

A SHORT COURSE ON ROBUST STATISTICS

David E. Tyler

Rutgers,
The State University of New Jersey

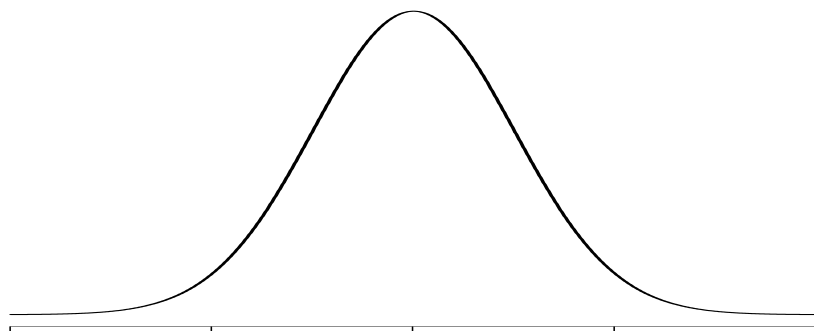
PART 1

CONCEPTS AND BASIC METHODS

MOTIVATION

- Univariate Data Set: X_1, X_2, \dots, X_n
- Parametric Model: $F(x_1, \dots, x_n \mid \theta)$
 - θ : Unknown parameter
 - F : Known function
- e.g. X_1, X_2, \dots, X_n i.i.d. $Normal(\mu, \sigma^2)$
- **Q:** Is it realistic to believe we don't know (μ, σ^2) , but we know e.g. the shape of the tails of the distribution?
- **A:** The model is assumed to be approximately true, e.g. symmetric and unimodal (past experience).
- **Q:** Are statistical methods which are good under the model reasonably good if the model is only approximately true? (Presumption underlying goodness-of-fit tests.)
- **ROBUST STATISTICS:** Formally addresses this issue.

CLASSIC EXAMPLE: MEAN .vs. MEDIAN

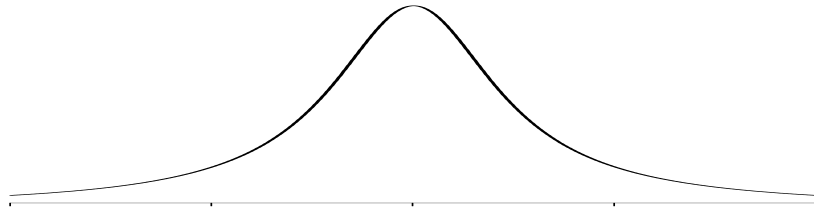


- Symmetric distributions: $\mu = \begin{cases} \text{population mean} \\ \text{population median} \end{cases}$
- Sample mean: $\bar{X} \approx \text{Normal} \left(\mu, \frac{\sigma^2}{n} \right)$
- Sample median: $\text{Median} \approx \text{Normal} \left(\mu, \frac{1}{n} \frac{1}{4f(\mu)^2} \right)$
- At normal: $\text{Median} \approx \text{Normal} \left(\mu, \frac{\sigma^2}{n} \frac{\pi}{2} \right)$
- Asymptotic Relative Efficiency of Median to Mean

$$ARE(\text{Median}, \bar{X}) = \frac{\text{avar}(\bar{X})}{\text{avar}(\text{Median})} = \frac{2}{\pi} = 0.6366$$

CAUCHY DISTRIBUTION

$$X \sim Cauchy(\mu, \sigma^2)$$



$$f(x; \mu, \sigma) = \frac{1}{\pi\sigma} \left(1 + \left(\frac{x - \mu}{\sigma} \right)^2 \right)^{-1/2}$$

- Mean: $\bar{X} \sim Cauchy(\mu, \sigma^2)$
- Median $\approx Normal\left(\mu, \frac{\pi^2\sigma^2}{4n}\right)$
- $ARE(Median, \bar{X}) = \infty$ or $ARE(\bar{X}, Median) = 0$
- For t on ν degrees of freedom:

$$ARE(Median, \bar{X}) = \frac{4}{(\nu - 2)\pi} \frac{\Gamma((\nu + 1)/2)^2}{\Gamma(\nu/2)}$$

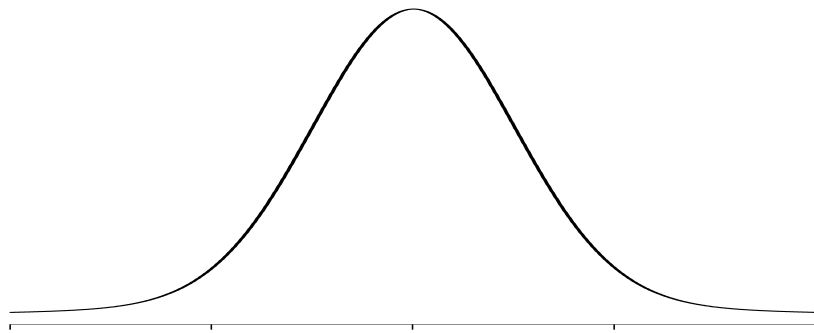
	$\nu \leq 2$	3	4	5
$ARE(Median, \bar{X})$	∞	1.621	1.125	0.960
$ARE(\bar{X}, Median)$	0	0.617	0.888	1.041

MIXTURE OF NORMALS

- *Theory of errors*: Central Limit Theorem gives plausibility to normal distributions.

- $$X \sim \begin{cases} \text{Normal}(\mu, \sigma^2) & \text{with probability } 1 - \epsilon \\ \text{Normal}(\mu, (3\sigma)^2) & \text{with probability } \epsilon \end{cases}$$

i.e. not all measurements are equally precise.



$$X \sim (1 - \epsilon) \text{Normal}(\mu, \sigma^2) + \epsilon \text{Normal}(\mu, (3\sigma)^2) \\ \text{for } \epsilon = 0.10$$

- Classic paper: Tukey (1960)
 - For $\epsilon > 0.10 \Rightarrow ARE(\text{Median}, \bar{X}) > 1$
 - The mean absolute deviation is more efficient than the sample standard deviation for $\epsilon > 0.01$.

PRINCETON ROBUSTNESS STUDIES

Andrews, et.al. (1972)

Other estimates of location.

- α -trimmed mean: Trim a proportion of α from both ends of the data set and then take the mean. (**Throwing away data?**)
- α -Windsorized mean: Replace a proportion of α from both ends of the data set by the next closest observation and then take the mean.
- Example: 2, 4, 5, 10, 200
Mean = 44.2 Median = 5
20% trimmed mean = $(4 + 5 + 10)/3 = 6.33$
20% Windsorized mean = $(4 + 4 + 5 + 10 + 10)/5 = 6.6$
- There exist estimates of location which are asymptotically most efficient for the center of any symmetric distribution. (Adaptive estimation, semi-parametrics). **Robust?**

Measuring the robustness of a statistics

- Relative Efficiency over a range of distributional models.
- Influence Function over a range of distributional models.
- Maximum Bias Function and the Breakdown Point.

Measuring the effect of an outlier (*not modeled*)

- *Good Data Set*: x_1, \dots, x_{n-1}
 - Statistic: $T_{n-1} = T(x_1, \dots, x_n)$
- *Contaminated Data Set*: $x_1, \dots, x_{n-1}, \mathbf{x}$
 - Contaminated Value: $T_n = T(x_1, \dots, x_n, \mathbf{x})$

THE SENSITIVITY CURVE

Tukey (1970)

$$SC_n(\mathbf{x}) = n(T_n - T_{n-1})$$

or

$$T_n = T_{n-1} + \frac{1}{n}SC_n(\mathbf{x})$$

THE INFLUENCE FUNCTION

Hampel (1969, 1974)

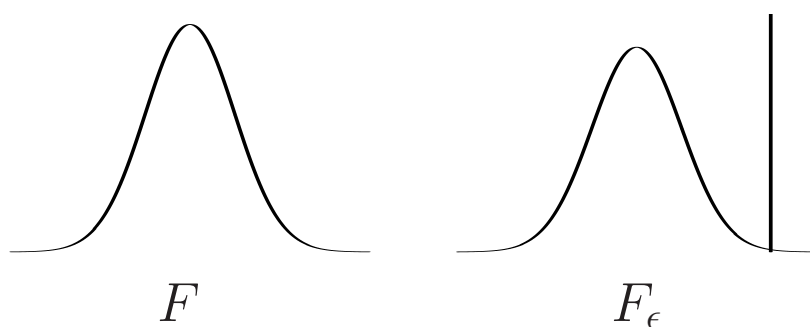
Population version of the sensitivity curve.

THE INFLUENCE FUNCTION

- Consider F and the ϵ -contaminated distribution

$$F_\epsilon = (1 - \epsilon)F + \epsilon\delta_{\mathbf{x}}$$

where $\delta_{\mathbf{x}}$ is the point mass distribution at \mathbf{x} .



- Compare Functional Values: $T(F)$.vs. $T(F_\epsilon)$
- Assume Qualitative Robustness (*Continuity*):

$$T(F_\epsilon) \rightarrow T(F) \text{ as } \epsilon \rightarrow 0$$

e.g. the mode is not qualitatively robust

- Influence Function (*Infinitesimal perturbation*)

$$IF(\mathbf{x}; T, F) = \lim_{\epsilon \rightarrow 0} \frac{T(F_\epsilon) - T(F)}{\epsilon}$$

EXAMPLES

- **Mean:** $T(F) = E_F[X]$.

$$\begin{aligned}T(F_\epsilon) &= E_{F_\epsilon}(X) \\ &= (1 - \epsilon)E_F[X] + \epsilon E[\delta_{\mathbf{x}}] \\ &= (1 - \epsilon)T[F] + \epsilon \mathbf{x}\end{aligned}$$

$$\begin{aligned}IF(\mathbf{x}; T, F) &= \lim_{\epsilon \rightarrow 0} \frac{(1-\epsilon)T(F) + \epsilon \mathbf{x} - T(F)}{\epsilon} \\ &= \mathbf{x} - T(F)\end{aligned}$$

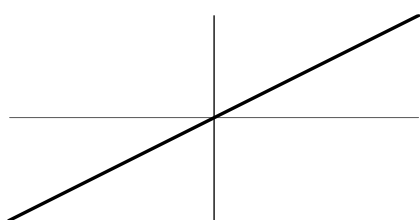
- **Median:** $T(F) = F^{-1}(1/2)$

$$IF(\mathbf{x}; T, F) = \{2 f(T(F))\}^{-1} \text{sign}(X - T(F))$$

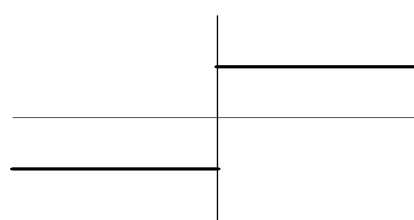
Plots of Influence Functions

Gives insight into the behavior of a statistic.

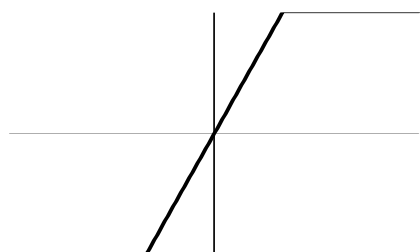
$$T_n \approx T_{n-1} + \frac{1}{n}IF(\mathbf{x}; T, F)$$



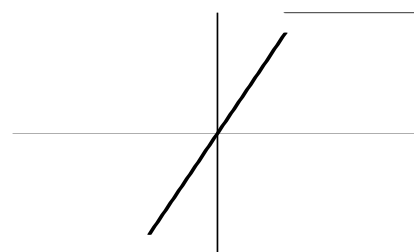
Mean



Median



α -trimmed mean



α -Winsorized mean
(*somewhat unexpected?*)

Desirable robustness properties of the influence function

- **SMALL**

- Gross Error Sensitivity

$$GES(T; F) = \sup_{\mathbf{x}} | IF(\mathbf{x}; T, F) |$$

$$GES < \infty \Rightarrow \text{B-robust (Bias-robust)}$$

- Asymptotic Variance

$$AV(T; F) = E_F[IF(X; T, F)^2]$$

- In general,

$$\sqrt{n}(T(X_1, \dots, X_n) - T(F)) \rightarrow Normal(0, AV(T; F))$$

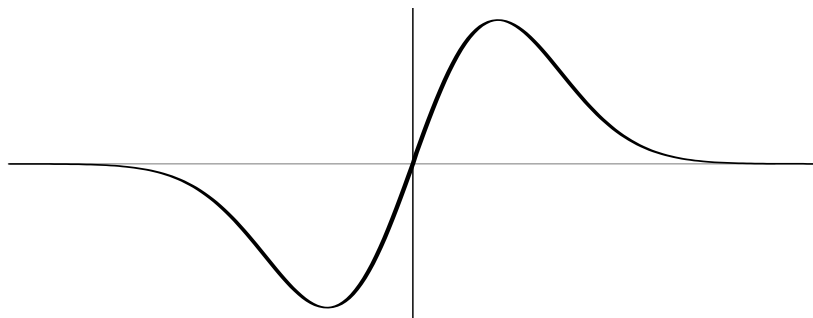
- Typically there is a trade-off at the normal model

- i.e. Smaller AV \leftrightarrow Larger GES

- **SMOOTH** (local shift sensitivity): protects e.g. against rounding error.

- **REDESCENDING** to 0.

REDESCENDING INFLUENCE FUNCTION



- Example: Data Set of Male Heights in cm
180, 175, 192, . . . , 185, 2020, 190, . . .
- Redescender = Automatic Outlier Detector

M-ESTIMATES *Huber (1964, 1967)*

Maximum likelihood type estimates
under non-standard conditions

- **One-Parameter Case.**

$$X_1, \dots, X_n \text{ i.i.d. } f(x; \theta) \quad \theta \in \Theta$$

- **Maximum likelihood estimates**

- Likelihood function.

$$L(\theta \mid x_1, \dots, x_n) = \prod_{i=1}^n f(x_i; \theta)$$

- Minimize the negative log-likelihood.

$$\min_{\theta} \sum_{i=1}^n \rho(x_i; \theta) \text{ where } \rho(x_i; \theta) = -\log f(x_i; \theta).$$

- Solve the likelihood equations

$$\sum_{i=1}^n \psi(x_i; \theta) = 0 \text{ where } \psi(x_i; \theta) = \frac{\partial \rho(x_i; \theta)}{\partial \theta}$$

DEFINITIONS OF M-ESTIMATES

- Objective function approach:

$$\hat{\theta} = \arg \min \sum_{i=1}^n \rho(x_i; \theta)$$

- M-estimating equation approach: $\hat{\theta}$ is a solution to

$$\sum_{i=1}^n \psi(x_i; \theta) = 0$$

- Basic examples.

- Mean: MLE for normal

$$\rho(x; \theta) = (x - \theta)^2 \quad \text{or} \quad \psi(x; \theta) = x - \theta$$

- Median: MLE for Double Exponential

$$\rho(x; \theta) = |x - \theta| \quad \text{or} \quad \psi(x; \theta) = \text{sign}(x - \theta)$$

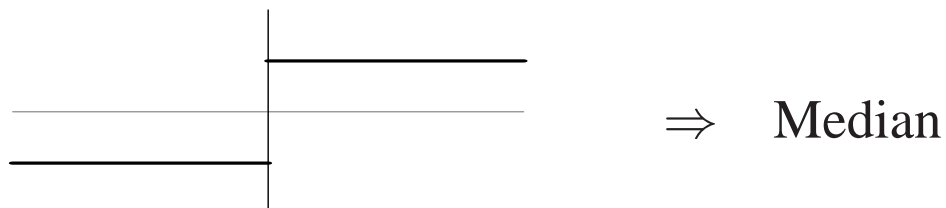
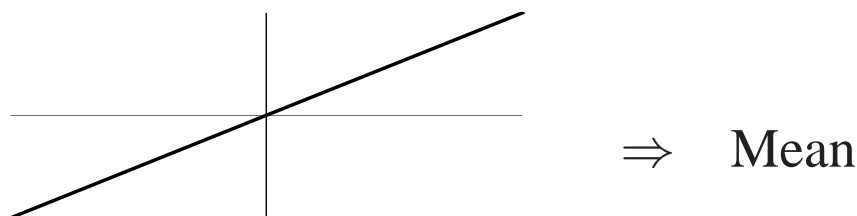
- ρ and ψ need not be related to any density or to each other.

INFLUENCE FUNCTION FOR M-ESTIMATES

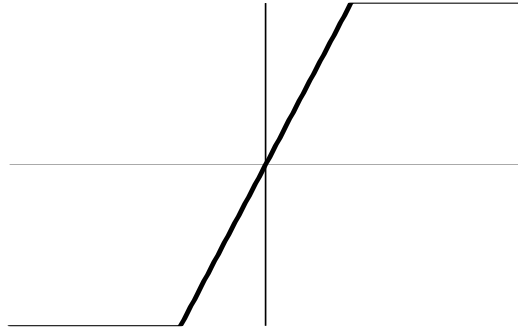
$$IF(\mathbf{x}; T, F) = c(T, F)\psi(\mathbf{x}; T(F))$$

where $c(t, f) = -1/E_F \left[\frac{\partial \psi(X; \theta)}{\partial \theta} \right]$ evaluated at $\theta = T(F)$.

- One can decide what shape is desired for the Influence Function and then construct an appropriate M-estimate.



EXAMPLE



- Choose

$$\psi(r) = \begin{cases} c & r \geq c \\ r & |r| < c \\ -c & r \leq -c \end{cases}$$

where c is a tuning constant.

- **Huber's M-estimate** = Adaptively trimmed mean.
i.e. *the proportion trimmed depends upon the data.*

M-estimates of location

Adaptively weighted means

- Translation equivariance.

$$X_i \rightarrow X_i + a \Rightarrow T_n \rightarrow T_n + a$$

Implies: $\rho(x; t) = \rho(x - t)$ and $\psi(x; t) = \psi(x - t)$

- Location M-estimates: translation equivariant and symmetric, i.e

$$\rho(-r) = \rho(r) \quad \text{or} \quad \psi(-r) = -\psi(r).$$

- Representation as adaptively weighted means.

Express $\psi(r) = ru(r)$ and let $w_i = u(x_i - \theta)$, then

$$0 = \sum_{i=1}^n \psi(x_i - \theta) = \sum_{i=1}^n (x_i - \theta)u(x_i - \theta) = \sum_{i=1}^n w_i \{x_i - \theta\}$$

$$\Rightarrow \hat{\theta} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

ADAPTIVELY WEIGHTED MEANS

The weights are determined by the data cloud.



$\hat{\theta}$



Heavily Down – weighted

COMPUTATIONS

- **IRLS.** *Iterative Re-weighted Least Squares Algorithm.*

$$\theta_{k+1} = \frac{\sum_{i=1}^n w_{i,k} x_i}{\sum_{i=1}^n w_{i,k}}$$

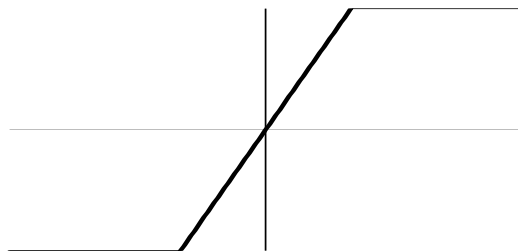
where

$$w_{i,k} = u(x_i - \theta_k)$$

SOME COMMON M-ESTIMATES OF LOCATION

Huber's M-estimate

$$\psi(r) = \begin{cases} c & r \geq c \\ r & |r| < c \\ -c & r \leq -c \end{cases}$$

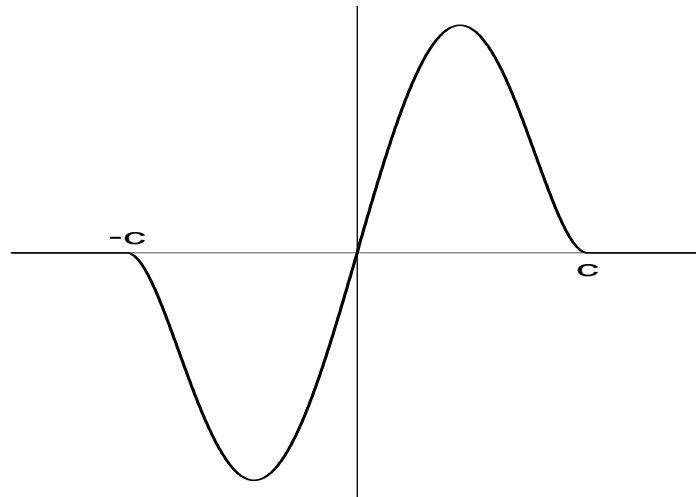


- For a given bound on the GES, it has maximum efficiency at the normal model
- MLE for the “least favorable distribution”, i.e. symmetric unimodal model with smallest Fisher Information within a “neighborhood” of the normal distribution.

Tukey's Bi-weight M-estimate (or bi-square)

$$u(r) = \left\{ \left(1 - \frac{r^2}{c^2} \right)_+ \right\}^2$$

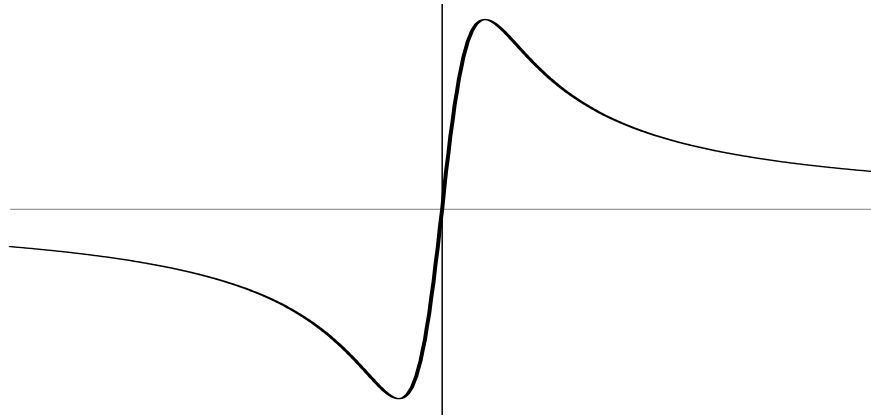
where $a_+ = \max\{0, a\}$.



- Linear near zero
- Smooth (continuous second derivatives)
- Strongly redescending to 0
- *NOT AN MLE.*

Cauchy MLE

$$\psi(r) = \frac{r/c}{(1 + r^2/c^2)^{1/2}}$$



Not a hard redescender.

PART 2

**MORE ADVANCED CONCEPTS AND
METHODS**

SCALE EQUIVARIANCE

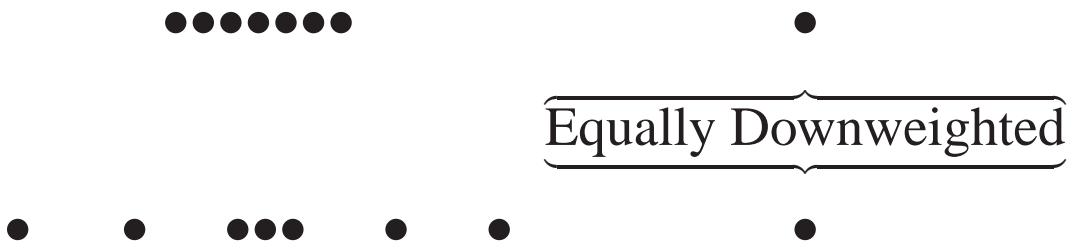
- M-estimates of location alone are not scale equivariant in general, i.e.

$$X_i \rightarrow X_i + a \Rightarrow \hat{\theta} \rightarrow \hat{\theta} + a \text{ location equivariant}$$

$$\text{but } X_i \rightarrow b X_i \not\Rightarrow \hat{\theta} \rightarrow b \hat{\theta} \text{ scale equivariant}$$

(Exceptions: the mean and median.)

- Thus, the adaptive weights are not dependent on the spread of the data



SCALE STATISTICS: s_n

$$X_i \rightarrow bX_i + a \Rightarrow s_n \rightarrow |b|s_n$$

- Sample standard deviation.

$$s_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- MAD (or more appropriately MADAM).

Median Absolute Deviation About the Median

$$s_n^* = \text{Median} | x_i - \text{median} |$$

$$s_n = 1.4826 s_n^* \rightarrow_p \sigma \text{ at } Normal(\mu, \sigma^2)$$

- **Example:** 2, 4, 5, 10, 12, 14, 200

– Median = 10

– Absolute Deviations: 8, 6, 5, 0, 2, 4, 190

– *MADAM* = 5 $\Rightarrow s_n = 7.413$

– (Standard deviation = 72.7661)

M-estimates of location with auxiliary scale

$$\hat{\theta} = \arg \min \sum_{i=1}^n \rho \left(\frac{x_i - \theta}{c s_n} \right)$$

or

$$\hat{\theta} \text{ solves } \sum_{i=1}^n \psi \left(\frac{x_i - \theta}{c s_n} \right) = 0$$

- c : tuning constant
- s_n : consistent for σ at normal

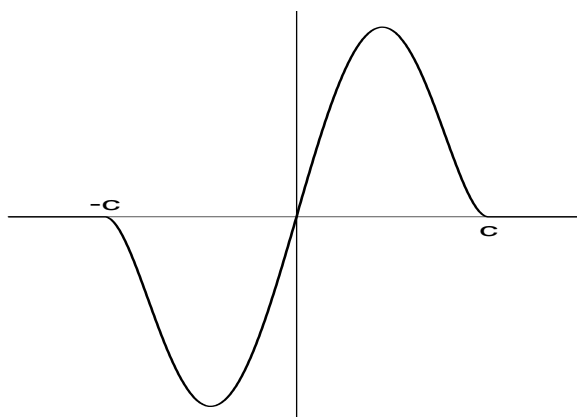
Tuning an M-estimate

- Given a ψ -function, define a class of M-estimates via

$$\psi_c(r) = \psi(r/c)$$

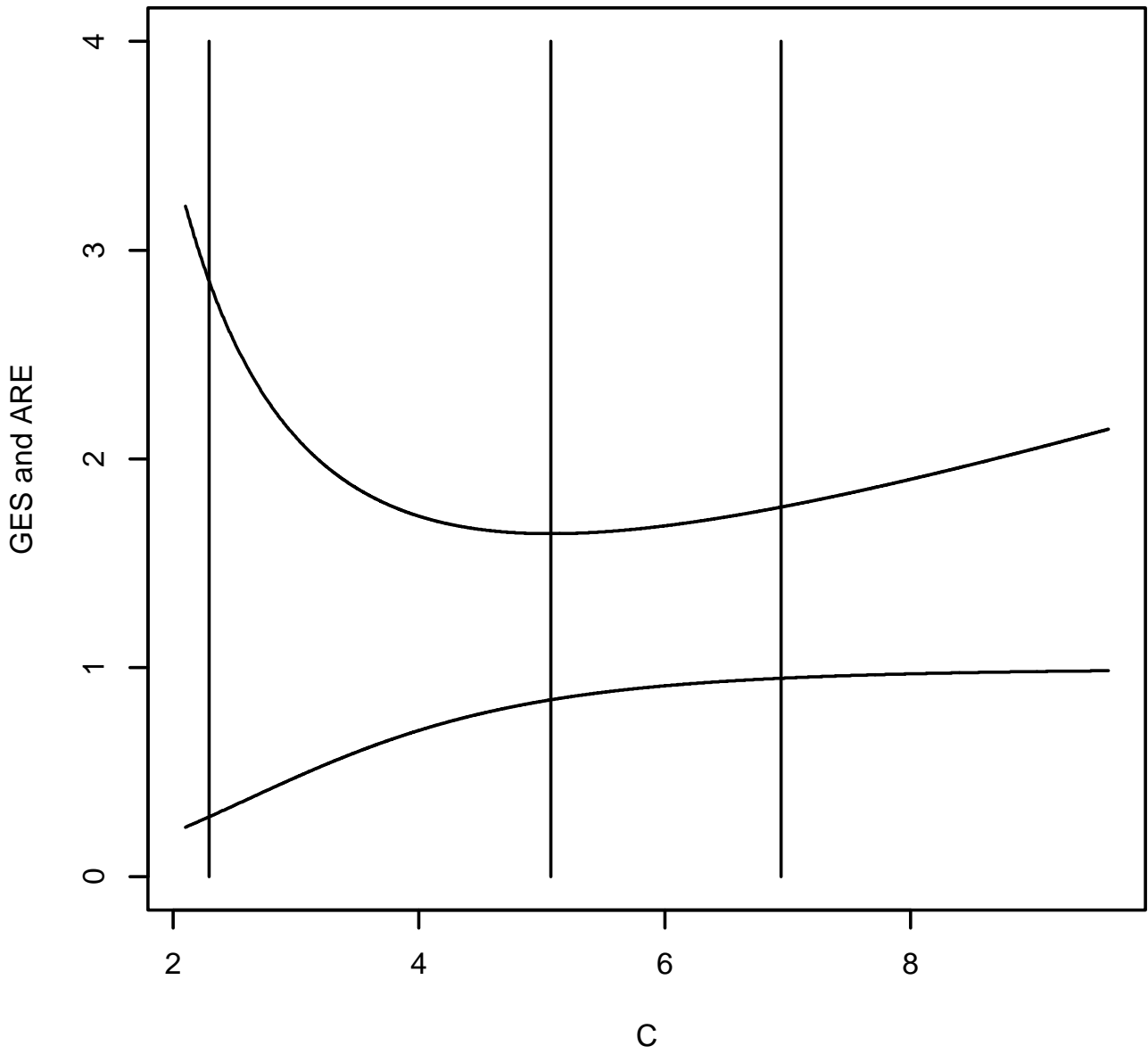
- Tukey's Bi-weight.

$$\psi(r) = r\{(1 - r^2)_+\}^2 \Rightarrow \psi_c(r) = \frac{r}{c} \left\{ \left(1 - \frac{r^2}{c^2}\right)_+ \right\}^2$$



- $c \rightarrow \infty \Rightarrow \approx$ Mean
- $c \rightarrow 0 \Rightarrow$ Locally very unstable

GES and ARE for Bi-Weight M-estimates of Location



THE BREAKDOWN POINT

- **Q:** What happens if we have more than one outlier?
 - GES: measure of local robustness.
 - Breakdown Point: measure of global robustness.
- $\mathcal{X}_n = \{X_1, \dots, X_n\}$: n “good” data points
- $\mathcal{Y}_m = \{Y_1, \dots, Y_m\}$: m “bad” data points
- $\mathcal{Z}_{n+m} = \mathcal{X}_n \cup \mathcal{Y}_m$: ϵ_m -contaminated sample. $\epsilon_m = \frac{m}{n+m}$
- Bias of a statistic: $|T(\mathcal{X}_n \cup \mathcal{Y}_m) - T(\mathcal{X}_n)|$
- Max-Bias under ϵ_m -contamination:

$$B(\epsilon_m; T, \mathbf{X}_n) = \sup_{\mathcal{Y}_m} |T(\mathcal{X}_n \cup \mathcal{Y}_m) - T(\mathcal{X}_n)|$$

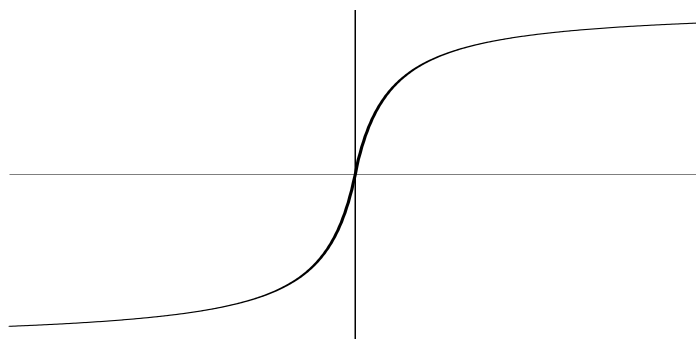
- Finite sample contamination breakdown point:
Donoho and Huber (1983)

$$\epsilon_c^*(T; \mathbf{X}_n) = \inf\{\epsilon_m \mid B(\epsilon_m; T, \mathbf{X}_n) = \infty\}$$

- Other concepts of breakdown (e.g. replacement).

EXAMPLES

- Mean: $\epsilon_c^* = \frac{1}{n+1}$
- Median: $\epsilon_c^* = \frac{1}{2}$
- M-estimate of location with monotone and bounded ψ function: $\epsilon_c^* = \frac{1}{2}$.



PROOF (sketch of lower bound): Let $K = \sup_r |\psi(r)|$

$$0 = \sum_{i=1}^{n+m} \psi(z_i - T_{n+m}) = \sum_{i=1}^n \psi(x_i - T_{n+m}) + \sum_{i=1}^m \psi(y_i - T_{n+m})$$

$$\left| \sum_{i=1}^n \psi(x_i - T_{n+m}) \right| = \left| \sum_{i=1}^m \psi(y_i - T_{n+m}) \right| \leq mK$$

Breakdown occurs $\Rightarrow |T_{n+m}| \rightarrow \infty$, say $T_{n+m} \rightarrow -\infty$

$$\Rightarrow \left| \sum_{i=1}^n \psi(x_i - T_{n+m}) \right| \rightarrow \left| \sum_{i=1}^n \psi(\infty) \right| \leq nK$$

Therefore, $m \geq n$ and so $\epsilon_c^* \geq 1/2$. \square

Population Version of Breakdown Point *under contamination neighborhoods*

- Model Distribution: F
- Contaminating Distribution: H
- ϵ -contaminated Distribution: $F_\epsilon = (1 - \epsilon)F + \epsilon H$
- Bias: $| T(F_\epsilon) - T(F) |$
- Max-Bias under ϵ -contamination:

$$B(\epsilon; T, F) = \sup_H | T(F_\epsilon) - T(F) |$$

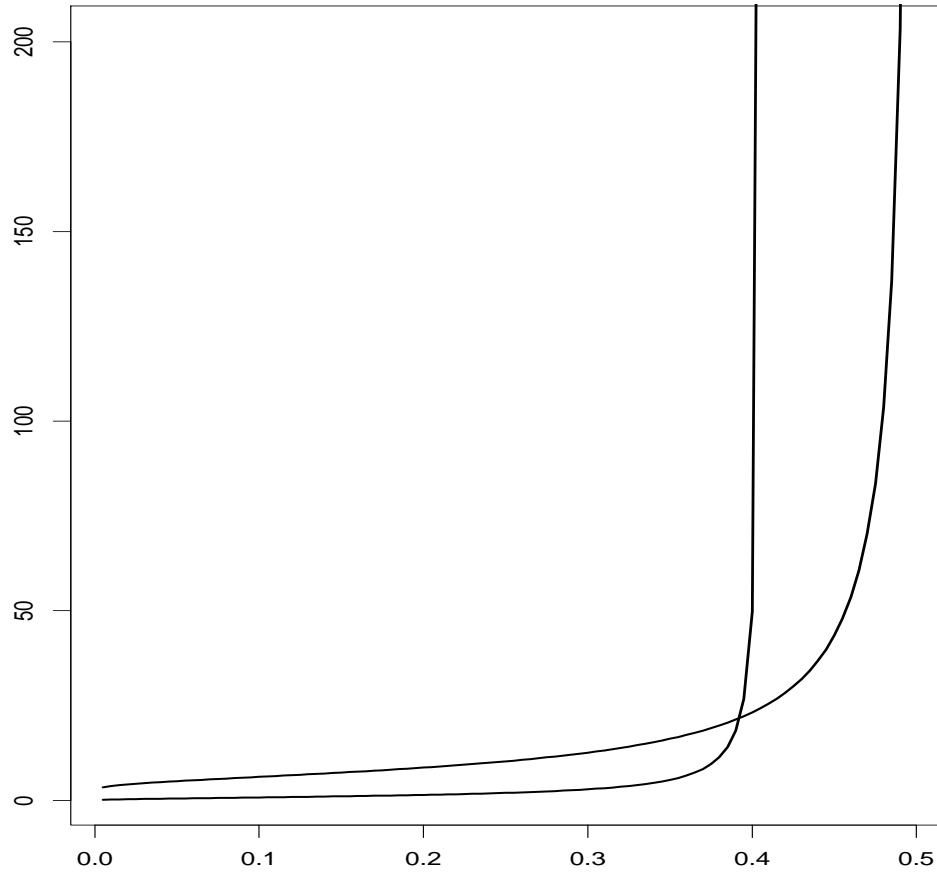
- Breakdown Point:

$$\epsilon^*(T; F) = \inf\{\epsilon \mid B(\epsilon; T, F) = \infty\}$$

- Examples

- Mean: $T(F) = E_F(X) \Rightarrow \epsilon^*(T; F) = 0$
- Median: $T(F) = F^{-1}(1/2) \Rightarrow \epsilon^*(T; F) = 1/2$

ILLUSTRATION



$$GES \approx \partial B(\epsilon; T, F) / \partial \epsilon \Big|_{\epsilon=0}$$
$$\epsilon^*(T; F) = \text{asymptote}$$

Heuristic Interpretation of Breakdown Point

subject to debate

- Proportion of bad data a statistic can tolerate before becoming arbitrary or meaningless,
- If 1/2 the data is bad then one cannot distinguish between the good and the bad data? $\Rightarrow \epsilon \leq 1/2?$

EXAMPLE

Redescending M-estimates of location with fixed scale

$$T(n_1, \dots, x_n) = \arg \min_t \sum_{i=1}^n \rho \left(\frac{|x_i - t|}{c} \right)$$

- Breakdown point. *Huber (1984)*
- For bounded increasing ρ (w.l.o.g. $\sup \rho(r) = 1$)

$$\epsilon_c^*(T; F) = \frac{1 - A(\mathcal{X}_n; c)/n}{2 - A(\mathcal{X}_n; c)/n}$$

where $A(\mathcal{X}_n; c) = \min_t \sum_{i=1}^n \rho\left(\frac{x_i - t}{c}\right)$

- Breakdown point depends on \mathcal{X}_n and c
- $\epsilon^* : 0 \rightarrow 1/2$ as $c : 0 \rightarrow \infty$
- But, for large c , $T(n_1, \dots, x_n) \approx \text{Mean}!!$

EXPLANATION

Relationship between redescending M-estimates of location and kernel density estimates

Chu, Glad, Godtliebsen and Marron (1998)

- Objective function:

$$\sum_{i=1}^n \rho\left(\frac{x_i - t}{c}\right)$$

- Kernel density estimate:

$$\hat{f}(x) \propto \sum_{i=1}^n \kappa\left(\frac{x_i - t}{h}\right)$$

- Relationship: $\kappa \propto 1 - \rho$ and tuning constant $c =$ window width h

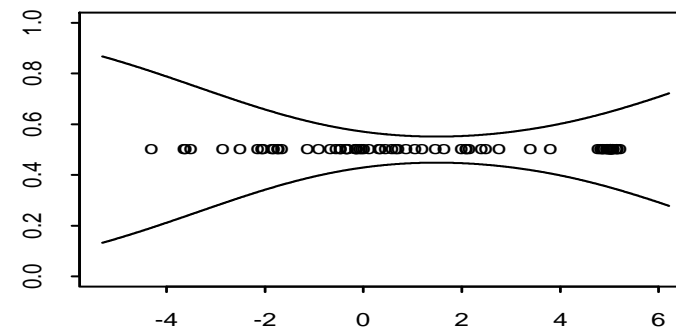
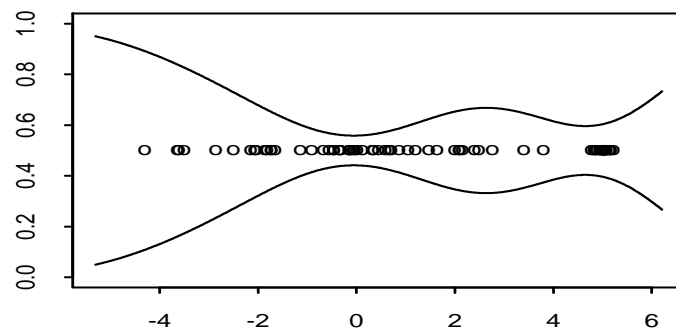
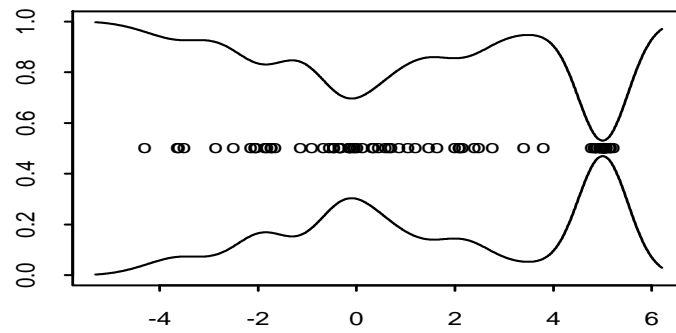
- Example:

$$\kappa(r) = \frac{1}{\sqrt{2\pi}} \exp^{-r^2/2} \quad \Rightarrow \quad \rho(r) = 1 - \exp^{-r^2/2}$$

Normal kernel Welsch's M-estimate

- Example: Epanechnikov kernel \Rightarrow skipped mean

Objective Function



Density Function

- If the “outliers” are less compact than the “good” data, then breakdown will not occur.
- This is not true for monotonic M-estimates.

PART 3

ROBUST REGRESSION

- Data: (Y_i, \mathbf{X}_i) $i = 1, \dots, n$
 - $Y_i \in \mathfrak{R}$ Response
 - $\mathbf{X}_i \in \mathfrak{R}^p$ Predictors
- Predict: Y by $\mathbf{X}'\boldsymbol{\beta}$
- Residual for a given $\boldsymbol{\beta}$: $r_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i'\boldsymbol{\beta}$
- M-estimates of regression
 - Generalization of MLE for symmetric error term
 - $Y_i = \mathbf{X}_i'\boldsymbol{\beta} + \epsilon_i$

M-ESTIMATES OF REGRESSION

- Objective function approach:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^n \rho(r_i(\boldsymbol{\beta}))$$

where $\rho(r) = \rho(-r)$ and $\rho \uparrow$ for $r \geq 0$

- M-estimating equation approach:

$$\sum_{i=1}^n \psi(r_i(\boldsymbol{\beta})) \mathbf{x}_i = 0$$

e.g. $\psi(r) = \rho'(r)$

- Interpretation: Adaptively weighted least squares.
 - Express $\psi(r) = ru(r)$ and $w_i = u(r_i(\hat{\boldsymbol{\beta}}))$

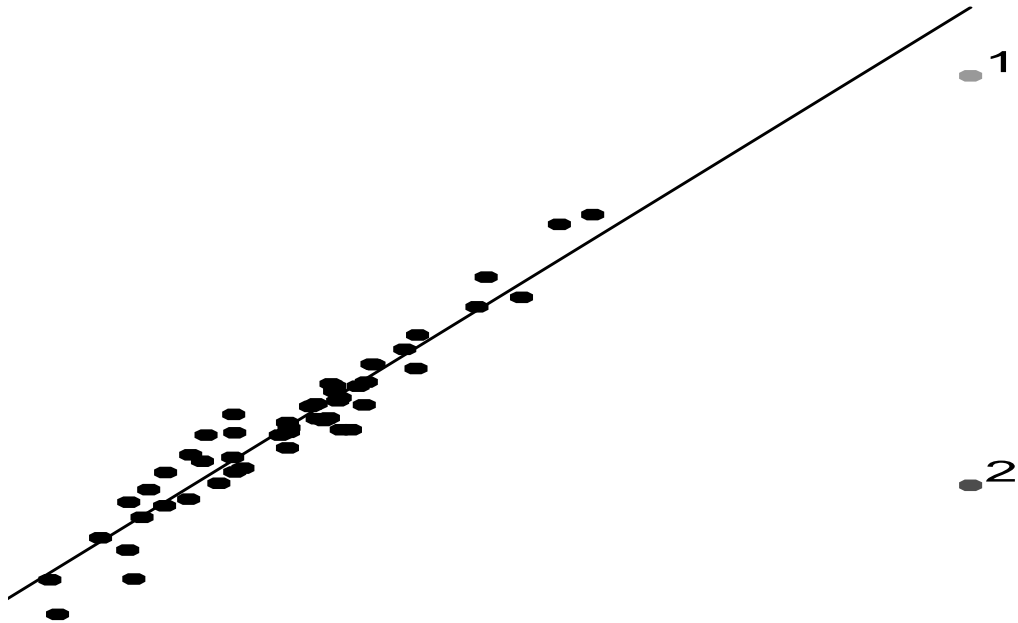
$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^n w_i \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \left[\sum_{i=1}^n w_i y_i \mathbf{x}_i \right]$$

- IRLS algorithm can be used for computations.

INFLUENCE FUNCTIONS FOR M-ESTIMATES OF REGRESSION

$$IF(y, \mathbf{x}; T, F) \propto \psi(r) \mathbf{x}$$

- $r = y - \mathbf{x}'T(F)$
- Due to residual: $\psi(r)$ • Due to Design: \mathbf{x}
- $GES = \infty$, i.e. unbounded influence.



- Outlier 1 is highly influential.
- Outlier 2 is highly influential for monotonic ψ functions, but not for redescending ψ functions.

BOUNDED INFLUENCE REGRESSION

- GM-estimates (Generalized M-estimates).

- Mallows-type (*Mallows, 1975*):

$$\sum_{i=1}^n w(\mathbf{x}_i) \psi(r_i(\boldsymbol{\beta})) \mathbf{x}_i = 0$$

Downweights outlying design points (leverage points), even if they are good leverage points.

- General form (*c.g. Maronna and Yohai, 1981*):

$$\sum_{i=1}^n \psi(\mathbf{x}_i, r_i(\boldsymbol{\beta})) \mathbf{x}_i = 0$$

- Breakdown points.

- M-estimates: $\epsilon^* = 0$

- GM-estimates: $\epsilon^* \leq 1/(p + 1)$

LATE 70's - EARLY 80's

- Open problem: Is high breakdown point regression possible?
- Yes. Repeated Median. *Siegel (1982)*.
 - Not regression equivariate
- Regression equivariance: for $a \in Re$ and A nonsingular

$$(Y_i, \mathbf{X}_i) \rightarrow (aY_i, A'\mathbf{X}_i) \Rightarrow \hat{\boldsymbol{\beta}} \rightarrow aA^{-1}\hat{\boldsymbol{\beta}}$$

i.e.

$$\hat{Y}_i \rightarrow a\hat{Y}_i$$

- Open problem: Is high breakdown point equivariate regression possible?
- Yes. Least Median of Squares. *Rousseeuw (1984)*

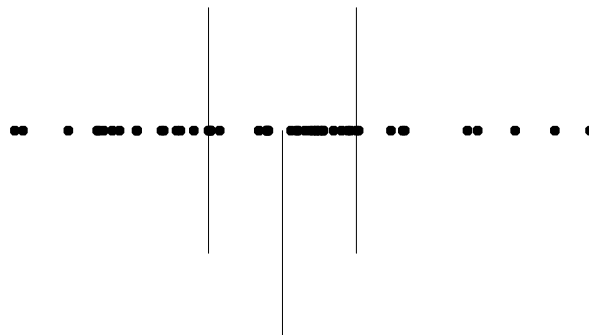
Least Median of Squares (LMS)

Hampel (1984), Rousseeuw, (1984)

- LMS:

$$\min_{\beta} \text{Median}\{r_i(\beta)^2 \mid i = 1, \dots, n\}$$

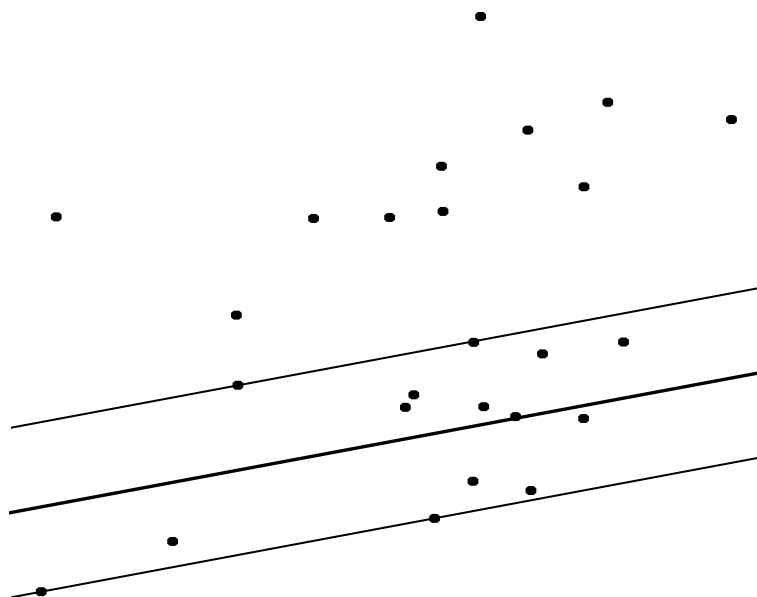
- Alternatively: $\min_{\beta} \text{MAD}\{r_i(\beta)\}$
- Breakdown point: $\epsilon^* = 1/2$
- Location version: SHORTH (*Princeton Robustness Study*)



Midpoint of the **SHORTest Half**.

LMS

- Mid-line of the shortest strip containing 1/2 of the data.



- **Problem:** Not \sqrt{n} -consistent, but only $\sqrt[3]{n}$ -consistent

$$\sqrt{n} \|\hat{\beta} - \beta\| \rightarrow_p \infty$$

$$\sqrt[3]{n} \|\hat{\beta} - \beta\| = O_p(1)$$

- **Not locally stable.** e.g. *Example is pure noise.*

S-ESTIMATES OF REGRESSION

Rousseeuw and Yohai (1984)

- For $S(\cdot)$ is an estimate of scale (about zero)

$$\min_{\beta} S(r_i(\beta))$$

- $S = MAD \Rightarrow$ LMS
- $S =$ sample standard deviation about 0 \Rightarrow least squares
- Bounded monotonic M-estimates of scale (about zero):

$$\sum_{i=1}^n \chi(|r_i|/s) = 0$$

for $\chi \uparrow$, and bounded above and below. Alternatively

$$\frac{1}{n} \sum_{i=1}^n \rho(|r_i|/s) = \epsilon$$

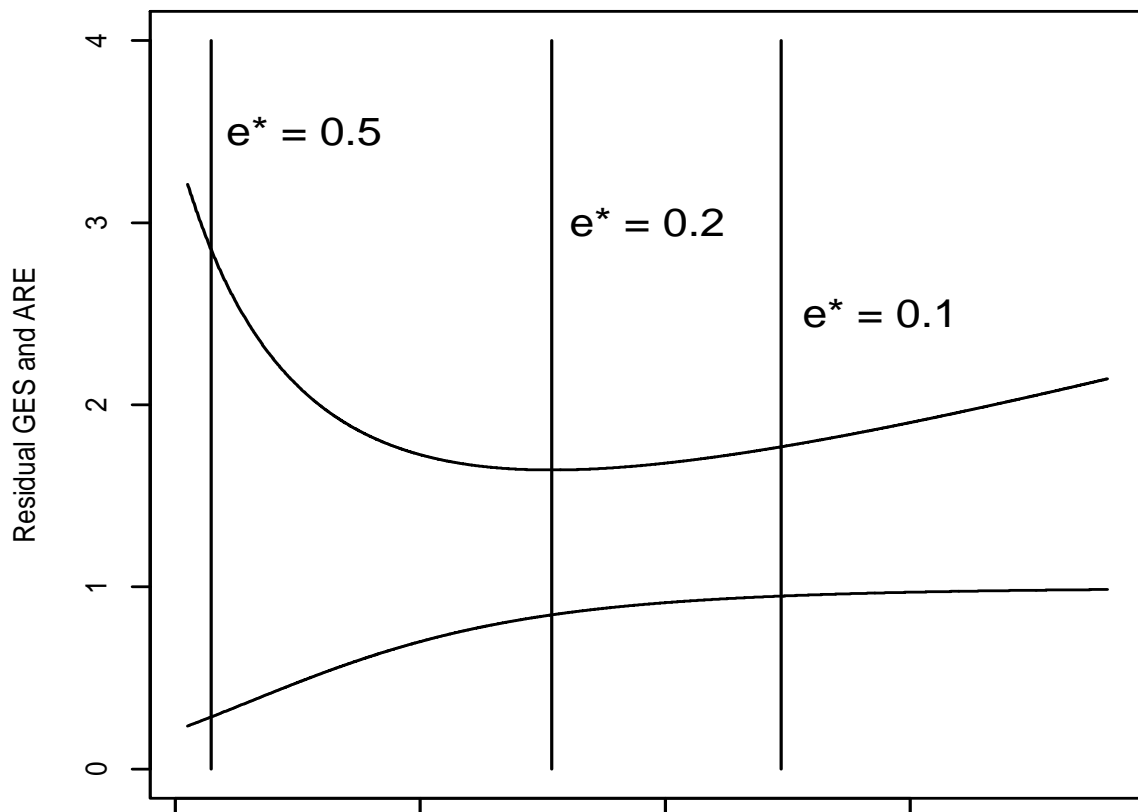
for $\rho \uparrow$, and $0 \leq \rho \leq 1$

- For LMS: $\rho = 0 - 1$ jump function and $\epsilon = 1/2$
- Breakdown point: $\epsilon^* = \min(\epsilon, 1 - \epsilon)$

S-estimates of Regression

- \sqrt{n} - consistent and Asymptotically normal.
- **TRADE-OFF:** Higher Breakdown point \Rightarrow Lower efficiency **and** lower gross error sensitivity for Normal errors.

Bi-Weight Regression S-estimates



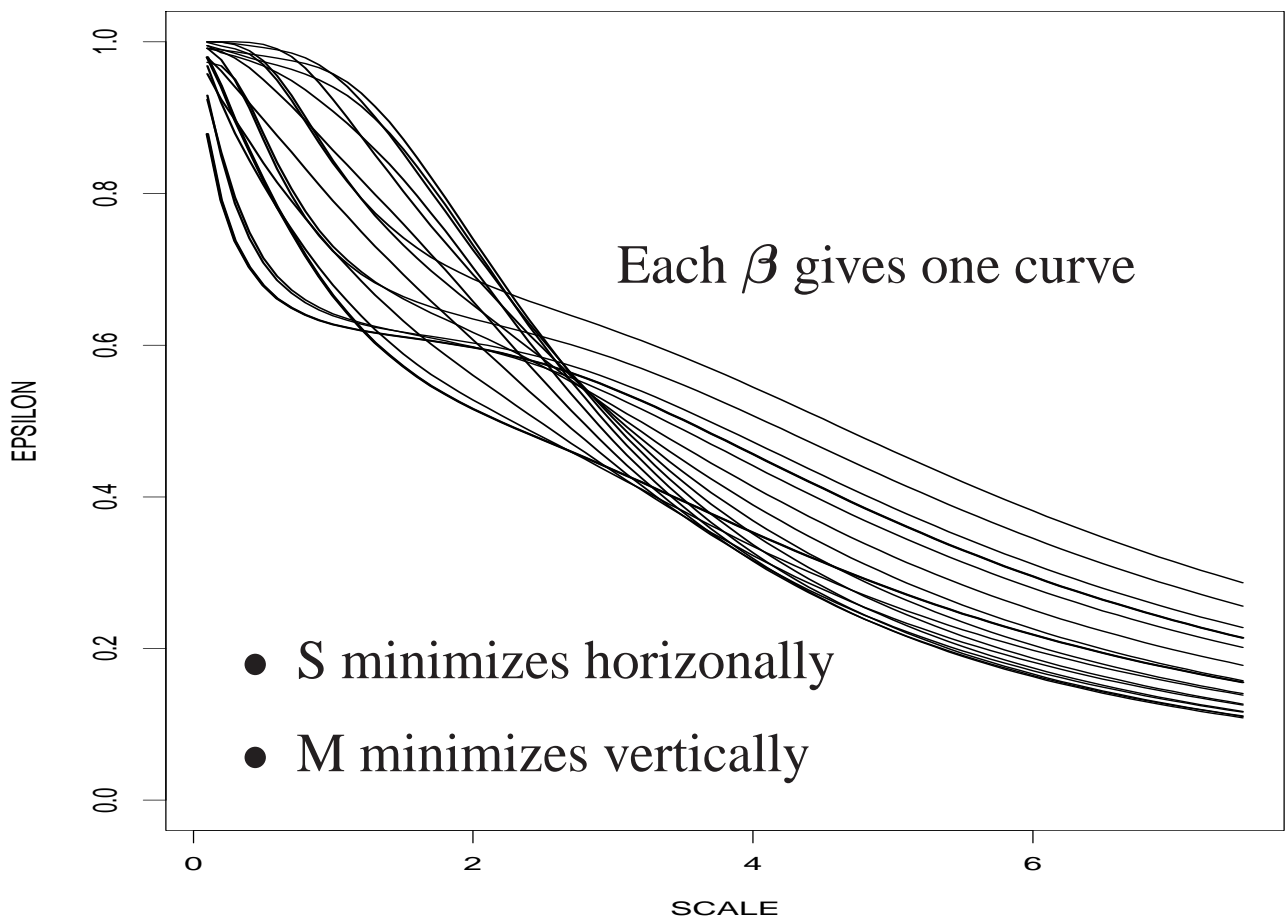
M-ESTIMATES OF REGRESSION WITH GENERAL SCALE

Martin, Yohai, and Zamar (1989)

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n \rho \left(\frac{|y_i - \mathbf{x}'_i \boldsymbol{\beta}|}{c s_n} \right)$$

- Parameter of interest: $\boldsymbol{\beta}$
- Scale statistic: s_n (consistent for σ at normal errors)
- Tuning constant: c
- Monotonic bounded $\rho \Rightarrow$ redescending M-estimate
- High Breakdown Point Examples
 - LMS and S-estimates
 - MM-estimates. *Yohai (1987)*

$$s_n \text{ .vs. } \sum_{i=1}^n \rho \left(\frac{|y_i - \mathbf{x}'_i \boldsymbol{\beta}|}{c s_n} \right)$$



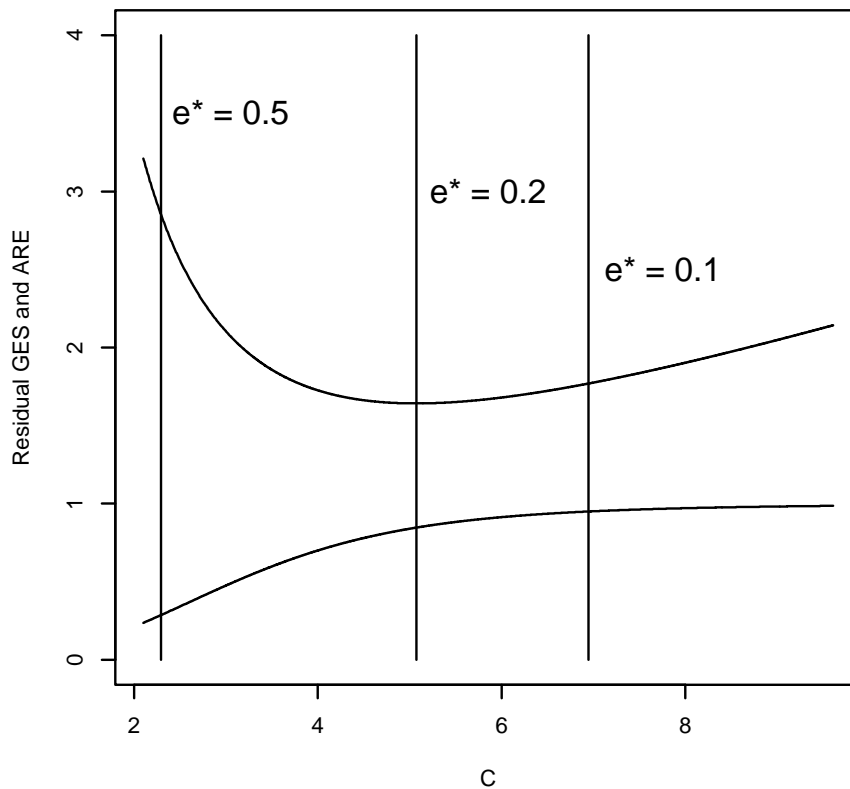
MM-ESTIMATES OF REGRESSION

Default robust regression estimate in S-plus (also SAS?)

- Begin with a preliminary estimate of regression with breakdown point $1/2$. *Usually an S-estimate of regression.*
- Compute a monotone M-estimate of scale about zero for the residuals: s_n *(Usually from the S-estimate.)*
- Compute the M-estimate of regression with scale s_n and desired tuning constant c
- Tune based upon the ARE and residual GES.
- Breakdown point: $\epsilon^* = 1/2$

TUNING

Bi-Weight Regression M-estimates



- High breakdown point S-estimates are badly tuned M-estimates.
- Tuning the S-estimates affects its breakdown point.
- MM-estimates can be tuned without affecting the breakdown point

Robust Multivariate location and Covariance Estimates for p-dimensional data.

Parallel developments

- Multivariate M-estimates. *Maronna* (1976). *Huber* (1977).
 - Adaptively weighted mean and covariances.
 - IRLS algorithms.
 - Breakdown point: $\epsilon^* \leq \frac{1}{p+1}$
- Minimum Volume Ellipsoid Estimates (MVE = LMS type). *Rousseeuw* (1985).
- Multivariate S-estimates. *Davies* (1987).
- Multivariate MM-estimates. *Tatsuoka and Tyler* (2000). *Tyler* (2002).

OTHER IMPORTANT TOPICS

Projection-based multivariate approaches

- Tukey's data depth. *Tukey* (1974).
- Stahel-Donoho's Estimate. *Stahel* (1981). *Donoho* 1982.
- Projection-estimates. *Maronna, Stahel and Yohai* (1994). *Tyler* (1996).

Computational Issues

- All known high breakdown point regression and multivariate estimates are computationally intensive.
- Random Elemental Subset Selection. *Rousseeuw* (1984).

References

- Forthcoming.