

**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: $F(x_1, \dots, x_n | \theta)$

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?

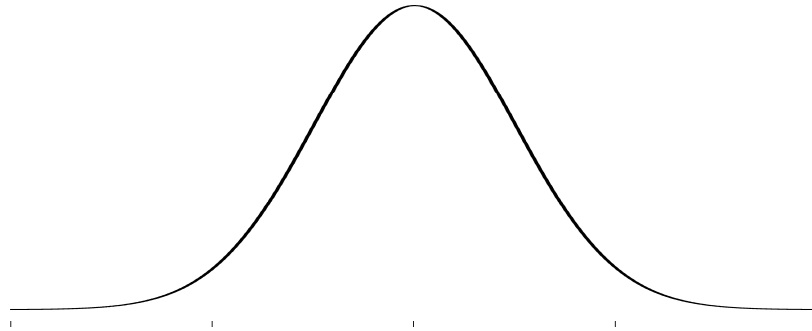
A: The model is assumed to be approximately true, e.g. symmetric and unimodal (past experience).

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

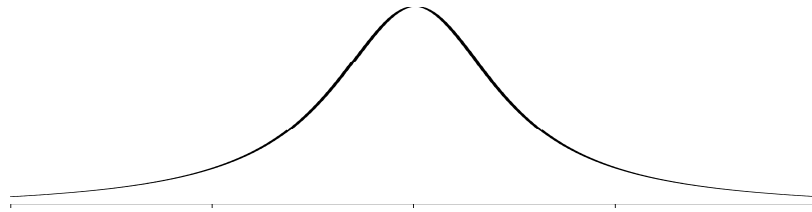


- Symmetric distributions: $\mu = \begin{cases} \text{population mean} \\ \text{population median} \end{cases}$
- Sample mean: $\bar{X} \approx \text{Normal} \left(\mu, \frac{\sigma^2}{n} \right)$
- Sample median: $\text{Median} \approx \text{Normal} \left(\mu, \frac{1}{n} \frac{1}{4f(\mu)^2} \right)$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

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



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

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

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

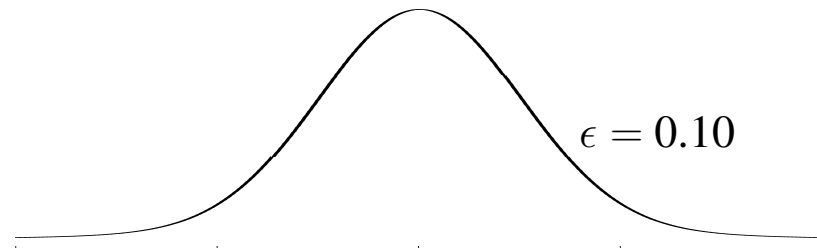
	$\nu \leq 2$	3	4	5
$ARE(\text{Median}, \bar{X})$	∞	1.621	1.125	0.960
$ARE(\bar{X}, \text{Median})$	0	0.617	0.888	1.041

MIXTURE OF NORMALS

Theory of errors: Central Limit Theorem gives plausibility to normality.

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

i.e. not all measurements are equally precise.



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

- **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}

$$\text{Statistic: } T_{n-1} = T(x_1, \dots, x_n)$$

- *Contaminated Data Set:* $x_1, \dots, x_{n-1}, \mathbf{x}$

$$\text{Contaminated Value: } T_n = T(x_1, \dots, x_n, \mathbf{x})$$

THE SENSITIVITY CURVE (Tukey, 1970)

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

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. $T(F_\epsilon)$
- Given **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:** $T(F) = E_F[X]$.

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

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

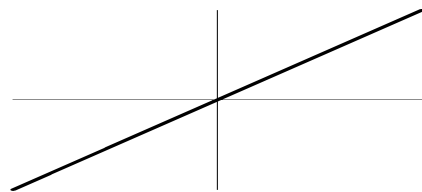
- **Median:** $T(F) = F^{-1}(1/2)$

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

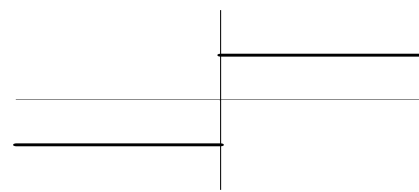
Plots of Influence Functions

Gives insight into the behavior of a statistic.

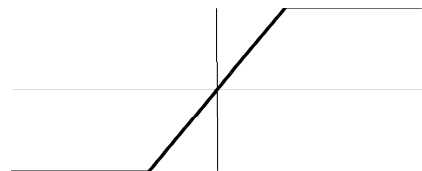
$$T_n \approx T_{n-1} + \frac{1}{n} IF(\mathbf{x}; T, F)$$



Mean



Median



α -trimmed mean



α -Winsorized mean
(*somewhat unexpected?*)

Desirable robustness properties for 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

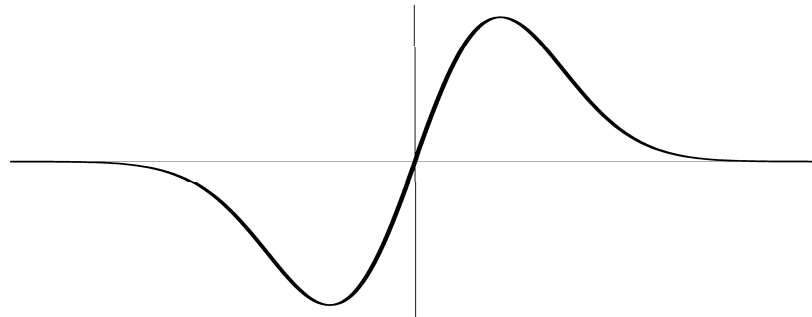
$$AV(T; F) = E_F[IF(X; T, F)^2]$$

- In general,

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

- Trade-off at the normal model: **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**

CLASSES OF ESTIMATES

L-statistics: Linear combination of order statistics

- Let $X_{(1)} \geq \dots \geq X_{(n)}$ represent the order statistics.

$$T(X_1, \dots, X_n) = \sum_{i=1}^n a_{i,n} X_{(i)}$$

where $a_{i,n}$ are constants.

- **Examples:** ○ **Mean.** $a_{i,n} = 1/n$

- **Median.**
$$a_{i,n} = \begin{cases} 1 & i = \frac{n+1}{2} \\ 0 & i \neq \frac{n+1}{2} \end{cases} \quad a_{i,n} = \begin{cases} \frac{1}{2} & i = \frac{n}{2}, \frac{n}{2} + 1 \\ 0 & \text{otherwise} \end{cases}$$

for n odd for n even

- α -trimmed mean.

- α -Winsorized mean.

- General form for the influence function exists
- Can obtain any desirable monotonic shape, but not redescending
- Do not readily generalize to other settings.

M-ESTIMATES

Huber (1964, 1967)

Maximum likelihood type estimates
under non-standard conditions

• **One-Parameter Case.** X_1, \dots, X_n i.i.d. $f(x; \theta)$ $\theta \in \Theta$

• **Maximum likelihood estimates**

– Likelihood function. $L(\theta | x_1, \dots, x_n) = \prod_{i=1}^n f(x_i; \theta)$

– Minimize the negative log-likelihood.

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

– Solve the likelihood equations

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

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.

M-ESTIMATES OF LOCATION

A symmetric and translation equivariant M-estimate.

Translation equivariance

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

gives

$$\rho(x; t) = \rho(x - t) \text{ and } \psi(x; t) = \psi(x - t)$$

Symmetric

$$X_i \rightarrow -X_i \Rightarrow T_n \rightarrow -T_n$$

gives

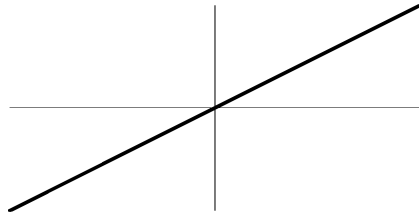
$$\rho(-r) = \rho(r) \text{ or } \psi(-r) = -\psi(r).$$

INFLUENCE FUNCTION FOR M-ESTIMATES

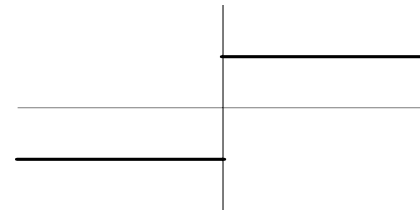
$$IF(\mathbf{x}; T, F) = c(T, F)\psi(\mathbf{x}; T(F))$$

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.

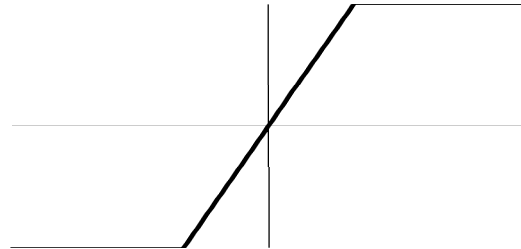


⇒ Mean



⇒ Median

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

- Recall translation equivariance and symmetry implies

$$\psi(x; t) = \psi(x - t) \quad \text{and} \quad \psi(-r) = -\psi(r).$$

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

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

- **W-estimates:** $w_i = u(x_i - \hat{\theta}_o)$
where $\hat{\theta}_o$ is a preliminary estimate of location, such as the median.

ADAPTIVELY WEIGHTED MEANS

The weights are determined by the data cloud.

●●●●●●●●
 $\hat{\theta}$

● ●
 $\overbrace{\hspace{10em}}$
Heavily Downweighted

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

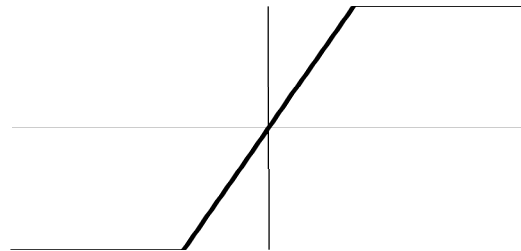
$$w_{i,k} = u(x_i - \theta_k)$$

- **W-estimate = One-step M-estimate.**

SOME COMMON M-ESTIMATES OF LOCATION

Huber's M-estimate

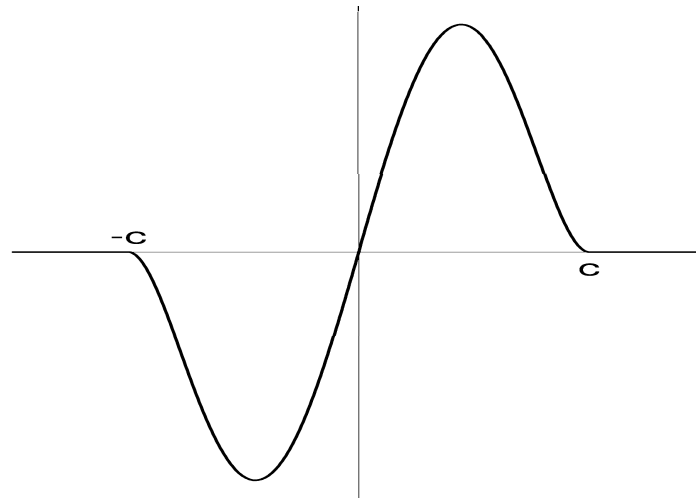
$$\psi(r) = \begin{cases} c & r \geq c \\ r & |r| < c \\ -c & r \leq -c \end{cases}$$



- 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.

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

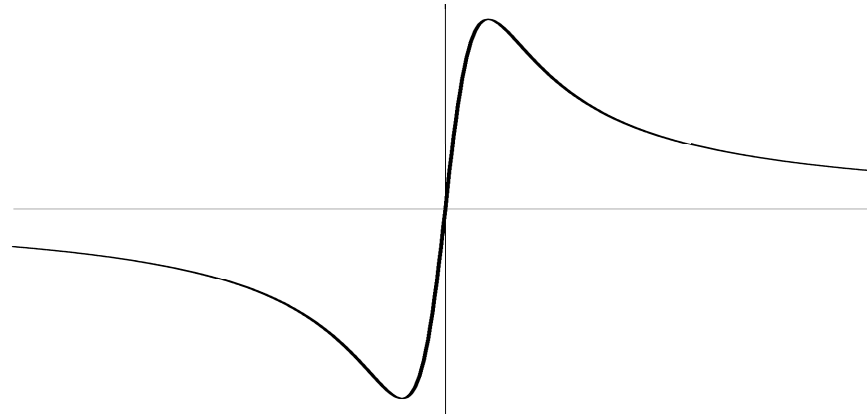
$$u(r) = \left\{ \left(1 - \frac{r^2}{c^2} \right)_+ \right\}^2 \quad \text{where } a_+ = \max\{0, a\}$$



- 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 STRONG REDESCENDER.

PART 2

MORE ADVANCED CONCEPTS AND METHODS

SCALE EQUIVARIANCE

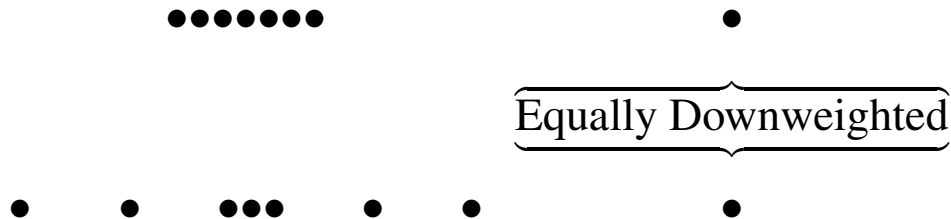
- M-estimates of location alone are not scale equivariant in general, i.e.

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

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

(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

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

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

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**

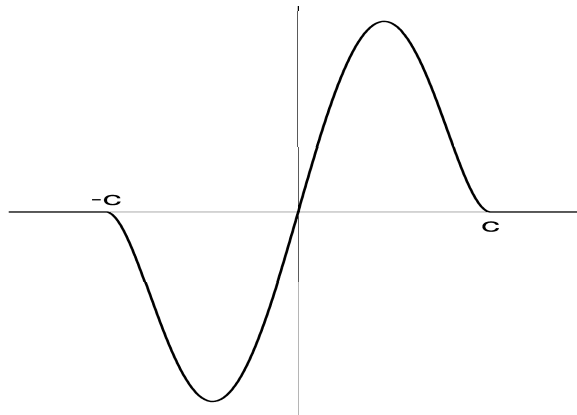
Tuning an M-estimate

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

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

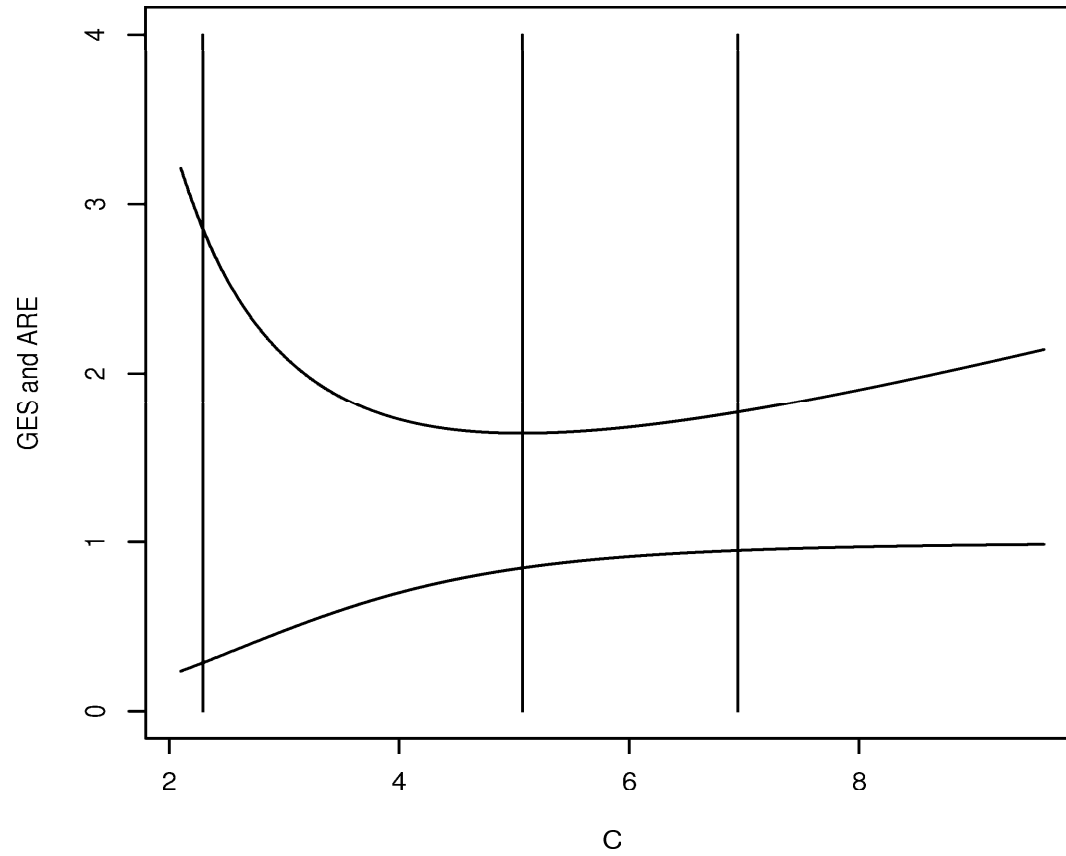
- Tukey's Bi-weight.

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



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

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



MORE THAN ONE OUTLIER

- GES: measure of local robustness.
- Measures 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 and THE BREAKDOWN POINT

- Max-Bias under ϵ_m -contamination:

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

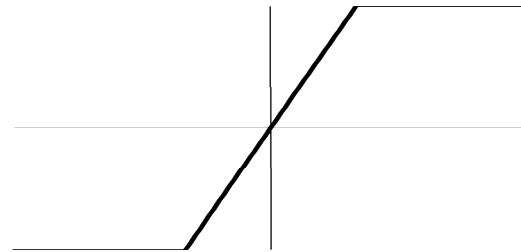
- Finite sample contamination breakdown point: *Donoho and Huber (1983)*

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

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

EXAMPLES

- Mean: $\epsilon_c^* = \frac{1}{n+1}$
- Median: $\epsilon_c^* = 1/2$
- α -trimmed mean: $\epsilon_c^* = \alpha$
- α -Windsorized mean: $\epsilon_c^* = \alpha$
- M-estimate of location with monotone and bounded ψ function:



$$\epsilon_c^* = 1/2$$

PROOF

(sketch of lower bound)

Let $K = \sup_r |\psi(r)|$.

$$\begin{aligned} 0 &= \sum_{i=1}^{n+m} \psi(z_i - T_{n+m}) = \sum_{i=1}^n \psi(x_i - T_{n+m}) + \sum_{i=1}^m \psi(y_i - T_{n+m}) \\ &\quad \left| \sum_{i=1}^n \psi(x_i - T_{n+m}) \right| = \left| \sum_{i=1}^m \psi(y_i - T_{n+m}) \right| \leq mK \end{aligned}$$

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: $F_\epsilon = (1 - \epsilon)F + \epsilon H$

• Max-Bias under ϵ -contamination:

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

• Breakdown Point: *Hampel* (1968)

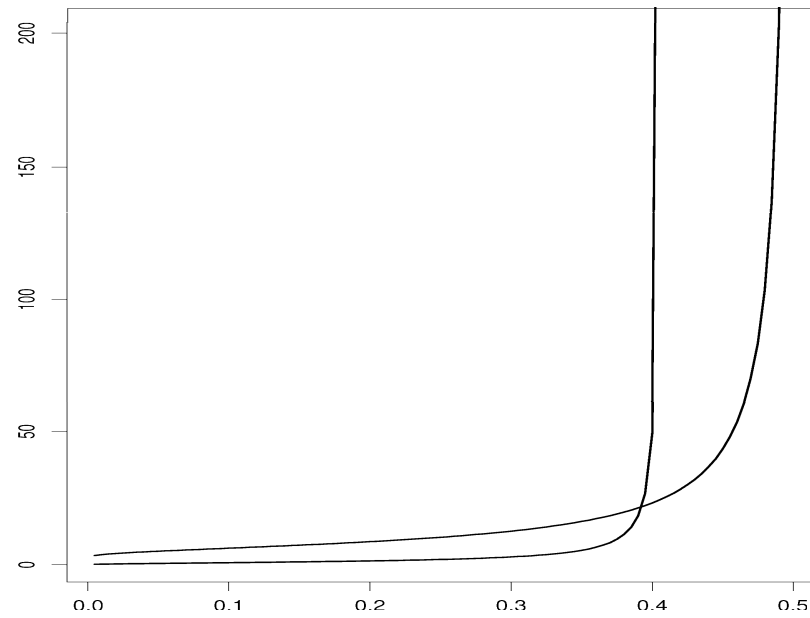
$$\epsilon^*(T; F) = \inf\{\epsilon \mid B(\epsilon; T, F) = \infty\}$$

• 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 data 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 ,

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

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

- Breakdown point depends on \mathcal{X}_n and c
- $\epsilon^* : 0 \rightarrow 1/2$ as $c : 0 \rightarrow \infty$
- 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:**

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

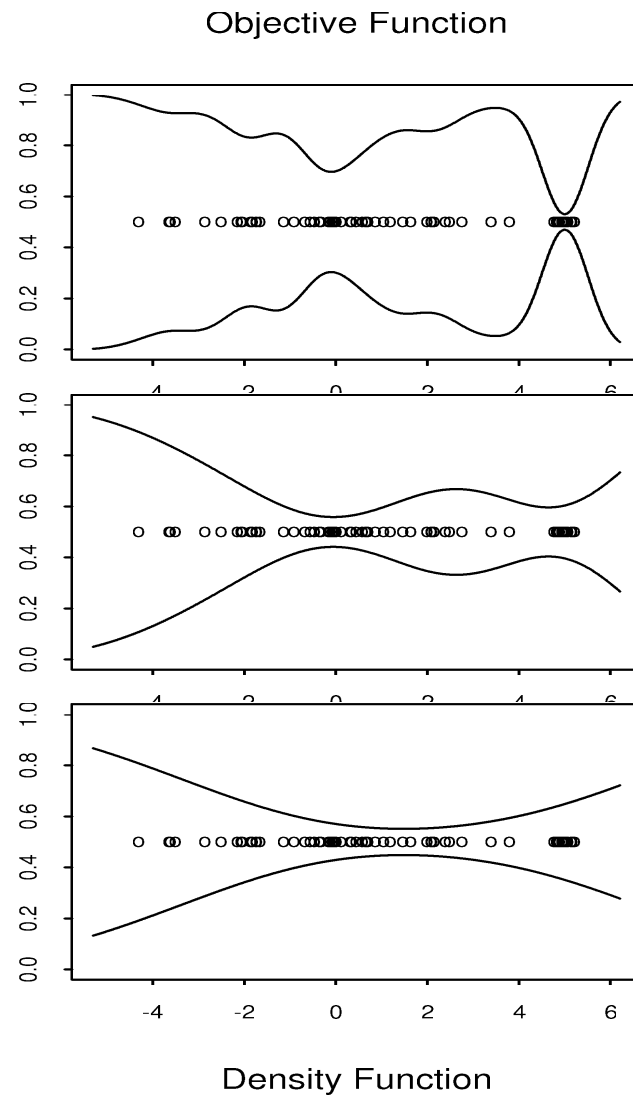
- **Kernel density estimate:**

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

- **Relationship:** $\kappa \propto 1 - \rho$ and $c = \text{window width } h$

- **Example:** $\kappa(r) = \frac{1}{\sqrt{2\pi}} \exp^{-r^2/2} \Rightarrow \rho(r) = 1 - \exp^{-r^2/2}$
Normal kernel \Rightarrow Welsch's M-estimate

- **Example:** Epanechnikov kernel \Rightarrow skipped mean



- “Outliers” less compact than “good” data \Rightarrow Breakdown will not occur.
- Not true for monotonic M-estimates.

PART 3
ROBUST REGRESSION

REGRESSION SETTING

- 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
 - $Y_i = \mathbf{X}'_i\boldsymbol{\beta} + \epsilon_i$ where ϵ_i are i.i.d. symmetric.

M-ESTIMATES OF REGRESSION

- **Objective function approach:**

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

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 $w_i = u(r_i(\hat{\boldsymbol{\beta}}))$

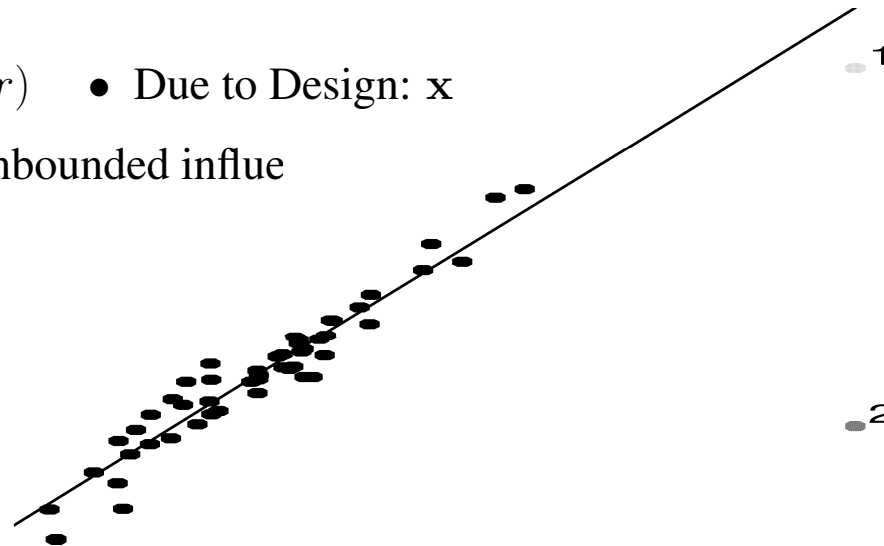
$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^n w_i \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \left[\sum_{i=1}^n w_i y_i \mathbf{x}_i \right]$$

- **Computations via IRLS algorithm.**

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 influe



- **Outlier 1:** highly influential for any ψ function.
- **Outlier 2:** highly influential for monotonic but not redescending ψ .

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

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

i.e.

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

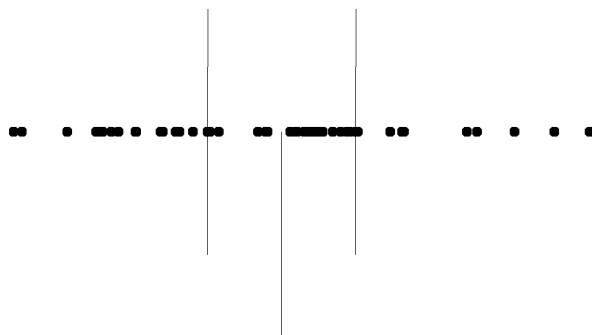
- **Open problem:** Is high breakdown point equivariate regression possible?
- **Yes.** Least Median of Squares. *Rousseeuw (1984)*

Least Median of Squares (LMS)

Hampel (1984), Rousseeuw, (1984)

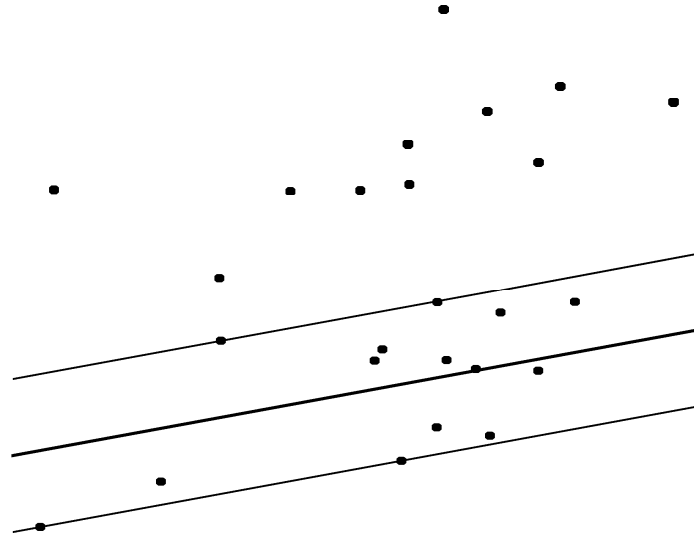
$$\min_{\beta} \text{Median}\{r_i(\beta)^2 \mid i = 1, \dots, n\}$$

- 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 $S(\cdot)$, 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,

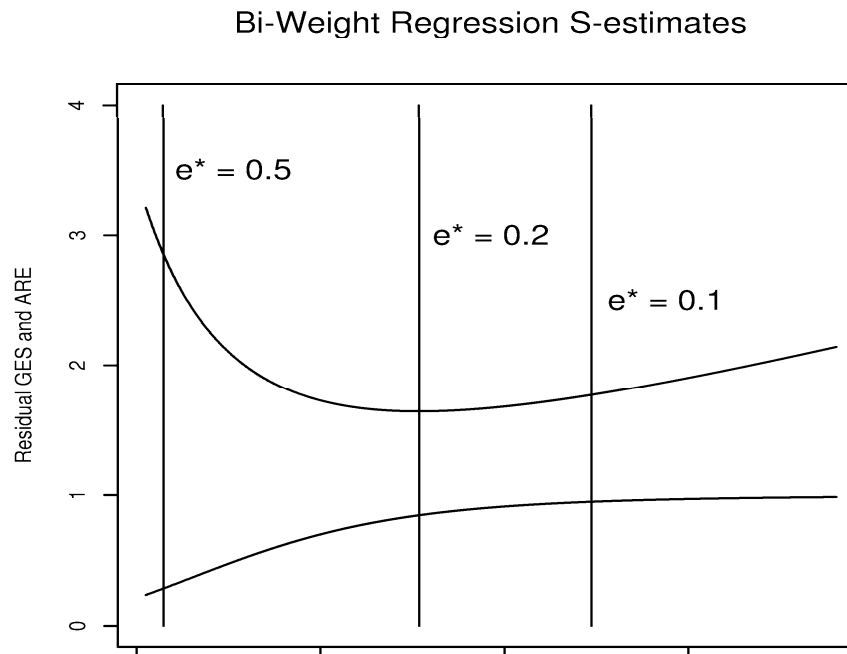
$$\frac{1}{n} \sum_{i=1}^n \rho(|r_i|/s) = \epsilon$$

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 residual gross error sensitivity for Normal errors.



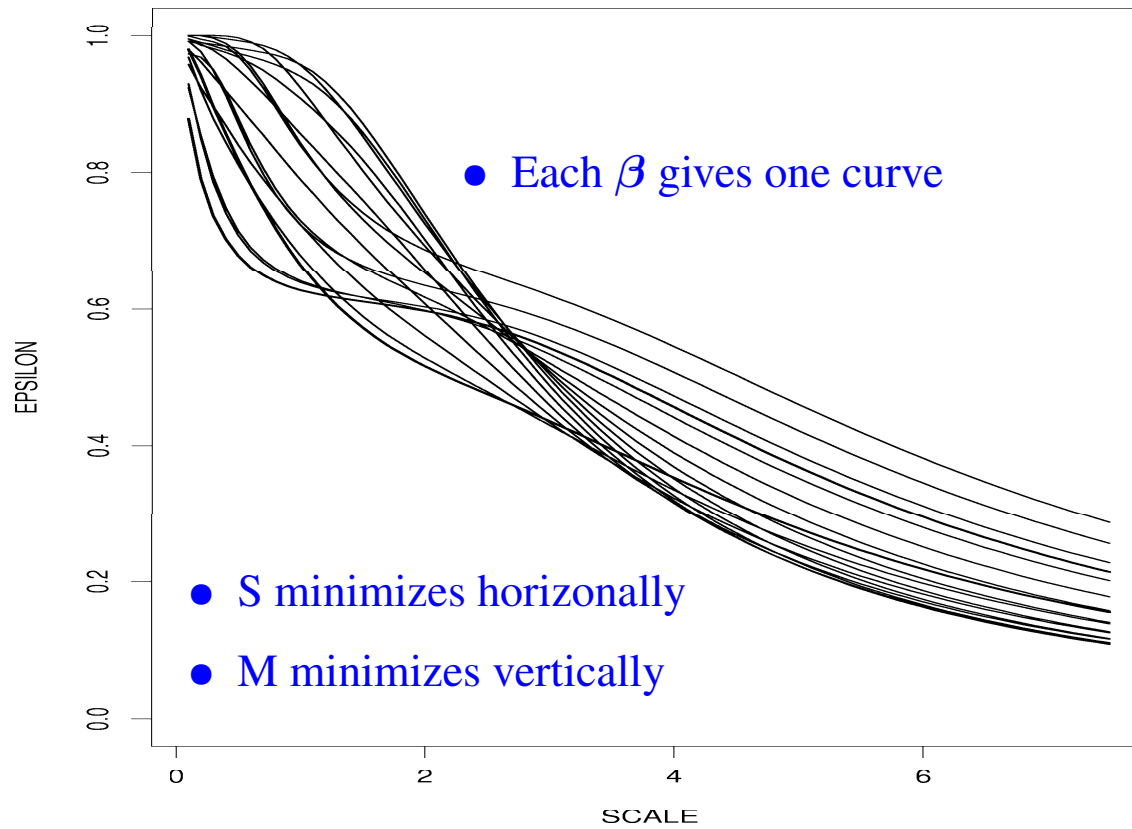
M-ESTIMATES OF REGRESSION WITH GENERAL SCALE

Martin, Yohai, and Zamar (1989)

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

- Parameter of interest: β
- 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)*

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



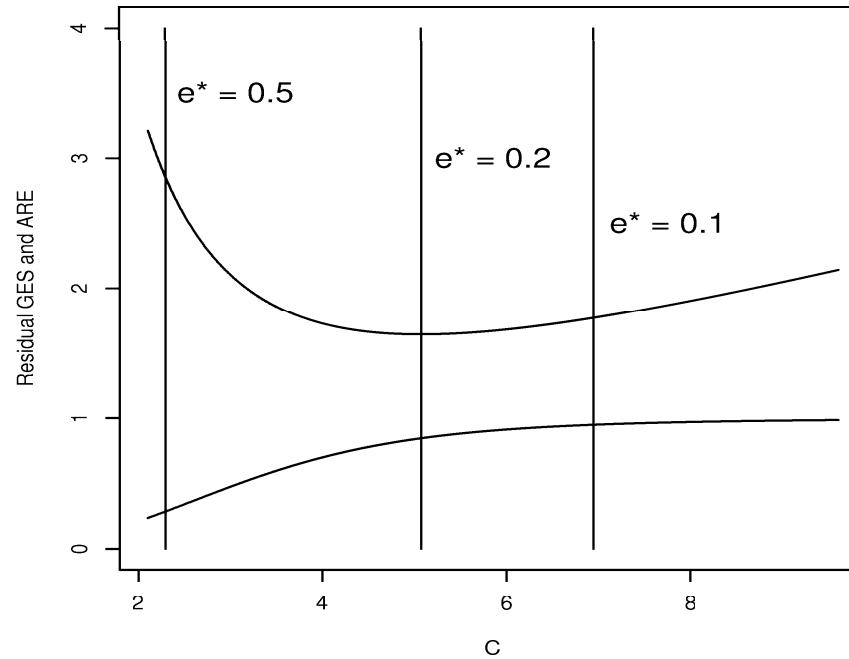
MM-ESTIMATES OF REGRESSION

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

- Given 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 an M-estimate of regression using the scale statistics s_n and with any desired tuning constant c .
- Tune to obtain reasonable 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

Parallel developments

- **Multivariate M-estimates.** *Maronna (1976). Huber (1977).*
 - Adaptively weighted mean and covariances.
 - IRLS algorithms.
 - Breakdown point: $\epsilon^* \leq 1/(p + 1)$, where p is the dimension of the data.
- **Minimum Volume Ellipsoid Estimates.** *Rousseeuw (1985).*
 - MVE is a multivariate version of LMS.
- **Multivariate S-estimates.** *Davies (1987), Lopuhuaä (1989)*
- **Multivariate MM-estimates.** *Tatsuoka and Tyler (2000). Tyler (2002).*

OTHER 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

- **Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.**
- **Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.**
- **Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006). *Robust Statistics: Theory and Methods*. Wiley, New York.**

Web-Site for this Short-Course

www.rci.rutgers.edu/dtyler/ShortCourse.pdf