Statistics 210B, Spring 1998

Class Notes

P.B. Stark

stark@stat.berkeley.edu

www.stat.berkeley.edu/~stark/index.html

January 27, 1999

Twelfth Set of Notes

1 Robustness and related topics

For references, see

Hampel, F.R., Rousseeuw, P.J., Ronchetti, E.M., and Strahel, W.A., 1986. *Robust Statistics: The approach based on influence functions*, Wiley, NY.; Huber, P.J., 1981. *Robust Statistics*, Wiley, N.Y. Bickel

1.1 Heuristics

The optimality theory we studied at the beginning of the course is predicated upon the assumption that $F \in \mathcal{F}$, which was often a parametric family of distributions. An estimator or test that is optimal for some family of distributions can have terrible performance if the "model" $F \in \mathcal{F}$ is wrong, even by a small amount (in some distance on measures). In a loose sense, an estimator or test is *robust* if its performance is good over a neighborhood of distributions including the family in which the truth is modeled to lie.

Virtually every real data set has some "gross outliers," which are in some sense corrupted measurements. Data can be transcribed incorrectly, data transmission can corrupt bits, cosmic rays can hit satellite-borne instruments, power supplies can have surges, operators can miss their morning cups of coffee, *etc.* Even without gross outliers, there is always a limit on the precision with which data are recorded, leading to truncation or rounding errors that make continuous models for data error distributions only approximate. Furthermore, parametric error models are rarely dictated by direct physical arguments; rather, the central limit theorem is invoked, or some historical becomes standard in the field.

One early attempts to deal with gross outliers is due to Sir Harold Jeffreys, who (in the 1930s) modeled data errors as a mixture of two Gaussian distributions, one with variance tending to infinity. The idea is that the mixture fraction of that Gaussian represents a fraction of gross outliers possibly present in the data; one wants to minimize the influence of such observations on the resulting estimate or test. Considerations in favor of fitting to minimize mean absolute deviation instead of least squares go back much further.

We will be looking at ways of quantifying robustness, and of constructing procedures whose performance when the model is true is not much worse than the optimal procedure, but whose performance when the model is wrong (by a little bit) is not too bad, and is often much better than the performance of the optimal procedure in that event.

The kinds of questions typically asked in the robustness literature are:

1. Is the procedure sensitive to small departures from the model?

2. To first order, what is the sensitivity?

3. How far from the model can one go before the procedure produces garbage?

The first issue is that of qualitative robustness; the second is quantitative robustness; the third is the "breakdown point."

1.2 Resistance and Breakdown Point

Resistance has to do with changes to the observed data, rather than to the theoretical distribution underlying the data. A statistic is *resistant* if arbitrary changes to a few data

(such as might be caused by *gross outliers*, or small changes to all the data (such as might be caused by rounding or truncation), result in only small changes to the value of the statistic.

Suppose we are allowed to change the values of the observations in the sample. What fraction would we need to change to make the estimator take an arbitrary value?

For example, consider the sample mean $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ as an estimator of the mean of F ($T(F) = \int x dF; \theta = T(F_\theta)$.) By corrupting a single observation, we can make the sample mean take any real value: the breakdown point is $\frac{1}{n}$. In contrast, consider the "α-trimmed mean," defined as follows: Let $k_\ell = \lfloor \alpha n \rfloor$ and $k_h = \lceil \alpha n \rceil$. Let $X_{(j)}$ be the jth order statistic of the data. Define

$$\bar{X}_\alpha = \frac{1}{k_h - k_\ell + 1} \sum_{j=k_\ell}^{k_h} X_{(j)}. \tag{1}$$

This measure of location is less sensitive to outliers than is the sample mean: the breakdown point is $\min(k_\ell, n - k_h + 1)/n$. An alternative, not necessarily equivalent, definition of the α-trimmed mean is through the functional

$$T(F) = \frac{1}{1-2\alpha} \int_\alpha^{1-\alpha} F^{-1}(t) dt. \tag{2}$$

This version has breakdown point α.

We shall make the notion of breakdown point more precise presently; a few definitions are required.

Definition 1 *The Lèvy distance between two distribution functions F and G on $\mathbf{R}$ is*

$$\lambda(F, G) = \inf\{\epsilon : F(x - \epsilon) - \epsilon \leq G(x) \leq F(x + \epsilon) + \epsilon \ \ \forall x \in \mathbf{R}\}. \tag{3}$$

Definition 2 *A* Polish space $\mathcal{X}$ *is a complete, separable topological space whose topology is metrizable by a metric d.*

Examples of Polish spaces include $\mathbf{R}^k$. Let $\mathcal{M}$ denote the space of probability measures the Borel σ-algebra $\mathcal{B}$ of subsets of $\mathcal{X}$. ($\mathcal{B}$ is the smallest σ-algebra containing all the open sets in $\mathcal{X}$.) Let $\mathcal{M}'$ denote the set of finite signed measures on $(\mathcal{X}, \mathcal{B})$; this is the linear space of measures generated by $\mathcal{M}$. The measures in $\mathcal{M}$ are *regular* in the sense that for every $F \in \mathcal{M}$,

$$\sup_{C \subset B;\, C \text{ compact}} F(C) = F(B) = \inf_{G \supset B;\, G \text{ open}} F(G). \tag{4}$$

The weak-star topology in $\mathcal{M}$ is the weakest topology for which the functional

$$\int \psi dF \tag{5}$$

is continuous for every continuous, bounded function $\psi : \mathcal{X} \to \mathbf{R}$.

In this section, we assume that $\mathcal{X}$ is a Polish space, and that all measures are defined on $\mathcal{B}$. An overbar (*e.g.,* $\bar{A}$) will denote topological closure, and the superscript c will denote complementation ($A^c = \{x \in \mathcal{X} : x \notin A\}$).

Definition 3 *For any subset A of the sample space $\mathcal{X}$ and any metric $d : \mathcal{X} \times \mathcal{X} \to \mathbf{R}^+$, the* closed ϵ-neighborhood of A *is*

$$A^\epsilon = \{x \in \mathcal{X} : \inf_{a \in A} d(x, a) \leq \epsilon\}. \tag{6}$$

It will be important presently that

$$A^\epsilon = \bar{A}^\epsilon = \bar{A}^\epsilon = \bar{\bar{A}}^\epsilon. \tag{7}$$

Definition 4 *The Prohorov distance between two measures F and G defined on a common algebra $\mathcal{A}$ of subsets of a metric space $\mathcal{X}$ is*

$$\pi(F, G) = \inf\{\epsilon \geq 0 : F(A) \leq G(A^\epsilon) + \epsilon \; \forall A \in \mathcal{A}\} \tag{8}$$

Expanding the events by ϵ to form A^ϵ corresponds to the measure G being "shifted" slightly from F, for example, by rounding. The addition of ϵ corresponds to a fraction ϵ of the observations being from a completely different distribution.

We shall verify that the Prohorov distance really is a metric if the sample space $\mathcal{X}$ is a Polish space. Clearly, it is nonnegative, and $\pi(F, F) = 0$. We need to show symmetry, the triangle inequality, and that $\{\pi(F, G) = 0\} \Rightarrow \{F = G\}$. The following proof follows that in Huber (1981).

Symmetry. This will follow immediately if we can show that if $F(A) \leq G(A^\epsilon) + \epsilon$ for all $A \in \mathcal{A}$, then $G(A) \leq F(A^\epsilon) + \epsilon$ for all $A \in \mathcal{A}$. Recall that because $\mathcal{A}$ is an algebra, if $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$ as well. Take any $\delta > \epsilon$, and consider $A = (B^\delta)^c = B^{\delta c}$ for any $B \in \mathcal{B}$. Note that $A \in \mathcal{B}$, so by the premise,

$$F(B^{\delta c}) \leq G(B^{\delta c \epsilon}) + \epsilon, \tag{9}$$

or

$$1 - F(B^\delta) \leq 1 - G(B^{\delta c \epsilon c}) + \epsilon \tag{10}$$

$$G(B^{\delta c \epsilon c}) \leq F(B^\delta) + \epsilon. \tag{11}$$

However, $B \subset B^{\delta c \epsilon c}$, as we shall see. This statement is equivalent to $B^{\delta c \epsilon} \subset B^c$. This is essentially immediate from $\delta > \epsilon$: if $x \in B^{\delta c \epsilon}$, then $\exists y \notin B^\delta$ s.t. $d(x,y) < \epsilon$ (typo in Huber here). Thus $x \in B^c$, because otherwise $d(x,y) > \delta > \epsilon$. Thus

$$G(B) \leq G(B^{\delta c \epsilon c}) \leq F(B^\delta) + \epsilon. \tag{12}$$

But $B^\epsilon = \cap_{\delta > \epsilon} B^\delta$, so the result follows.

To show that $\pi(F,G) = 0 \Rightarrow F = G$, note that the closure of A is $\bar{A} = \cap_{\epsilon > 0} A^\epsilon$. Thus $\pi(F,G) = 0$ implies $F(A) \leq G(A)$ and $G(A) \leq F(A)$ for all closed sets $A \in \mathcal{A}$.

Triangle inequality. If $\pi(F,G) \leq \epsilon$ and $\pi(G,H) \leq \delta$, then for every $A \in \mathcal{B}$,

$$F(A) \leq G(A^\epsilon) + \epsilon \leq H((A^\epsilon)^\delta) + \epsilon + \delta. \tag{13}$$

But by the triangle inequality for the metric d on $\mathcal{X}$, $(A^\epsilon)^\delta \subset A^{\epsilon+\delta}$, so we are done.

Note that the Prohorov distance between $\hat{F}_n$ and the "contaminated" empirical distribution one gets by changing k of the data by an arbitrary amount is $\frac{k}{n}$.

Theorem 1 *(Strassen, 1965; see Huber Thm 2.3.7.) The following are equivalent:*

1. *$F(A) \leq G(A^\delta) + \epsilon$ for all $A \in \mathcal{B}$*
2. *There are dependent $\mathcal{X}$-valued random variables X, Y such that $\mathcal{L}(X) = F$, $\mathcal{L}(Y) = G$, and $\mathbf{P}\{d(X,Y) \leq \delta\} \geq 1 - \epsilon$. (here $\mathcal{L}(X)$ denotes the probability law of X,* etc.

Definition 5 *Suppose that the distance function d on $\mathcal{X} \times \mathcal{X}$ is bounded by one (one can replace d by $d(x,y)/(1+d(x,y))$ to make this so). The* bounded Lipschitz metric *on $\mathcal{M}$ is*

$$d_{BL}(F,G) = \sup_{\phi : |\psi(x) - \psi(y)| \leq d(x,y)} \left| \int \psi dF - \int \psi dG \right|. \tag{14}$$

This, too, is truly a metric on $\mathcal{M}$.

Theorem 2 *The set of regular Borel measures $\mathcal{M}$ on a Polish space $\mathcal{X}$ is itself a Polish space with respect to the weak topology, which is metrizable by the Prohorov metric and by the bounded Lipschitz metric.*

Consider a collection of probability distributions indexed by ϵ, such as the Prohorov neighborhood

$$\mathcal{P}_\pi(\epsilon; F) = \{G \in \mathcal{M} : \pi(F, G) \leq \epsilon\} \tag{15}$$

or the "gross error contamination neighborhood" (not truly an neighborhood in the weak topology)

$$\mathcal{P}_{\text{gross error}}(\epsilon; F) = \{G \in \mathcal{M} : G = (1-\epsilon)F + \epsilon H, H \in \mathcal{M}\}. \tag{16}$$

Let $M(G, T_n)$ denote the median of the distribution of $T_n(G) - T(F)$. Let $A(G, T_n)$ denote some fixed percentile of the distribution of $|T_n(G) - T(F)|$.

Definition 6 *Consider a sequence $\{T_n\}$ of estimators that is Fisher consistent and converges in probability to a functional statistic T. The* maximum bias of $\{T_n\}$ at F over the collection $\mathcal{P}(\epsilon)$ *is*

$$b_1(\epsilon) = b_1(\epsilon, \mathcal{P}, F) = \sup_{G \in \mathcal{P}(\epsilon)} |T(G) - T(F)|. \tag{17}$$

The maximum asymptotic bias of $\{T_n\}$ at F over the collection $\mathcal{P}(\epsilon)$ *is*

$$b(\epsilon) = b(\epsilon, \mathcal{P}, F) = \lim_{n \to \infty} \sup_{G \in \mathcal{P}(\epsilon)} |M(G, T_n)|. \tag{18}$$

If b_1 is well defined, $b(\epsilon) \geq b_1(\epsilon)$. Note that for the gross-error model and for the Lèvy and Prohorov distances, $b(\epsilon) \leq b(1)$, because the set $\mathcal{P}(1) = \mathcal{M}$.

Definition 7 *(Following Huber, 1981.) For a given collection $\mathcal{P}(\epsilon)$ of distributions indexed by $\epsilon \geq 0$, the* asymptotic breakdown point *of T at F is*

$$\epsilon^* \equiv \epsilon^*(F, T, \mathcal{P}(\cdot)) = \sup\{\epsilon : b(\epsilon, \mathcal{P}(\epsilon), F) < b(1)\}. \tag{19}$$

Definition 8 *(Following Hampel* et al.*, 1986.) The* breakdown point *of a sequence of estimators $\{T_n\}$ of a parameter $\theta \in \Theta$ at the distribution F is*

$$\epsilon^*(T_n, F) \equiv \sup\{\epsilon \leq 1 : \exists K_\epsilon \subseteq \Theta,\ K_\epsilon \text{ compact, s.t. } \pi(F, G) < \epsilon \Rightarrow G(\{T_n \in K_\epsilon\}) \to 1 \ \text{ as } n \to \infty\}. \tag{20}$$

That is, the breakdown point is the largest Prohorov distance from F a distribution can be, and still have the estimators almost surely take values in a given compact as $n \to \infty$.

Definition 9 *The* finite-sample breakdown point of the estimator T_n at $x = (x_j)_{j=1}^n$ *is*

$$\epsilon^*(T_n, x) \equiv \frac{1}{n} \max \left\{ m : \max_{i_1,\ldots,i_m} \sup_{y_1,\ldots,y_m} |T(z_1, \ldots, z_n)| < \infty \right\}, \tag{21}$$

where

$$z_j = \begin{cases} x_j, & j \notin \{i_k\}_{k=1}^m \\ y_k, & j = i_k \text{ for some } k. \end{cases} \tag{22}$$

This definition makes precise the notion that corrupting some fraction of the measurements can corrupt the value of the statistic arbitrarily. Note that the finite-sample breakdown point is a function not of a distribution, but of the sample and the estimator. Typically, its value does not depend on the sample. It is this "breakdown point" that we saw was zero for the sample mean; it is a measure of *resistance*, not *robustness*.

Definition 10 *A sequence of estimators* $\{T_n\}$ *is* qualitatively robust at F *if for every* $\epsilon > 0$, $\exists \delta > 0$ *such that for all* $G \in \mathcal{M}$ *and all* n,

$$\pi(F, G) < \delta \Rightarrow \pi(\mathcal{L}_F(T_n), \mathcal{L}_G(T_n)) < \epsilon. \tag{23}$$

That is, $\{T_n\}$ *is qualitatively robust at* F *if the distributions of* T_n *are equicontinuous w.r.t.* n.

1.3 The Influence Function

The next few sections follow Hampel *et al.* (Ch2) fairly closely. Consider an estimator of a real parameter $\theta \in \Theta$, where Θ is an open subset of $\mathbf{R}$, based on a sample X_n of size n. Each observation takes values in $\mathcal{X}$. We consider a family $\mathcal{F}$ of distributions $\{F_\theta : \theta \in \Theta\}$, which are assumed to have densities $\{f_\theta : \theta \in \Theta\}$ with respect to a common dominating measure.

Consider estimators $\hat{\theta} = T_n(\hat{F}_n)$ that asymptotically can be replaced by functional estimators. (That is, either $T_n(\hat{F}_n) = T(\hat{F}_n)$ for all n, or there is a functional $T : \mathrm{dom}(T) \to \mathbf{R}$ such that if the components of X_n are iid G,

$$T_n(\hat{G}_n) \to T(G) \tag{24}$$

in probability as $n \to \infty$. $T(G)$ is the asymptotic value of $\{T_n\}$ at G. Suppose that for $G \in \mathrm{dom}(T)$, , where $T(F_\theta) = \theta \ \forall \theta \in \Theta$ (this is Fisher consistency of the estimator).

Definition 11 *A functional T defined on probability measures is Gâteaux differentiable at the measure F in* $\mathrm{dom}(T)$ *if there exists $a : \mathbf{R} \to \mathbf{R}$ s.t. $\forall G \in \mathrm{dom}(T)$,*

$$\lim_{t \to 0} \frac{T((1-t)F + tG) - T(F)}{t} = \int a(x) dG(x). \tag{25}$$

That is,

$$\frac{\partial}{\partial t} T((1-t)F + tG)|_{t=0} = \int a(x) dG(x). \tag{26}$$

Gâteaux differentiability is weaker than Fréchet differentiability. Essentially, Gâteaux differentiability at F ensures that the directional derivatives of T exist in all directions that (at least infinitesmally) leave one in the domain of T.

Let δ_x be a point mass at x.

Definition 12 *The* influence function *of T at F is*

$$\mathrm{IF}(x; T, F) \equiv \lim_{t \to 0} \frac{T((1-t)F + t\delta_x) - T(F)}{t} \tag{27}$$

at the points $x \in \mathcal{X}$ where the limit exists.

The influence function can exist even when the Gâteaux derivative does not, because the set of directions considered is less rich. The influence function gives the effect on T of an infinitesmal perturbation to the data at the point x. This leads to a "Taylor series" expansion of T at F:

$$T(G) = T(F) + \int \mathrm{IF}(x; T, F) d(G - F)(x) + \text{ remainder}. \tag{28}$$

(Note that $\int \mathrm{IF}(x; T, F) dF(x) = 0$.)

1.3.1 Heuristics using $\mathrm{IF}(x; T, F)$

Consider what happens for large sample sizes. The ecdf $\hat{F}_n$ tends to F, and $T_n(\hat{F}_n)$ tends to $T(\hat{F}_n)$. We have $\hat{F}_n = \frac{1}{n} \sum_{j=1}^n \delta_{X_j}$, so

$$(T_n(\hat{F}_n) - T(F)) \approx \frac{1}{n} \sum_{j=1}^n \mathrm{IF}(X_j; T, F) + \text{ remainder}, \tag{29}$$

or

$$\sqrt{n}(T_n(\hat{F}_n) - T(F)) \approx \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \mathrm{IF}(X_j; T, F) + \text{ remainder.} \tag{30}$$

Now $\{\mathrm{IF}(X_j; T, F)\}_{j=1}^n$ are n iid random variables with mean zero, so the sum is asymptotically normal. If the remainder vanishes asymptotically (not that easy to verify, typically, but often true), then $\sqrt{n}(T_n(\hat{F}_n) - T(F))$ is also asymptotically normal, with asymptotic variance

$$V(T, F) = \int \mathrm{IF}(x; T, F)^2 dF(x). \tag{31}$$

When the relationship holds, an asymptotic form of the Cramér-Rao bound relates asymptotic efficiency to the influence function. Suppose T is Fisher consistent. The Fisher information at F_{θ_0} is

$$I(F_{\theta_0}) = \int \left(\frac{\partial}{\partial \theta} \ln f_\theta(x) \bigg|_{\theta_0} \right)^2 dF_{\theta_0}. \tag{32}$$

Then T is asymptotically efficient only if

$$\mathrm{IF}(x; T, F) = I(F_{\theta_0})^{-1} \frac{\partial}{\partial \theta} (\ln f_\theta(x))|_{\theta_0}. \tag{33}$$

1.3.2 Relation to the Jackknife

There is an asymptotic connection between the jackknife estimate of variance and the influence function as well. Recall that for a functional statistic $T_n = T(\hat{F}_n)$, we define the jth pseudovalue by

$$T_{nj}^* = nT(\hat{F}_n) - (n-1)T(\hat{F}_{(j)}), \tag{34}$$

where $\hat{F}_{(j)}$ is the cdf of the data with the jth datum omitted. The jackknife estimate is

$$T_n^* = \frac{1}{n} \sum_{j=1}^{n} T_{nj}^*. \tag{35}$$

The (sample) variance of the pseudovalues is

$$V_n = \frac{1}{n-1} \sum_{j=1}^{n} (T_{nj}^* - T_n^*)^2. \tag{36}$$

The jackknife estimate of the variance of T_n is $\frac{1}{n} V_n$. Plugging into the definition 27 and taking $t = \frac{-1}{n-1}$ gives

$$\begin{aligned} \frac{n-1}{-1} \left(T\left(1 - \frac{-1}{n-1}\hat{F}_n + \frac{-1}{n-1}\delta_{x_j}\right) - T(\hat{F}_n) \right) &= (n-1)[T(\hat{F}_n) - T(\hat{F}_{(j)})] \\ &= T_{nj}^* - T(\hat{F}_n). \end{aligned} \tag{37}$$

The jackknife is thus an approximation to the influence function.

1.3.3 Robustness measures defined from $\mathrm{IF}(x;T,F)$

The *gross error sensitivity of* T *at* F is

$$\gamma^* = \gamma^*(T,F) = \sup_{\{x:\mathrm{IF}(x;T,F) \text{ exists}\}} |\mathrm{IF}(x;T,F)|. \tag{38}$$

This measures the maximum change to T a small perturbation to F at a point can induce, which is a bound on the asymptotic bias of T in a neighborhood of F. If $\gamma^*(T,F) < \infty$, T is *B-robust* at F (B is for bias). For Fisher-consistent estimators, there is typically a minimum possible value of $\gamma^*(T,F)$, leading to the notion of *most B-robust* estimators. There is typically a tradeoff between efficiency and B-robustness: for a given upper bound on γ^*, there is a most efficient estimator.

The gross error sensitivity measures what can happen when the difference between what is observed and F can be anywhere (perturbing part of an observation by an arbitrary amount). There is a different notion of robustness related to changing the observed values slightly. The (infinitesmal) effect of moving an observation from x to y is $\mathrm{IF}(y;T,F) - \mathrm{IF}(x;T,F)$. This can be standardized by the distance from y to x to give the *local shift sensitivity*

$$\lambda^* = \lambda^*(T,F) = \sup_{\{x \neq y:\mathrm{IF}(x;T,F) \text{ and } \mathrm{IF}(y;T,F) \text{ both exist}\}} \frac{|\mathrm{IF}(y;T,F) - \mathrm{IF}(x;T,F)|}{|y-x|}. \tag{39}$$

This is the Lipschitz constant of the influence function. (A real function f with domain $\mathrm{dom}(f)$ is *Lipschitz continuous at* x if there exists a constant $C > 0$ such that $|f(x) - f(y)| \leq C|x-y|$ for all $y \in \mathrm{dom}(f)$. A real function f is Lipschitz continuous if it is Lipschitz continuous at every $x \in \mathrm{dom}(f)$. The Lipschitz constant of a Lipschitz-continuous function f is the smallest C such that $|f(x) - f(y)| \leq C|x-y|$ for all $x, y \in \mathrm{dom}(f)$. These definitions extend to functions defined on metric spaces, and by taking exponent of $|x-y|$ to be other than unity.)

Another measure of robustness involving the influence function is the *rejection point*

$$\rho^* = \rho^*(T,F) = \inf\{r > 0 : \mathrm{IF}(x;T,F) = 0 \ \ \forall |x| > r\}. \tag{40}$$

This measures how large an observation must be before the estimator ignores it completely. If very large observations are deemed almost certainly to be gross errors, it is good for ρ^* to be finite.

1.3.4 Example

Suppose $X \sim N(\theta, 1)$; $\theta \in \Theta = \mathbf{R}$; $\theta_0 = 0$; $F = \Phi$; $T_n(X_n) = \bar{X} = \frac{1}{n}\sum_{j=1}^n X_j$;

$$T(G) \equiv \int x dG(x). \tag{41}$$

The functional T is Fisher-consistent. Calculating IF gives

$$\begin{aligned} \mathrm{IF}(x; T, F) &= \lim_{x \to 0} \frac{\int u d((1-t)\Phi + t\delta_x)(u) - \int u d\Phi(u)}{t} \\ &= \lim_{t \to 0} \frac{(1-t)\int u d\Phi(u) + t\int u d\delta_x(u) - \int u d\Phi(u)}{t} \\ &= \lim_{t \to 0} \frac{tx}{t} \\ &= x. \end{aligned} \tag{42}$$

The Fisher information of the standard normal with unknown mean is $I(\Phi) = 1$, so

$$\int \mathrm{IF}^2(x; T, \Phi) d\Phi = 1 = I^{-1}(\Phi), \tag{43}$$

and IF $\propto (\partial/\partial\theta)(\ln f_\theta)|_0$. The arithmetic mean is the MLE and is asymptotically efficient. The gross-error sensitivity is $\gamma^* = \infty$, the local shift sensitivity is $\lambda^* = 1$, and the rejection point is $\rho^* = \infty$.

1.4 M-estimators

We shall consider in more detail what is needed for an M-estimator to be robust. An M estimator is one of the form

$$T_n(X) = \arg\min_{U_n} \sum_{j=1}^n \rho(X_j, U_n), \tag{44}$$

where ρ is a function on $\mathcal{X} \times \Theta$. This is a generalization of maximum likelihood, where one would minimize the negative log-likelihood, which for independent data, would be of the

form just given, with $\rho(X_j, U_n) = -\log f(x;\theta)$. If ρ is differentiable in θ, with

$$\psi(x,\theta) = \frac{\partial}{\partial\theta}\rho(x,\theta), \tag{45}$$

then T_n is a stationary point, and so satisfies

$$\sum_{j=1}^{n} \psi(X_j, T_n) = 0. \tag{46}$$

An estimator that can be characterized as the solution of an equation of either form 44 or 46 is an *M-estimator*. In such cases, the estimator is identified with ρ or ψ.

Take the second form. Define T implicitly as the functional for which

$$\int \psi(x, T(G))dG = 0 \tag{47}$$

for all G for which the integral exists; then the estimator is $T(\hat{F}_n)$. Let's calculate the IF at F of an estimator of this form. Define $F_{t,x} \equiv (1-t)F + t\delta_x$. We want the limit as $t \to 0$ of

$$T' \equiv \lim_{t\to 0} \frac{T(F_{t,x}) - T(F)}{t}. \tag{48}$$

Assuming we can interchange integration and differentiation, we can differentiate **??** to get

$$\int \psi(x, T(F))d(\delta_x - F) + \int \frac{\partial}{\partial\theta}\psi(x,\theta)\Big|_{T(F)} dF \cdot \frac{\partial}{\partial t}T(F_{t,x})\Big|_{t=0} = 0. \tag{49}$$

This yields

$$\mathrm{IF}(x;\psi,F) = \frac{\psi(x,T(F))}{-\int \frac{\partial}{\partial\theta}\psi(y,\theta)\Big|_{T(F)} dF(y)}, \tag{50}$$

assuming the denominator is not zero. The estimator given by ψ is B-robust iff $\psi(\cdot, T(F))$ is bounded.

In the special case of a location estimate, the dependence of ψ on x and θ is $\psi(x;\theta) = \psi(x-\theta)$, which yields

$$\mathrm{IF}(x;F,T) = \frac{\psi(x - T(F))}{\int \psi'(x - T(F))dF}. \tag{51}$$

The influence function is proportional to ψ.

Therefore, M-estimates have finite gross error sensitivity only if ψ is bounded, finite rejection point only if ψ *redescends* to zero for large values of its argument. For the mean, this does not occur.

1.4.1 Robustness of M-estimates

See Huber, Ch3 for more details; this is drawn from there.

Let's calculate $b_1(\epsilon)$ for the Lévy metric for an M-estimate of location, with $\psi(x;t) = \psi(x-t)$, with ψ monotonically increasing. We will use λ to denote something new in this section, so let $d_L(F,G)$ denote the Lévy distance between distributions (or cdfs) F and G. Accordingly, we take $\mathcal{P}(\epsilon) \equiv \{G : d_L(F,G) \leq \epsilon\}$. Assume that $T(F) = 0$. Define

$$b_+(\epsilon) = \sup_{G \in \mathcal{P}(\epsilon)} T(G) \tag{52}$$

and

$$b_-(\epsilon) = \inf_{G \in \mathcal{P}(\epsilon)} T(G). \tag{53}$$

Then $b_1(\epsilon) = \max\{-b_-(\epsilon), b_+(\epsilon)\}$ (because $T(F) = 0$). Define

$$\lambda(t;G) = E_G \psi(X-t) = \int \psi(x-t) dG(x). \tag{54}$$

Because ψ is monotonic, λ is decreasing in t, but not necessarily strictly decreasing, so the solution of $\lambda(t;G) = 0$ is not necessarily unique. Define

$$T^*(G) = \sup\{t : \lambda(t;G) > 0\} \tag{55}$$

and

$$T^{**}(G) = \inf\{t : \lambda(T;G) < 0\}. \tag{56}$$

Then $T^*(G) \leq T(G) \leq T^{**}(G)$. Note that $\lambda(t;G)$ increases if G is made stochastically larger. The stochastically largest element of $\{G : d_L(F,G) \leq \epsilon\}$ is

$$F_1(x) = (F(x-\epsilon) - \epsilon)_+ = \begin{cases} 0, & x \leq x_0 + \epsilon \\ F(x-\epsilon) - \epsilon & x > x_0 + \epsilon, \end{cases} \tag{57}$$

where x_0 solves $F(x_0) = \epsilon$. (Assume that x_0 exists; the discontinuous case introduces some additional bookkeeping.) Note that this distribution puts mass ϵ at $x = \infty$. For $G \in \mathcal{P}(\epsilon)$,

$$\lambda(t;G) \leq \lambda(t;F_1) = \int_{x_0}^{\infty} \psi(x-t+\epsilon) dF(x) + \epsilon\psi(\infty). \tag{58}$$

It follows that

$$\begin{aligned} b_+(\epsilon) &= \sup_{G\in\mathcal{P}(\epsilon)} T(G) \\ &= T^{**}(F_1) \\ &= \inf\{t : \lambda(t;F_1) < 0\}. \end{aligned} \tag{59}$$

Note that

$$\ell_+ = \lim_{t\to\infty} \lambda(t;F_1) = (1-\epsilon)\psi(-\infty) + \epsilon\psi(\infty). \tag{60}$$

Provided $\ell_+ < 0$ and $\psi(\infty) < \infty$, $b_+(\epsilon) < b_+(1) = \infty$. Thus to avoid breakdown from above we need

$$\frac{\epsilon}{1-\epsilon} < -\frac{\psi(-\infty)}{\psi(\infty)}. \tag{61}$$

We can calculate $b_-(\epsilon)$ in the same way: the stochastically smallest element of $\mathcal{P}_(\epsilon)$ is

$$F_{-1} = (F(x+\epsilon)+\epsilon)\wedge 1 = \begin{cases} f(x+\epsilon)+\epsilon, & x \le x_1 - \epsilon \\ 1, & x > x_1 - \epsilon, \end{cases} \tag{62}$$

where x_1 solves $F(x_1) = 1-\epsilon$. This distribution assigns mass ϵ to $x = -\infty$. We have

$$\lambda(t;G) \ge \lambda(t;F_{-1}) = \epsilon\psi(-\infty) + \int_{-\infty}^{x_1} \psi(x-t-\epsilon)dF(x). \tag{63}$$

Thus

$$\begin{aligned} b_-(\epsilon) &= \inf_{G\in\mathcal{P}(\epsilon)} T(G) \\ &= T^*(F_{-1}) \\ &= \sup\{t : \lambda(t;F_{-1}) > 0\}. \end{aligned} \tag{64}$$

Note that

$$\ell_- = \lim_{t\to-\infty} \lambda(t;F_{-1}) = \epsilon\psi(-\infty) + (1-\epsilon)\psi(\infty). \tag{65}$$

To avoid breakdown from below, we need $\psi(-\infty) > -\infty$ and $\ell_- > 0$, which leads to

$$\frac{\epsilon}{1-\epsilon} > -\frac{\psi(\infty)}{\psi(-\infty)}. \tag{66}$$

Combining this with 61 gives

$$-\frac{\psi(\infty)}{\psi(-\infty)} < \frac{\epsilon}{1-\epsilon} < -\frac{\psi(-\infty)}{\psi(\infty)}. \tag{67}$$

Define

$$\eta \equiv \min\left\{-\frac{\psi(-\infty)}{\psi(\infty)}, -\frac{\psi(\infty)}{\psi(-\infty)}\right\}. \tag{68}$$

The breakdown point is then

$$\epsilon^* = \frac{\eta}{1+\eta}. \tag{69}$$

The maximum possible value $\epsilon^* = 1/2$ is attained if $\psi(\infty) = -\psi(-\infty)$. The breakdown point is $\epsilon^* = 0$ if ψ is unbounded.

This calculation also shows that if ψ is bounded and $\lambda(t; F)$ has a unique zero at $t = T(F)$, then T is continuous at F; otherwise, T is not continuous at F.

For non-monotone functions ψ, things are much more complicated. This is the case for "redescending influence functions," to which we shall turn presently.

1.4.2 Minimax Properties for location estimates

It is straightforward to find the minimax bias location estimate for symmetric unimodal distributions; the solution is the sample median (see Huber, §4.2). Minimizing the maximum variance is somewhat more difficult. Define

$$v_1(\epsilon) = \sup_{G \in \mathcal{P}(\epsilon)} A(G, T), \tag{70}$$

where $A(G, T)$ is the asymptotic variance of T at G. Assume the observations are iid $G(\cdot - \theta)$. The shape varies over the family $\mathcal{P}(\epsilon)$; the parameter varies over the reals. Such families are not typically compact in the weak topology. Huber uses the *vague* topology to surmount this problem. The *vague* topology is the weakest topology on the set $\mathcal{M}_+$ of sub-probability measures for which $F \to \int \psi dF$ is continuous for all continuous functions ψ with compact support. (A subprobability measure can have total mass less than one, but is otherwise the same as a probability measure.) Because $\mathbf{R}$ is locally compact, $\mathcal{M}_+$ is compact.

Define F_0 to be the distribution in $\mathcal{P}(\epsilon)$ with smallest Fisher information

$$I(G) = \sup_{\psi \in \mathcal{C}_K^1} \frac{(\int \psi' dG)^2}{\int \psi^2 dG}, \tag{71}$$

where $\mathcal{C}_K^1$ is the set of all compactly supported, continuously differentiable functions ψ s.t. $\int \psi^2 dF > 0$. This extends the definition of the Fisher information beyond measures that

have densities; in fact, $I(F) < \infty$ iff F has an absolutely continuous density w.r.t. Lebesgue measure, and $\int (f'/f)^2 f dx < \infty$.

Proof. (Following Huber, pp78ff.) By assumption, $\int \psi^2 dx < \infty$. If $\int (f'/f)^2 f dx < \infty$,

$$\begin{aligned} \left(\int \psi' f dx \right)^2 &= \left(\psi f |_{-\infty}^{\infty} - \int \psi \frac{f'}{f} f dx \right)^2 \\ &= \left(\int \psi \frac{f'}{f} f dx \right)^2 \\ &\leq \left(\int \psi^2 f dx \right) \left(\int \left(\frac{f'}{f} \right)^2 f dx \right), \end{aligned} \tag{72}$$

by the weighted Cauchy-Schwarz inequality. Thus $I(F) \leq \int (f'/f)^2 f dx < \infty$. Now suppose $I(F) < \infty$. Then $L(\psi) = -\int \psi' dF$ is a bounded linear functional on the (dense) subset $\mathcal{C}_K^1$ of $L_2(F)$, the Hilbert space of square-integrable functions w.r.t. F. By continuity, L can be extended to a continuous linear functional on all of $L_2(F)$. By the Riesz Representation Theorem, there then exists a function $g \in L_2(F)$ such that for all $\psi \in L_2(F)$,

$$L\psi = \int \psi g dF. \tag{73}$$

Clearly, $L1 = \int g dF = 0$. Define

$$f(x) \equiv \int_{y<x} g(y) F(dy) = \int 1_{y<x} g(y) F(dy). \tag{74}$$

By the Cauchy-Schwarz inequality,

$$|f(x)|^2 \leq (\int 1_{y<x}^2 F(dy))(\int g(y)^2 F(dy)) = F((-\infty, x)) \int g^2 F(dy), \tag{75}$$

which tends to zero as $x \to -\infty$; $|f(x)|$ also tends to zero as $x \to \infty$. For $\psi \in \mathcal{C}_K^1$,

$$-\int \psi'(x) f(x) dx = -\int_{y<x} \int \psi'(x) g(y) F(dy) dx = \int \psi(y) g(y) F(dy) = L\psi \tag{76}$$

by Fubini's theorem. Thus the measure $f(x)dx$ and the measure $F(dx)$ give the same linear functional on derivatives of functions in $\mathcal{C}_K^1$. This set is dense in $L_2(F)$, so they define the same measure, and so f is a density of F. We can now integrate the definition of the functional L by parts to show that

$$L\psi = -\int \psi' f dx = \int \psi \frac{f'}{f} f dx. \tag{77}$$

Thus

$$I(F) = \|L\|^2 = \int g^2 dF = \int \left(\frac{f'}{f}\right)^2 f dx. \tag{78}$$

The functional $I(G)$ is lower-semicontinuous with respect to the vague topology, so $I(G)$ attains its infimum on any vaguely compact set. Furthermore, $I(G)$ is a convex function of G.

Theorem 3 *(Huber, Proposition 4.5) Let $\mathcal{P}$ be a set of measures on $\mathbf{R}$. Suppose $\mathcal{P}$ is convex, $F_0 \in \mathcal{P}$ minimizes $I(G)$ over $\mathcal{P}$, $0 < I(F_0) < \infty$, and the set where the density f_0 of F_0 is strictly positive is (a) convex and (b) contains the support of every distribution in $\mathcal{P}$.*

Then F_0 is the unique minimizer of $I(G)$ over $\mathcal{P}$.

The reciprocal of $I(F_0)$ lower-bounds the (worst) asymptotic variance of any estimator over all $G \in \mathcal{P}$, so if one can find an estimator whose asymptotic variance is $1/I(F_0)$, it is minimax (for asymptotic variance).

Finding F_0 can be cast as a variational problem; see Huber, §4.5.

The least-informative distributions in neighborhoods of the normal tend to have thinner tails than the normal. If one believes that outliers might be a problem, it makes sense to abandon minimaxity in favor of estimators that do somewhat better when the truth has thicker tails than the normal. That leads to considering *redescending influence functions*, for which $\psi = 0$ for x sufficiently large.

One can develop "minimax" estimators in this restricted class. For example, we could seek to minimize the asymptotic variance subject to $\psi(x) = 0$, $|x| > c$. For the ϵ-contamination neighborhood of a normal, the minimax ψ in this class is

$$\psi(x) = -\psi(-x) = \begin{cases} x, & 0 \le x \le a \\ b \tanh\left(\frac{b(c-x)}{2}\right) & a \le x \le c \\ 0, & x \ge c. \end{cases} \tag{79}$$

The values of a and b depend on ϵ.

Other popular redescending influence functions include Hampel's piecewise linear influ-

ence functions:

$$\psi(x) = -\psi(-x) = \begin{cases} x, & 0 \le x \le a \\ a & a \le x \le b \\ a\frac{c-x}{c-b} & b \le x \le c \\ 0, & x \ge c, \end{cases} \tag{80}$$

and Tukey's "biweight"

$$\psi(x) = \begin{cases} x(1-x^2)^2, & |x| \le 1 \\ 0, & |x| > 1. \end{cases} \tag{81}$$

A complication in using redescending influence functions is that scaling (some form of Studentizing) is much more important for them to be efficient than it is for monotone influence functions. The slope of the influence function in the descending regions can also inflate the asymptotic variance (recall that $(\int \psi' dF)^2$ is in the denominator of $A(F,T)$).

1.5 Estimates of Scale

We require that a scale estimate S_n be equivariant under changes of scale, so that

$$S_n(aX) = aS_n(X) \ \ \forall a > 0. \tag{82}$$

It is common also to require that a scale estimate be invariant under sign changes and translations, so that

$$S_n(-X) = S_n(X) = S_n(X + b\mathbf{1}), \tag{83}$$

where $b \in \mathbf{R}$ and $\mathbf{1}$ is an n-vector of ones. The most common need for a scale estimate is to remove scale as a nuisance parameter in a location estimate, by Studentizing.

It turns out that the bias properties of a scale estimate are more important for studentizing than the variance properties. That leads to considering something involving the median deviation. The single most widely used robust scale estimator is the median absolute deviation (MAD). Let $M_n(x)$ be the median of the list of the elements of x: $\text{med}\{x_j\}_{j=1}^n$. Then

$$\text{MAD}_n(x) = \text{med}\{|x_j - M_n(x)|\}_{j=1}^n. \tag{84}$$

The breakdown point of the MAD is $\epsilon^* = 1/2$. Typically, the MAD is multiplied by a 1.4826 $(1/\Phi^{-1}(3/4))$ to make its expected value unity for a standard normal.

1.6 Robust Regression