. Then the M-estimator $\hat{\theta}$ defined by ψ has breakdown point $\frac{1}{2} - \frac{1}{n}$ if n is even and $\frac{1}{2} - \frac{1}{2n}$ if n is odd. The same holds for the scale-adjusted M-estimator $\hat{\theta}_S$ where we consider $\psi((X_i - \theta)/S)$ for the S just defined.

April 23, 2003

3.5 Consistency of approximate M-estimators of ψ type. As in Sec. 3.3, let $(X, \mathcal{A}, P)$ be a probability space and Θ a locally compact separable metric space. Let $\psi(\theta, x)$ be a function of x in X and $\theta \in \Theta$ with values in a Euclidean space $\mathbb{R}^m$. Let $X_1, X_2, \dots$ be independent with values in X and distribution P. A sequence of estimators $T_n := T_n(X_1, \dots, X_n)$ with values in Θ will be called *approximate M-estimators of ψ type* if

$$\frac{1}{n}\sum_{i=1}^n \psi(T_n, X_i) \to 0 \text{ almost surely as } n \to \infty. \tag{3.5.1}$$

Recall that if T_n are M-estimators of ψ type, the expression on the left in (3.5.1) equals 0, at least with probabilities converging to 1. Convergence of T_n to some θ_0 holds under some assumptions:

.....

(B-3) $\lambda(\theta) := E\psi(\theta, \cdot)$ is defined and finite for all θ, and for some θ_0, $\lambda(\theta_0) = 0$, while $\lambda(\theta) \neq 0$ for all $\theta \neq \theta_0$.

....

In (B-3), $\psi(\theta, \cdot)$ will be integrable for all P and θ if it is a bounded function of x for each θ. If ψ is bounded uniformly in x and θ, as for the classes of ψ functions with $-A \le \psi(\theta, x) \le A < +\infty$ considered in the 1-dimensional location case, so much the better.

......

3.5.4 Theorem. Let $\{T_n\}$ be a sequence of measurable approximate M-estimators of ψ type. If (B-1), (B-2), (B-3) and (B-4) hold, then $T_n \to \theta_0$ almost surely.

May 8, 2003

3.6 Asymptotic normality of M-estimates. First, let's note some of the conditions under which nonlinear functions of sample averages are asymptotically normal. Let f be a function from an open interval containing a point μ into $\mathbb{R}$. Suppose the derivative $f'(\mu)$ exists and is not 0. Let $X_1, X_2, \ldots ,$ be i.i.d. variables with mean μ and variance $0 < \sigma^2 := \sigma^2(X_1) < \infty$. Let $S_n := X_1 + \cdots + X_n$, and $\overline{X}_n := S_n/n$. Then $|\overline{X}_n - \mu| = O_p(n^{-1/2})$ by Chebyshev's inequality, and

$$\begin{aligned} f(\overline{X}_n) &= f(\mu) + f'(\mu)(\overline{X}_n - \mu) + o(|\overline{X}_n - \mu|), \text{ so} \\ n^{1/2}(f(\overline{X}_n) - f(\mu)) &= f'(\mu)((S_n - n\mu)/n^{1/2}) + o_p(1). \end{aligned}$$

Thus the distribution of the left side converges to $N(0, f'(\mu)^2\sigma^2)$ by the central limit theorem. This kind of reasoning is known as the "delta-method." To extend the method to vector-valued random variables, let f be a real-valued function on an open set $U \subset \mathbb{R}^k$. Then f is said to be *Fréchet differentiable* at a point $t \in U$ if there is a vector $v := f'(t) \in \mathbb{R}^k$ such that

$$f(u) = f(t) + v \cdot (u - t) + o(|u - t|)$$

as $u \to t$. For $k = 1$, this is equivalent to the usual derivative. For $k > 1$, the components of $f'(t)$ will be the partial derivatives $\partial f(u)/\partial u_i|_{u=t}$, forming the gradient of f at t. Each partial derivative is a directional derivative in the direction of a coordinate axis. Existence of the Fréchet derivative means that not only these partial derivatives exist, but the graph of f has a tangent hyperplane at $(t, f(t)) \in \mathbb{R}^{k+1}$, see Problem 2.

Using the central limit theorem in $\mathbb{R}^k$, the delta-method extends straightforwardly to $\mathbb{R}^k$-valued random variables having finite second moments.

If f takes values in $\mathbb{R}^m$ then the definition of Fréchet derivative is formally the same, but with the vector v replaced by a linear transformation from $\mathbb{R}^k$ into $\mathbb{R}^m$, given by an $m \times k$ matrix. The Fréchet differentiability of f is equivalent to that of each of its m component real-valued functions.

Next, asymptotic normality can be shown in quite general cases. Let Θ be an open subset of $\mathbb{R}^m$, $(X, \mathcal{A}, P)$ a probability space, and ψ a function from $\Theta \times X$ into $\mathbb{R}^m$. Let $X_1, X_2, \ldots$ be i.i.d. with values in X and distribution P. It is true under further assumptions that for a sequence $T_n = T_n(X_1, \ldots , X_n)$ of statistics with values in Θ, if

$$n^{-1/2} \textstyle\sum_{i=1}^n \psi(T_n, X_i) \to 0 \text{ in probability,} \tag{3.6.1}$$

then the distribution of T_n will converge to some normal law.

......

3.6.13 Lemma. Assume that (AN-1) through (AN-5) hold and T_n are estimators satisfying (3.6.1). Then

$$n^{1/2} \left(\lambda(T_n) + \tfrac{1}{n} \textstyle\sum_{i=1}^n \psi(\theta_0, X_i)\right) \to 0 \text{ in probability.}$$

.....

Since $\lambda(\cdot)$ takes the open set $\Theta \subset \mathbb{R}^m$ into $\mathbb{R}^m$, its Fréchet derivative at θ_0, if it exists, is a linear transformation A from $\mathbb{R}^m$ into itself, given by an $m \times m$ matrix. Note that if A exists and is non-singular, then (AN-4)(i) follows. Let B' denote the transpose of a matrix B.

3.6.15 Theorem. Assume (AN-1) through (AN-5), that T_n satisfy (3.6.1), and that λ has a non-singular Fréchet derivative A at θ_0. Then $n^{1/2}(T_n - \theta_0)$ is asymptotically normal with mean 0 and covariance matrix $A^{-1}C(A^{-1})'$, where C is the covariance matrix of $\psi(\theta_0, x)$.

May 8, 2003

3.7 Efficiency of estimators. In this and the following two sections the distribution of the data is assumed to belong to a parametric family $\{P_\theta,\ \theta \in \Theta\}$, having densities $f(\theta, x)$.

The information inequality or Fréchet-Cramér-Rao lower bound, when Θ is an open interval in $\mathbb{R}$ and g is a differentiable real-valued function on Θ, is

$$\mathrm{var}_\theta(T_n) \ \geq \ g'(\theta)^2/(nI_1(\theta)),$$

where $I_1(\theta) \ := \ E_\theta((\partial f(\theta,x)/\partial\theta)^2)$, as was proved in Theorem 2.4.10 under some regularity conditions when T_n is an unbiased estimator of $g(\theta)$. But by Theorem 2.4.15, if $\log f(\theta, x)$ is C^1 in θ, the lower bound is attained for all θ only when the family of distributions is exponential of order 1 with $T(x)$ equal to the given estimator $T_n(x)$ where $x = (x_1, \ldots, x_n)$. When this is true for one function $T(\cdot)$, the only other functions for which it holds are $aT(\cdot)+b$ where $a \neq 0$ and b are constants. So the only functions having unbiased estimators attaining the information inequality lower bound for all θ are $ag(\theta) + b$ where now a and b are any constants and g is the specific function $d \log K(\theta)/d\theta$ for which T is the unbiased estimator, by Corollary 2.5.9. Even for exponential families of order 1, unique unbiased, admissible estimators (for other functions) may be unsatisfactory, as in the example at the end of Sec. 2.5.

If the information inequality provided best possible lower bounds for mean-square errors only for estimating functions $ag(\theta)+b$ as just described, it would not be very useful. There is, however, an *asymptotic* lower bound,

$$\liminf_{n\to\infty} E_\theta([n^{1/2}(T_n - g(\theta))]^2) \ \geq \ g'(\theta)^2/I(\theta), \tag{3.7.1}$$

where $I(\theta) \equiv I_1(\theta)$, which is valid under rather general conditions, without unbiasedness, as will be shown here first for $g(\theta) \equiv \theta$, so $g'(\theta) \equiv 1$, in Theorem 3.7.3, then for more general g in Theorem 3.7.9.

......

For the family of laws $P_\theta = U[\theta, \theta+1]$ on $\mathbb{R}$, there exist (unbiased) estimators of θ with mean-square error of order $1/n^2$ (Sec. 2.4, Problem 3). Thus some regularity conditions (equivalence, differentiability in θ) cannot both just be removed.

Let $L(\theta, x) \ := \ \log f(\theta, x)$. Derivatives with respect to θ will be denoted by primes, so that $L'(\theta, x) \ := \ \partial L(\theta, x)/\partial\theta$, etc. Then by (AV-1) and (AV-2), $L(\theta, x)$ is a C^2 function of θ for any $x \in B$. The Fisher information $I(\theta) = E_\theta(L'(\theta, x)^2)$ as defined in Sec. 2.4.

.....

Let $\{T_n\}_{n\geq 1}$ be a sequence of estimators (statistics), so that for each n, T_n is measurable from X^n into Θ. It will be assumed that the T_n are consistent estimators of θ, at least in probability, and are asymptotically normal:

(AV-6) For each θ, there is a $v(\theta)$ with $0 < v(\theta) < \infty$ such that as $n \to \infty$, the distribution of $n^{1/2}(T_n - \theta)$ under P_θ^n converges to $N(0, v(\theta))$.

.....

Assuming asymptotic normality (AV-6), if

$$v(\theta) \;\geq\; 1/I(\theta) \tag{3.7.2}$$

holds, then so does (3.7.1) for $g(\theta) \;\equiv\; \theta$.

The next theorem will give an almost everywhere lower bound on efficiency of estimators of a 1-dimensional parameter.

.....

3.7.3 Theorem. Under assumptions (AV-1) through (AV-6), (3.7.2) holds for almost all θ in the open interval Θ for Lebesgue measure.

......

Next, Theorem 3.7.3 will be extended to estimators of functions $g(\theta)$, by the delta-method. The factor $g'(\theta)^2$ is familiar from information inequalities (Section 2.4). Note that (AV-1) through (AV-5) don't mention any estimators T_n.

3.7.9 Theorem. Assume (AV-1) through (AV-5). Let g be a C^1 function: $\Theta \to \mathbb{R}$. Suppose that for each $\theta \in \Theta$, there is a $w(\theta) \geq 0$ such that for each θ with $g'(\theta) \neq 0$, $0 < w(\theta) < \infty$ and the distribution of $\sqrt{n}(T_n - g(\theta))$ under P_θ^n converges to $N(0, w(\theta))$. Then for Lebesgue almost all $\theta \in \Theta$, $w(\theta) \geq g'(\theta)^2/I(\theta)$.

.....

If A is a $k \times m$ matrix, then A' denotes its transpose, with $(A')_{ij} \;:=\; A_{ji}$ for $i = 1, \ldots, m$, $j = 1, \ldots, k$. In particular, if x is a row vector $(x_1, \ldots, x_m)$ then x' is the corresponding column vector, and vice versa. In fact, elements of $\mathbb{R}^m$ will usually be taken as column vectors y, so that y' is the corresponding row vector. Matrix multiplication is written by juxtaposition. Thus for $x, y \in \mathbb{R}^m$, $x'y = \sum_{j=1}^m x_j y_j$ is the usual dot product $x \cdot y$. If $x, y \in \mathbb{R}^m$ and C is an $m \times m$ matrix, then $x'Cy$ is the number $\sum_{i,j=1}^m C_{ij} x_i y_j$.

The Fisher information for a single parameter extends to the Fisher information matrix for several parameters, defined as follows. Let Θ be an open set in $\mathbb{R}^m$. For $\theta \;:=\; (\theta_1, \ldots, \theta_m)$, let

$$I(\theta)_{ij} \;:=\; E_\theta\left(\frac{\partial L(\theta, x)}{\partial \theta_i}\frac{\partial L(\theta, x)}{\partial \theta_j}\right) \tag{3.7.10}$$

if the partial derivatives exist and have finite variances. Alternate forms of I_{ij} are

$$I(\theta)_{ij} \;=\; E_\theta\left(\left(\frac{\partial R_{\phi,\theta}}{\partial\phi_i}\frac{\partial R_{\phi,\theta}}{\partial\phi_j}\right)\Big|_{\phi=\theta}\right) \;=\; \int \frac{\partial f(\theta,x)}{\partial\theta_i}\frac{\partial f(\theta,x)}{\partial\theta_j}\frac{1}{f(\theta,x)}d\nu(x).$$

......

(AC-3) Let $L(\theta,x) \;:=\; \log f(\theta,x)$. For each $\theta\in\Theta$, the Fisher information matrix $I(\theta)$ as defined by (3.7.10) exists and is strictly positive definite. Also, $E_\theta(\nabla_\theta L(\theta,x)) = 0$ for the gradient of L.

(AC-4) $\{E_\theta\partial^2 L(\theta,x)/\partial\theta_i\partial\theta_j\}_{i,j=1}^m \;=\; -I(\theta)$ for all $\theta\in\Theta$.

.....

(AC-6) T_n are estimators of $\theta\in\Theta$ such that for each θ, the distribution of $n^{1/2}(T_n-\theta)$ under $\Pr_\theta$ converges as $n\to\infty$ to some multivariate normal law $N(0,v(\theta))$ where $v(\theta)$ is a nonnegative definite symmetric matrix.

3.7.11 Theorem. Assume (AC-1) through (AC-6). Then for Lebesgue almost all $\theta\in\Theta$, $v(\theta)-I^{-1}(\theta)$ is nonnegative definite. Thus, $v(\theta)$ is positive definite.

May 6, 2003

3.8 Efficiency of maximum likelihood estimators. Let $K >> M$ for $m\times m$ matrices K, M mean that $K-M$ is nonnegative definite. Let T_n be a sequence of estimators such that the distribution of $\sqrt{n}(T_n-\theta)$ under $\Pr_\theta$ is asymptotically $N(0,v(\theta))$. By Theorem 3.7.11, under its assumptions, $v(\theta) >> I(\theta)^{-1}$ for Lebesgue almost all θ. Thus, the sequence $\{T_n\}$ will be called "efficient" if for all θ, under $\Pr_\theta$, $\sqrt{n}(T_n-\theta)$ is asymptotically $N(0,v(\theta))$ with $I(\theta)^{-1} >> v(\theta)$. In practice, efficient estimators will have $v(\theta) = I(\theta)^{-1}$ for all θ. The definition allows for superefficiency for some set of θ which, under the conditions of Sec. 3.7, will have Lebesgue measure 0. The efficiency of maximum likelihood estimators with $v(\theta)\equiv I(\theta)^{-1}$ holds under some assumptions.

.....

The observations $X_1, X_2, \ldots,$ are i.i.d. (P_{θ_0}) for some $\theta_0\in\Theta$. Let $L(\theta,x) \;:=\; \log f(\theta,x)$ and $\psi(\theta,x) \;:=\; \nabla_\theta L(\theta,x)$ where ∇_θ denotes gradient with respect to θ.

(EML-2) For each $x\in X$, $f(\cdot,x)$ is C^1 with respect to θ, and the Fisher information matrix $I(\cdot)$ exists on Θ and is continuous and non-singular at θ_0.

If $E_\theta(\nabla_\theta L(\theta,x) = 0$, which is shown in Theorem 3.8.1 to follow from the given assumptions, then $I(\theta)$ is the covariance matrix C of $\psi(\theta,x)$.

(EML-3) $\{T_n\}$ is a sequence of maximum likelihood estimators and is consistent, in other words $T_n\to\theta$ in $\Pr_\theta$-probability as $n\to\infty$ for all θ.

Conditions for consistency of M-estimators were given in Sections 3.3 and 3.5.

Conditions (AN-4) and (AN-5)(ii) in Section 3.6 will be assumed, locally uniformly in θ_0. Specifically, recall that for $\delta>0$ small enough, depending on θ,

$$u(\theta,x,\delta) \;:=\; \sup\{|\psi(\eta,x)-\psi(\theta,x)| :\; |\eta-\theta|\le\delta\}.$$

(EML-4) (i) For each $\theta, \phi \in \Theta$, $\lambda_\phi(\theta) := E_\phi \psi(\theta, x)$ exists in $\mathbb{R}^m$. Let $\lambda(\cdot) := \lambda_{\theta_0}(\cdot)$.

.....

As in Theorem 3.6.15, let A be the Fréchet derivative of $\lambda(\cdot)$ at θ_0 if it exists.

3.8.1 Theorem. Assume (EML-1) through (EML-5). Then $\lambda(\theta_0) = 0$ and A exists with $A = -I(\theta_0)$. Also, the distribution of $\sqrt{n}(T_n - \theta_0)$ converges to $N(0, I(\theta_0)^{-1})$ as $n \to \infty$.

May 12, 2003

3.9 A likelihood ratio test for nested composite hypotheses: Wilks's theorem. Let Θ be a d-dimensional parameter space, specifically, an open set in $\mathbb{R}^d$. Let H_0 be a k-dimensional subset of Θ, in a sense to be made more precise below, for some $k < d$. For example, H_0 could be the intersection with Θ of a k-dimensional flat hyperplane. Let $\{P_\theta, \ \theta \in \Theta\}$ be an equivalent family of laws on a sample space $(X, \mathcal{B})$ with a likelihood function $f(\theta, x) > 0$ for all $\theta \in \Theta$ and $x \in X$.

Assume that observations $X_1, \dots, X_n$ are i.i.d. P_θ for some $\theta \in \Theta$. We want to test the hypothesis that $\theta \in H_0$. S. S. Wilks proposed the following test: let $L(\theta, x) := \log f(\theta, x)$ be the log likelihood. For n observations, let the maximum log likelihoods over Θ and H_0 be respectively

$$MLL_d := \sup_{\theta \in \Theta} \sum_{j=1}^n L(\theta, X_j), \qquad MLL_k := \sup_{\theta \in H_0} \sum_{j=1}^n L(\theta, X_j).$$

Let $W := 2(MLL_d - MLL_k)$. Wilks found that if the hypothesis H_0 is true, then the distribution of W converges as $n \to \infty$ to a χ^2 distribution with $d - k$ degrees of freedom, not depending on the true $\theta = \theta_0 \in H_0$. Thus, H_0 would be rejected if W is too large in terms of the tabulated χ^2_{d-k} distribution.

It turns out that Wilks's conclusion can be proved under the same assumptions as are used to prove the lower bounds on asymptotic efficiency of estimators in Section 3.7 and efficiency of maximum likelihood estimators in Section 3.8.

.....

3.9.1 Theorem (Wilks's theorem). Assume (AC-1) through (AC-5) in Section 3.7 and (EML-1) through (EML-5) in Section 3.8 for Θ where in (EML-3), T_n are maximum likelihood estimators of $\theta \in \Theta$. Let H_0 be a smooth k-dimensional subset of Θ containing θ_0 for some $k < d$. Let U_n be maximum likelihood estimators of η in H_0, assumed to exist and be unique with probability converging to 1 as $n \to \infty$. Assume also that $U_n \to \theta_0$ in probability as $n \to \infty$.

Then as $n \to \infty$, the distribution of W converges to a χ^2_{d-k} distribution.

Here is a summary of section 4.1 on consistency of posteriors: Posteriors $\pi_{n,x}$ are said to be consistent if for every neighborhood U of the true parameter θ_0, $\pi_{x,n}(U) \to 1$ almost surely as $n \to \infty$. Theorem 4.1.4 says that under very general conditions, this holds for π-almost all θ_0 where π is the prior. Theorem 4.1.1 gives sufficient conditions, similar to those for consistency of M-estimates in section 3.3, so that we have consistency of posteriors for all $\theta_0 \in \Theta$. Proposition 4.1.2 and an example after it show, however, that consistency can fail at an individual θ_0 if as $\theta_m \to \theta_0$, although $f(\theta_m, x) \to f(\theta_0, x)$ for all x, the densities $f(\theta_m, \cdot)$ are moving away from $f(\theta_0, \cdot)$ in terms of Kullback-Leibler distance, and if the prior π gives probabilities to neighborhoods U of θ_0 that approach 0 very fast as U shrinks to θ_0.