A SHORT COURSE ON ROBUST STATISTICS

David E. Tyler

Rutgers

The State University of New Jersey

Web-Site

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

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.**

PART 1

CONCEPTS AND BASIC METHODS

MOTIVATION

- Data Set: $X_1, X_2, \ldots, X_n$
- Parametric Model: $F(x_1, \ldots, x_n \mid \theta)$

θ: Unknown parameter
F: Known function

- e.g. $X_1, X_2, \ldots, X_n$ 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?

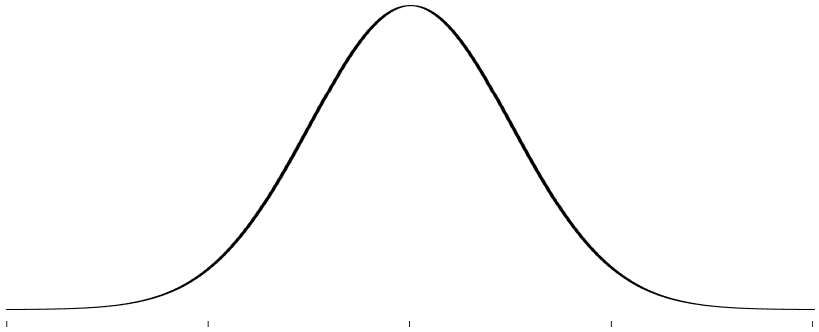
A: The model is assumed to be approximately true, e.g. symmetric and unimodal (past experience).

Q: Are statistical methods which are good under the model reasonably good if the model is only approximately true?

ROBUST STATISTICS

Formally addresses this issue.

CLASSIC EXAMPLE: MEAN .vs. MEDIAN

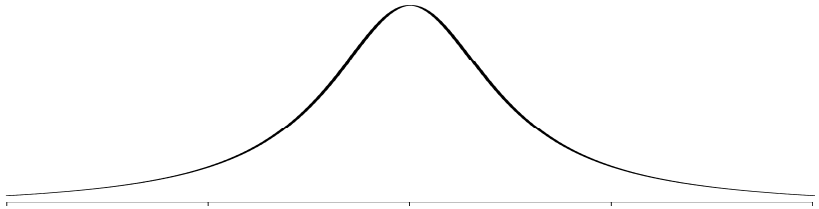


- Symmetric distributions: $\mu = \begin{cases} \text{population mean} \\ \text{population median} \end{cases}$
- Sample mean: $\overline{X} \approx Normal\left(\mu, \frac{\sigma^2}{n}\right)$
- Sample median: $Median \approx Normal\left(\mu, \frac{1}{n}\ \frac{1}{4f(\mu)^2}\right)$
- At normal: $Median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

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

CAUCHY DISTRIBUTION

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



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

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

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

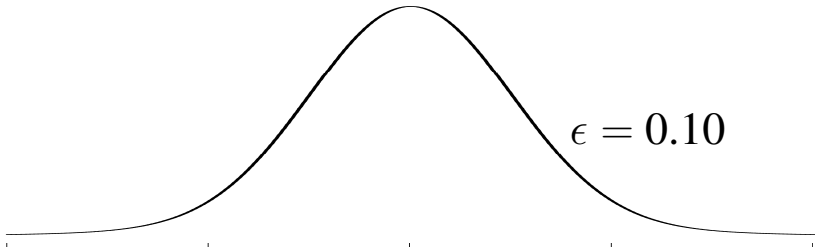
	$\nu \leq 2$	3	4	5
$ARE(Median, \overline{X})$	∞	1.621	1.125	0.960
$ARE(\overline{X}, Median)$	0	0.617	0.888	1.041

MIXTURE OF NORMALS

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

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

i.e. not all measurements are equally precise.



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

- Classic paper: Tukey (1960), A survey of sampling from contaminated distributions.
  - For $\epsilon > 0.10 \Rightarrow ARE(Median, \overline{X}) > 1$
  - The mean absolute deviation is more effiicent 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

Measuring the robustness of a statistics

- **Relative Efficiency over a range of distributional models.**
  - There exist estimates of location which are asymptotically most efficient for the center of any symmetric distribution. (Adaptive estimation, semi-parametrics).
  - **Robust?**

- 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*: $x_1, \ldots, x_{n-1}$

  Statistic: $T_{n-1} = T(x_1, \ldots, x_{n-1})$

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

  Contaminated Value: $T_n = T(x_1, \ldots, x_{n-1}, \mathbf{x})$

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)

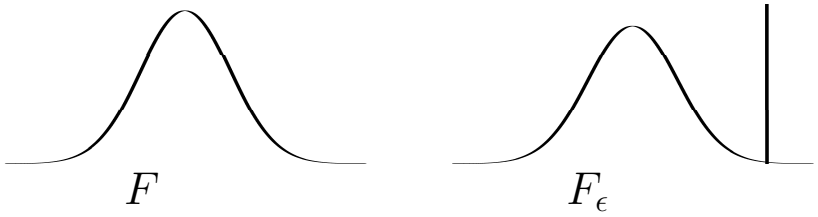
Population version of the sensitivity curve.

THE INFLUENCE FUNCTION

- Statistic $T_n = T(F_n)$ estimates $T(F)$.
- 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*):

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

(*e.g. the mode is not qualitatively robust*)

- Influence Function (*Infinitesimal pertubation: Gâteaux Derivative*)

$$IF(\mathbf{x}; T, F) = \lim_{\epsilon \to 0} \frac{T(F_\epsilon) - T(F)}{\epsilon} = \frac{\partial}{\partial \epsilon} T(F_\epsilon)\Big|_{\epsilon=0}$$

EXAMPLES

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

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

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

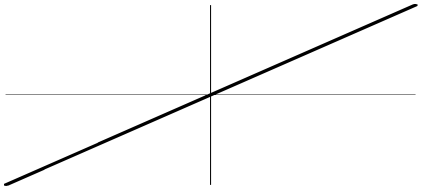
- **Median:** $T(F) = F^{-1}(1/2)$

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

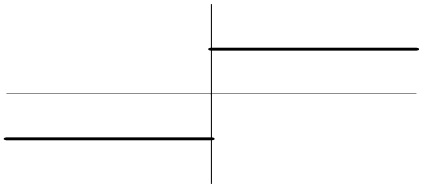
Plots of Influence Functions

Gives insight into the behavior of a statistic.

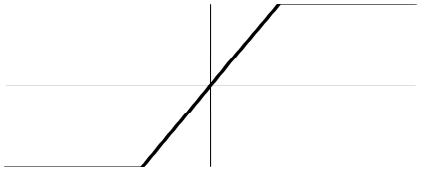
$$T_n \approx T_{n-1} + \frac{1}{n} IF(\mathbf{x}; T, F)$$



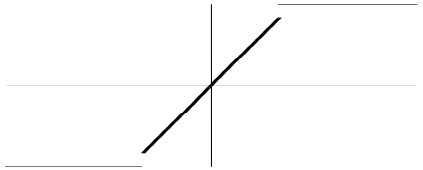
Mean



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

$$Note : E_F[\, IF(X;T,F) \,] = 0$$

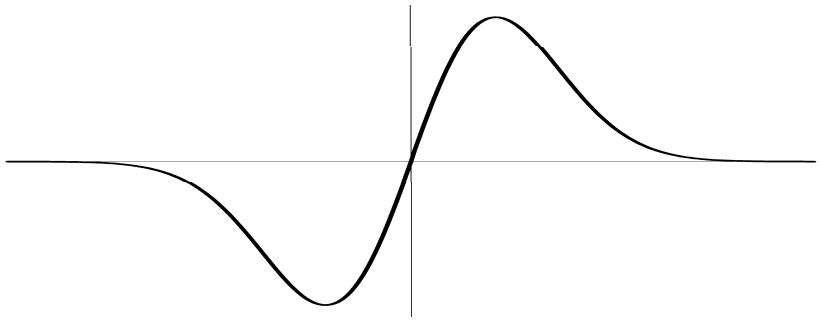
$$AV(T;F) = E_F[\, IF(X;T,F)^2 \,]$$

- Under general conditions, e.g. *Frèchet differentiability*,

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

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

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

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

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

$$a_{i,n} = \begin{cases} 1 & i = \frac{n+1}{2} \\ 0 & i \neq \frac{n+1}{2} \end{cases} \quad \text{for n odd}$$

$$a_{i,n} = \begin{cases} \frac{1}{2} & i = \frac{n}{2}, \frac{n}{2}+1 \\ 0 & \text{otherwise} \end{cases} \quad \text{for n even}$$

  - ◦ 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.** $X_1, \ldots, X_n$ i.i.d. $f(x;\theta), \quad \theta \in \Theta$
- **Maximum likelihood estimates**
  - Likelihood function. $L(\theta \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\theta)$
  - Minimize the negative log-likelihood.

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

  - 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: $\widehat{\theta} = \arg\min \Sigma_{i=1}^{n} \rho(x_i; \theta)$

- M-estimating equation approach: $\Sigma_{i=1}^{n} \psi(x_i; \widehat{\theta}) = 0.$

  – *Note*: Unique solution when $\psi(x; \theta)$ is strictly monotone in θ.

- **Basic examples.**

  – **Mean.** MLE for Normal: $f(x) = (2\pi)^{-1/2} e^{-\frac{1}{2}(x-\theta)^2}$ for $x \in \Re$.

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

  – **Median.** MLE for Double Exponential: $f(x) = \frac{1}{2} e^{-|x-\theta|}$ for $x \in \Re$.

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

- ρ and ψ need not be related to any density or to each other.
- Estimates can be evaluated under various distributions.

M-ESTIMATES OF LOCATION

A symmetric and translation equivariant M-estimate.

Translation equivariance

$$X_i \to X_i + a \Rightarrow T_n \to T_n + a$$

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

Symmetric

$$X_i \to -X_i \Rightarrow T_n \to -T_n$$

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

Alternative derivation

Generalization of MLE for center of symmetry for a given family of symmetric distributions.

$$f(x;\theta) = g(|x-\theta|)$$

INFLUENCE FUNCTION OF M-ESTIMATES

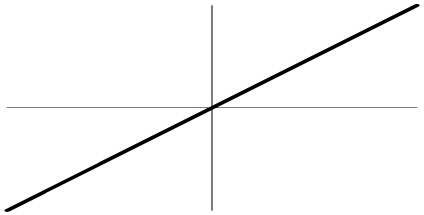
M-FUNCTIONAL: $T(F)$ is the solution to $E_F[\psi(\mathbf{X}; T(F))] = 0$.

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

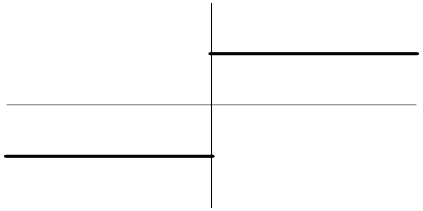
where $c(t, f) = -1/E_F\left[\frac{\partial\psi(\mathbf{X};\theta)}{\partial\theta}\right]$ evaluated at $\theta = T(F)$.

Note: $E_F[\, IF(X; T, F)\,] = 0$.

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

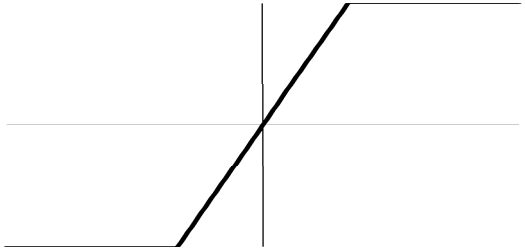


$\Rightarrow$ Mean



$\Rightarrow$ Median

EXAMPLE



Choose

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

where c is a tuning constant.

Huber's M-estimate

Adaptively trimmed mean.

i.e. *the proportion trimmed depends upon the data.*

DERIVATION OF INFLUENCE FUNCTION FROM M-ESTIMATES

Sketch

Let $T_\epsilon = T(F_\epsilon)$, and so

$$0 = E_{F_\epsilon}[\psi(\mathbf{X}; T_\epsilon)] = (1-\epsilon)E_F[\psi(\mathbf{X}; T_\epsilon)] + \epsilon\, \psi(\mathbf{x}; T_\epsilon).$$

Taking the derivative with respect to $\epsilon \Rightarrow$

$$0 = -E_F[\psi(\mathbf{X}; T_\epsilon)] + (1-\epsilon)\frac{\partial}{\partial\epsilon}E_F[\psi(\mathbf{X}; T_\epsilon)] + \psi(\mathbf{x}; T_\epsilon) + \epsilon\frac{\partial}{\partial\epsilon}\psi(\mathbf{x}; T_\epsilon).$$

Let $\psi'(x,\theta) = \partial\psi(x,\theta)/\partial\theta$. Using the chain rule $\Rightarrow$

$$0 = -E_F[\psi(\mathbf{X}; T_\epsilon)] + (1-\epsilon)E_F[\psi'(\mathbf{X}; T_\epsilon)]\frac{\partial T_\epsilon}{\partial\epsilon} + \psi(\mathbf{x}; T_\epsilon) + \epsilon\psi'(\mathbf{x}; T_\epsilon)\frac{\partial T_\epsilon}{\partial\epsilon}.$$

Letting $\epsilon \to 0$ and using qualitative robustness, i.e. $T(F_\epsilon) \to T(F)$, then gives

$$0 = E_F[\psi'(\mathbf{X}; T)]IF(\mathbf{x}; T(F)) + \psi(\mathbf{x}; T(F)) \Rightarrow \text{RESULTS}.$$

ASYMPTOTIC NORMALITY OF M-ESTIMATES

Sketch

Let $\theta = T(F)$. Using Taylor series expansion on $\psi(x;\hat{\theta})$ about θ gives

$$0 = \sum_{i=1}^{n} \psi(x_i;\hat{\theta}) = \sum_{i=1}^{n} \psi(x_i;\theta) + (\hat{\theta}-\theta)\sum_{i=1}^{n} \psi'(x_i;\theta) + \ldots,$$

or

$$0 = \sqrt{n}\{\frac{1}{n}\sum_{i=1}^{n} \psi(x_i;\theta)\} + \sqrt{n}(\hat{\theta}-\theta)\{\frac{1}{n}\sum_{i=1}^{n} \psi'(x_i;\theta)\} + O_p(1/\sqrt{n}).$$

By the CLT, $\sqrt{n}\{\frac{1}{n}\Sigma_{i=1}^{n}\psi(x_i;\theta)\} \rightarrow_d \mathbf{Z} \sim Normal(0, E_F[\psi(\mathbf{X};\theta)^2])$.

By the WLLN, $\frac{1}{n}\Sigma_{i=1}^{n}\psi'(x_i;\theta) \rightarrow_p E_F[\psi'(\mathbf{X};\theta)]$.

Thus, by Slutsky's theorem,

$$\sqrt{n}(\hat{\theta}-\theta) \rightarrow_d -\mathbf{Z}/E_F[\psi'(\mathbf{X};\theta)] \sim Normal(0,\sigma^2),$$

where

$$\sigma^2 = E_F[\psi(\mathbf{X};\theta)^2]/E_F[\psi'(\mathbf{X};\theta)]^2 = E_F[IF(\mathbf{X};TF)^2]$$

- NOTE: Proving Frèchet differentiability is not necessary.

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

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

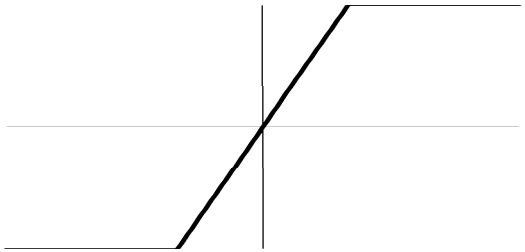
$$\Rightarrow \quad \widehat{\theta} = \frac{\Sigma_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{w_i}\ \mathbf{x_i}}{\Sigma_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{w_i}}$$

The weights are determined by the data cloud.

SOME COMMON M-ESTIMATES OF LOCATION

Huber's M-estimate

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

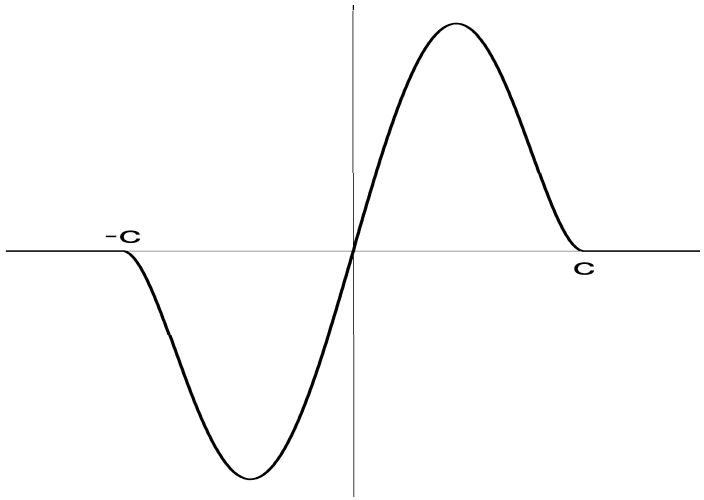


- Given bound on the GES, it has maximum efficiency at the normal model.
- MLE for the "*least favorable distribution*", i.e. symmetric unimodal model with smallest Fisher Information within a "neighborhood" of the normal.

  LFD = Normal in the middle and double exponential in the tails.

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

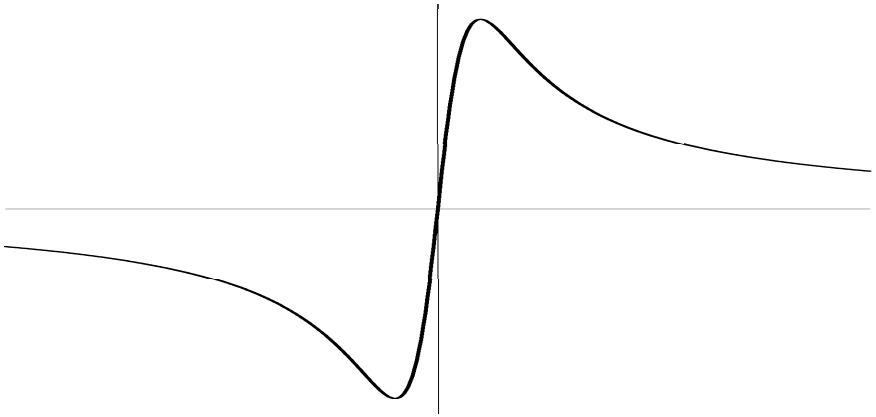
$$u(r) = \left\{ \left(1 - \frac{r^2}{c^2} \right)_+ \right\}^2 \quad \text{where} \quad a_+ = \max\{0, a\}$$



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

CAUCHY MLE

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



NOT A STRONG REDESCENDER.

COMPUTATIONS

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

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

where $w_{i,k} = u(x_i - \hat{\theta}_k)$ and $\hat{\theta}_o$ is any intial value.

- **Re-weighted mean = One-step M-estimate**

$$\hat{\theta}_1 = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{w_{i,o}}\ \mathbf{x_i}}{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{w_{i,o}}},$$

where $\hat{\theta}_o$ is a preliminary robust estimate of location, such as the median.

PROOF OF CONVERGENCE

Sketch

$$\hat{\theta}_{k+1} - \hat{\theta}_k = \frac{\Sigma_{i=1}^n w_{i,k} x_i}{\Sigma_{i=1}^n w_{i,k}} - \hat{\theta}_k = \frac{\Sigma_{i=1}^n w_{i,k}(x_i - \hat{\theta}_k)}{\Sigma_{i=1}^n w_{i,k}} = \frac{\Sigma_{i=1}^n \psi(x_i - \hat{\theta}_k)}{\Sigma_{i=1}^n u(x_i - \hat{\theta}_k)}$$

Note: If $\hat{\theta}_k > \hat{\theta}$, then $\hat{\theta}_k > \hat{\theta}_{k+1}$ and $\hat{\theta}_k < \hat{\theta}$, then $\hat{\theta}_k < \hat{\theta}_{k+1}$.

Decreasing objective function

Let $\rho(r) = \rho_o(r^2)$ and suppose $\rho_o'(s) \geq 0$ and $\rho_o''(s) \leq 0$. Then

$$\sum_{i=1}^n \rho(x_i - \hat{\theta}_{k+1}) < \sum_{i=1}^n \rho(x_i - \hat{\theta}_k).$$

- **Examples:**
  - Mean $\Rightarrow \rho_o(s) = s$
  - Median $\Rightarrow \rho_o(s) = \sqrt{s}$
  - Cauchy MLE $\Rightarrow \rho_o(s) = \log(1 + s)$.
- Include redescending M-estimates of location.
- Generalization of EM algorithm for mixture of normals.

PROOF OF MONOTONE CONVERGENCE

Sketch

Let $r_{i,k} = (x_i - \hat{\theta}_k)$. By Taylor series with remainder term,

$$\rho_o(r^2_{i,k+1}) = \rho_o(r^2_{i,k}) + (r^2_{i,k+1} - r^2_{i,k})\rho'_o(r^2_{i,k}) + \frac{1}{2}(r^2_{i,k+1} - r^2_{i,k})^2\rho''_o(r^2_{i,*})$$

So,

$$\sum_{i=1}^{n} \rho_o(r^2_{i,k+1}) < \sum_{i=1}^{n} \rho_o(r^2_{i,k}) + \sum_{i=1}^{n}(r^2_{i,k+1} - r^2_{i,k})\rho'_o(r^2_{i,k})$$

Now,

$$ru(r) = \psi(r) = \rho'(r) = 2r\rho'_o(r^2) \quad \text{and so} \quad \rho'_o(r^2) = u(r)/2.$$

Also,

$$(r^2_{i,k+1} - r^2_{i,k}) = (\hat{\theta}_{k+1} - \hat{\theta}_k)^2 - 2(\hat{\theta}_{k+1} - \hat{\theta}_k)(x_i - \hat{\theta}_k).$$

Thus,

$$\sum_{i=1}^{n} \rho_o(r^2_{i,k+1}) < \sum_{i=1}^{n} \rho_o(r^2_{i,k}) - \frac{1}{2}(\hat{\theta}_{k+1} - \hat{\theta}_k)^2 \sum_{i=1}^{n} u(r_{i,k}) < \sum_{i=1}^{n} \rho_o(r^2_{i,k})$$

- **Slow but Sure** $\Rightarrow$ Switch to Newton-Raphson after a few iterations.

PART 2

MORE ADVANCED CONCEPTS AND METHODS

SCALE EQUIVARIANCE

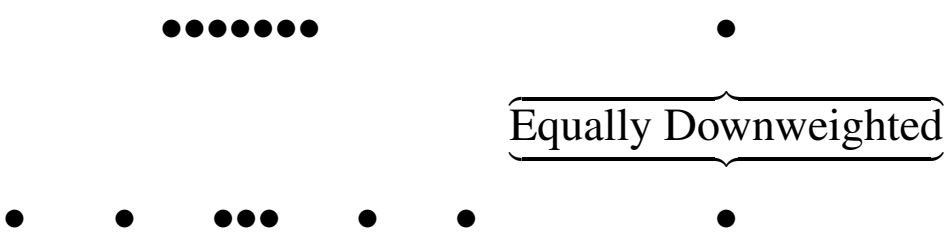
- M-estimates of location alone are not scale equivariant in general, i.e.

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

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

(*Exceptions: the mean and median.*)

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



SCALE STATISTICS: s_n

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

- Sample standard deviation.

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

- MAD (or more approriately MADAM).

Median Absolute Deviation About the Median

$$s_n^* = Median \mid x_i - median \mid$$

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

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

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

or

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

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

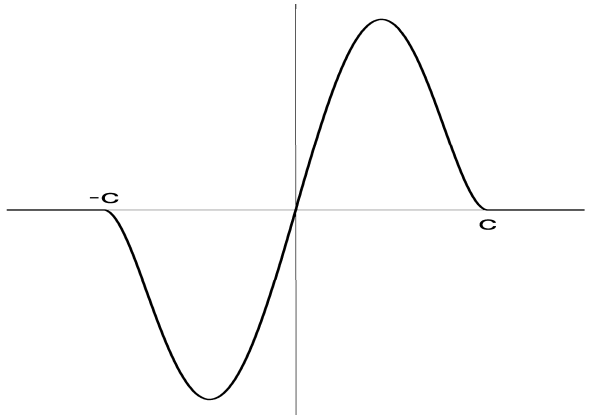
Tuning an M-estimate

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

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

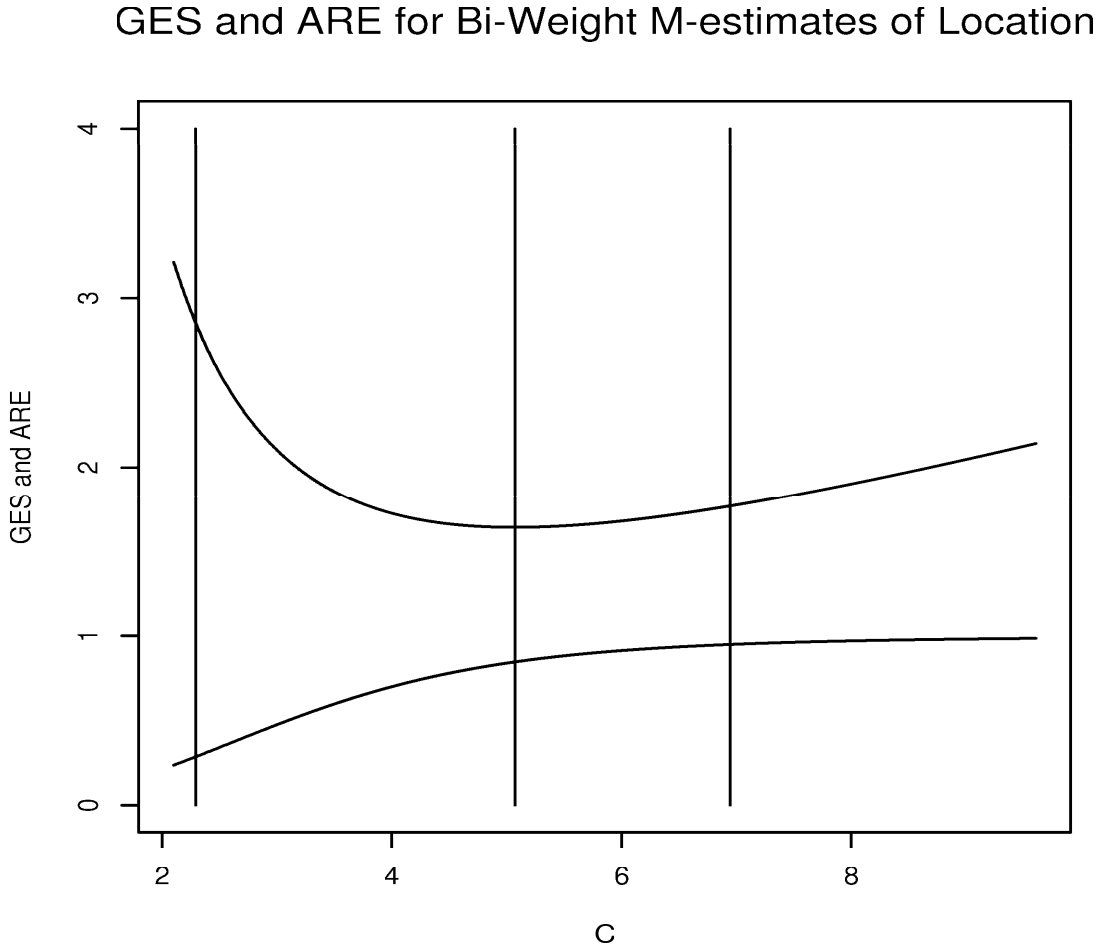
- Tukey's Bi-weight.

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



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

Normal distribution. Auxiliary scale.



MORE THAN ONE OUTLIER

- GES: measure of local robustness.
- Measures of global robustness?

$$\mathcal{X}_n = \{X_1, \ldots, X_n\}\text{: n "good" data points}$$

$$\mathcal{Y}_m = \{Y_1, \ldots, Y_m\}\text{: m "bad" data points}$$

$$\mathcal{Z}_{n+m} = \mathcal{X}_n \cup \mathcal{Y}_m\text{: } \epsilon_m\text{-contaminated sample. } \epsilon_m = \frac{m}{n+m}$$

$$\text{Bias of a statistic: } \mid T(\mathcal{X}_n \cup \mathcal{Y}_m) - T(\mathcal{X}_n) \mid$$

MAX-BIAS and THE BREAKDOWN POINT

- Max-Bias under ϵ_m-contamination:

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

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

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

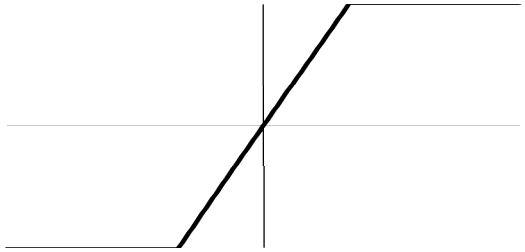
- Other concepts of breakdown (e.g. replacement).
- The definition of BIAS can be modified so that BIAS $\to \infty$ as $T(\mathcal{X}_n \cup \mathcal{Y}_m)$ goes to the boundary of the parameter space.

  Example: If T represents scale, then choose

$$BIAS = \mid \log\{T(\mathcal{X}_n \cup \mathcal{Y}_m)\} - \log\{T(\mathcal{X}_n)\} \mid.$$

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

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

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

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

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

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

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:

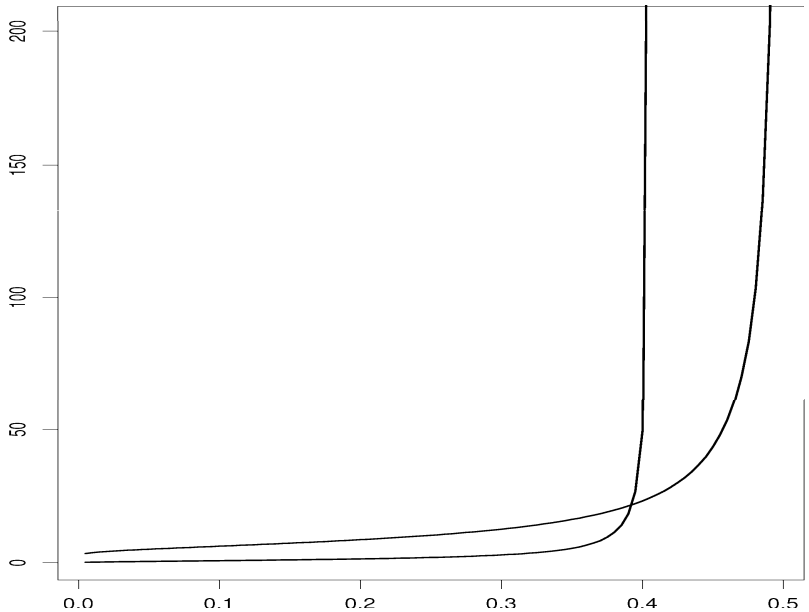
$$B(\epsilon; T, F) = \sup_H \mid T(F_\epsilon) - T(F) \mid$$

- 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



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

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?** $\Rightarrow$ $\epsilon \leq 1/2$**?**

- **Discussion: Davies and Gather (2005),** ***Annals of Statistics*****.**

Example: *Redescending M-estimates of location with fixed scale*

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

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

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

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:**

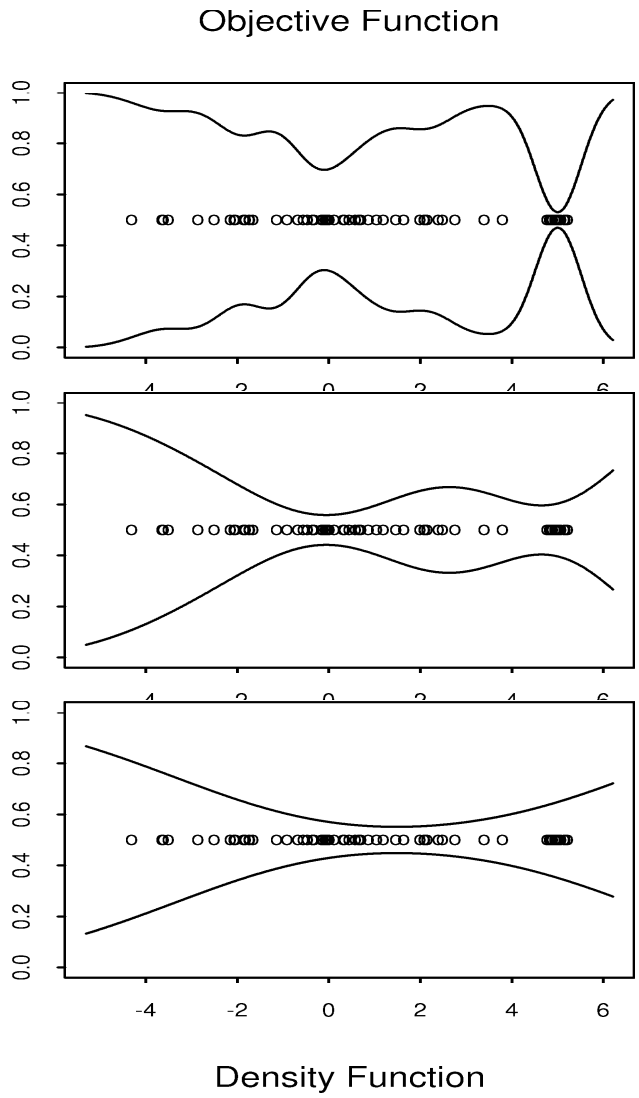
$$\widehat{f}(x) \propto \sum_{i=1}^{n} \kappa\left(\frac{x_i - t}{h}\right)$$

- **Relationship:** $\kappa \propto 1 - \rho$ and c = 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 — 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 AND MULTIVARIATE STATISTICS

REGRESSION SETTING

- Data: $(Y_i, \mathbf{X}_i)\ i = 1, \ldots, n$
  - $Y_i \in \Re$ Response
  - $\mathbf{X}_i \in \Re^p$ 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
  - Generalizion 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 $w_i = u(r_i(\widehat{\boldsymbol{\beta}}))$

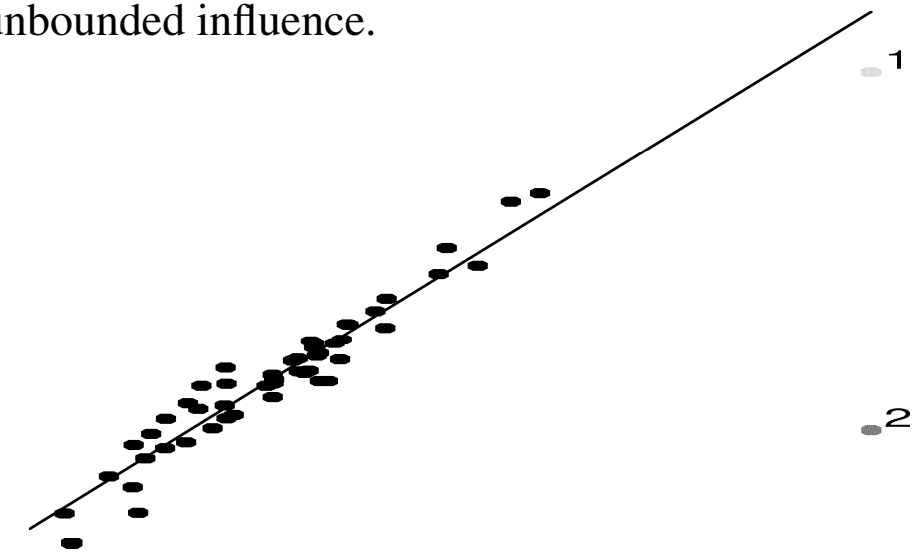
$$\widehat{\boldsymbol{\beta}} = [\sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i']^{-1} [\sum_{i=1}^{n} w_i y_i \mathbf{x}_i]$$

- **Computations via IRLS algorithm.**

INFLUENCE FUNCTIONS FOR M-ESTIMATES OF REGRESSION

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

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



- 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*):

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

    Downweights outlying design points (leverage points), even if they are good leverage points.
  - General form (*c.g. Maronna and Yohai, 1981*):

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

- **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

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

  i.e.

$$\widehat{Y}_i \to a\widehat{Y}_i$$

- **Open problem:** Is high breakdown point equivariate regression possible?

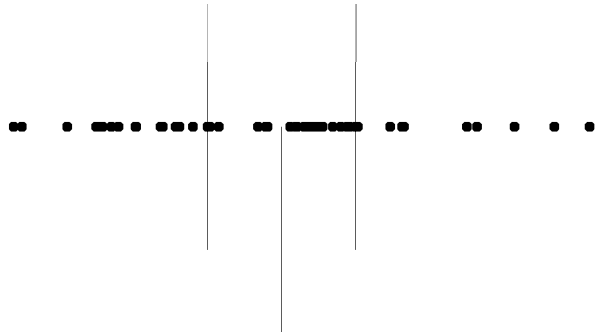
- **Yes.** Least Median of Squares. *Hampel (1984), Rousseeuw, (1984)*

Least Median of Squares (LMS)

Hampel (1984), Rousseeuw, (1984)

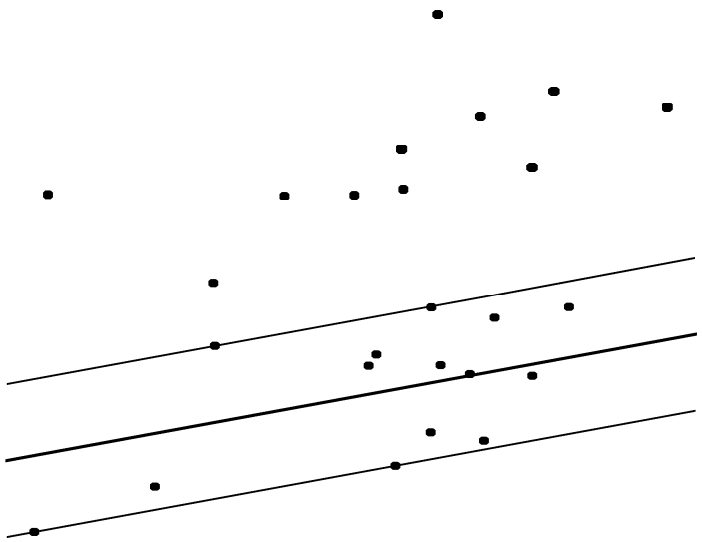
$$\min_{\boldsymbol{\beta}} \boldsymbol{Median}\{\boldsymbol{r}_i(\boldsymbol{\beta})^2 \mid \boldsymbol{i} = \mathbf{1}, \ldots, \boldsymbol{n}\}$$

- Alternatively: $\min_{\boldsymbol{\beta}} MAD\{\, r_i(\boldsymbol{\beta}) \,\}$
- 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

$$\sqrt{n}|| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} || \rightarrow_p \infty$$

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

- **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): $\quad \mathbf{min}_{\boldsymbol{\beta}}\, \boldsymbol{S}(\boldsymbol{r_i}(\boldsymbol{\beta}))$
- $S = MAD \;\Rightarrow\;$ LMS
- $S =$ sample standard deviation about 0 $\;\Rightarrow\;$ Least Squares
- Bounded monotonic M-estimates of scale (about zero):

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

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

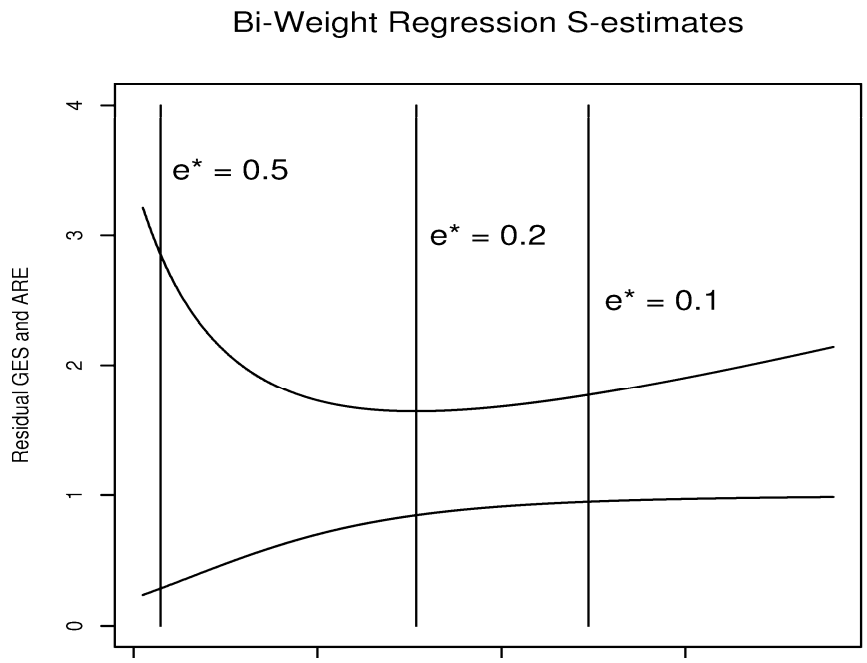
$$\frac{1}{n} \sum_{i=1}^{n} \rho\left(\mid r_i \mid /s\right) = \epsilon$$

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

- For LMS: ρ : 0 - 1 jump function and $\epsilon = 1/2$
- Breakdown point: $\quad \boldsymbol{\epsilon^* = \mathbf{min}(\epsilon, 1-\epsilon)}$

S-estimates of Regression

- $\sqrt{n}$- consistent and Asymptotically normal.
- Trade-off: Higher breakdown point $\Rightarrow$ Lower efficiency **and** higher residual gross error sensivity for Normal errors.



- One Resolution: One-step M-estimate via IRLS. (Note: R, S-plus, Matlab.)

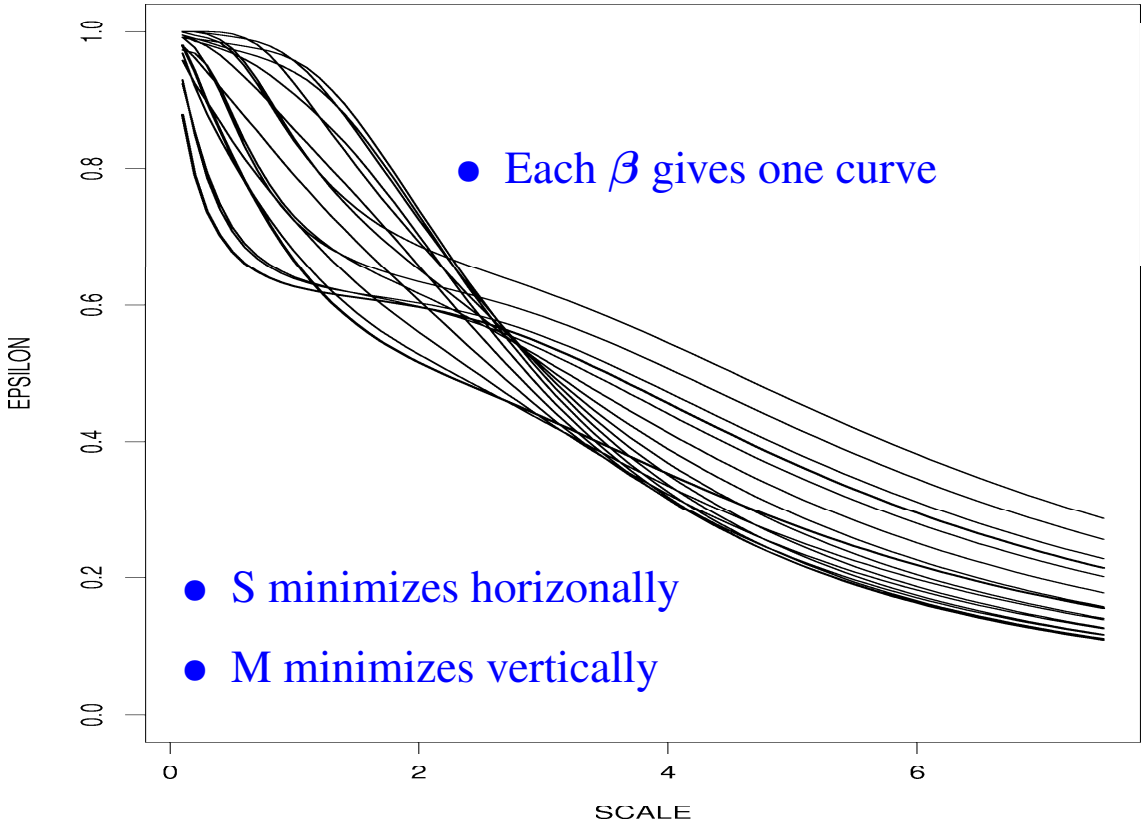
M-ESTIMATES OF REGRESSION WITH GENERAL SCALE

Martin, Yohai, and Zamar (1989)

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

- Parameter of interest: $\boldsymbol{\beta}$
- Scale statistic: $\boldsymbol{s_n}$ (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).
  - **CM-estimates,** *Mendes and Tyler* (1994).

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



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.

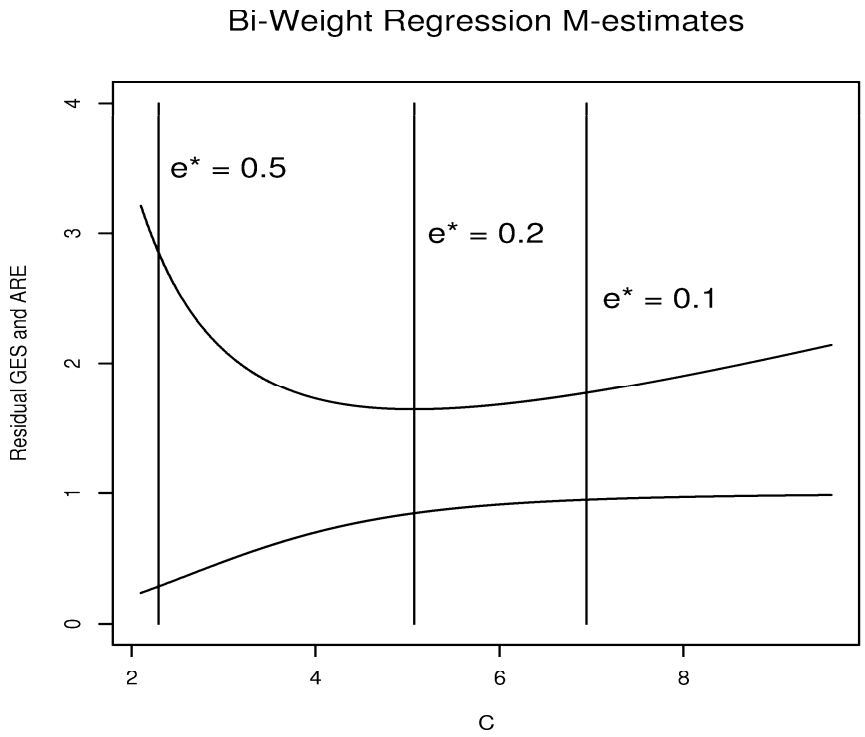
- Compute a monotone M-estimate of scale about zero for the residuals.
  s_n: ***Usually from the S-estimate.***

- 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



- 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

COMPUTATIONAL ISSUES

- All known high breakdown point regression estimates are computationally intensive. *Non-convex optimization problem.*

- Approximate or stochastic algorithms.

- Random Elemental Subset Selection. *Rousseeuw* (1984).
  - Consider exact fit of lines for p of the data points $\Rightarrow \begin{pmatrix} n \\ p \end{pmatrix}$ such lines.
  - Optimize the criterion over such lines.
  - For large n and p, randomly sample from such lines so that there is a good chance, say e.g. 95% chance, that one of the elemental subsets will contain only good data even if half the data is contaminated.

MULTIVARIATE DATA

Robust Multivariate Location and Covariance Estimates

Parallel developments

- **Multivariate M-estimates.** *Maronna* (1976). *Huber* (1977).
  - Adaptively weighted mean and covariances.
  - IRLS algorithms.
  - Breakdown point: $\epsilon^* \leq 1/(d+1)$, where d 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 CM-estimates.** *Kent and Tyler* (1996).

- **Multivariate MM-estimates.** *Tatsuoka and Tyler* (2000). *Tyler* (2002).

MULTIVARIATE DATA

- **d-dimensional data set:** $\mathbf{X_1, \ldots X_n}$

- **Classical summary statistics:**

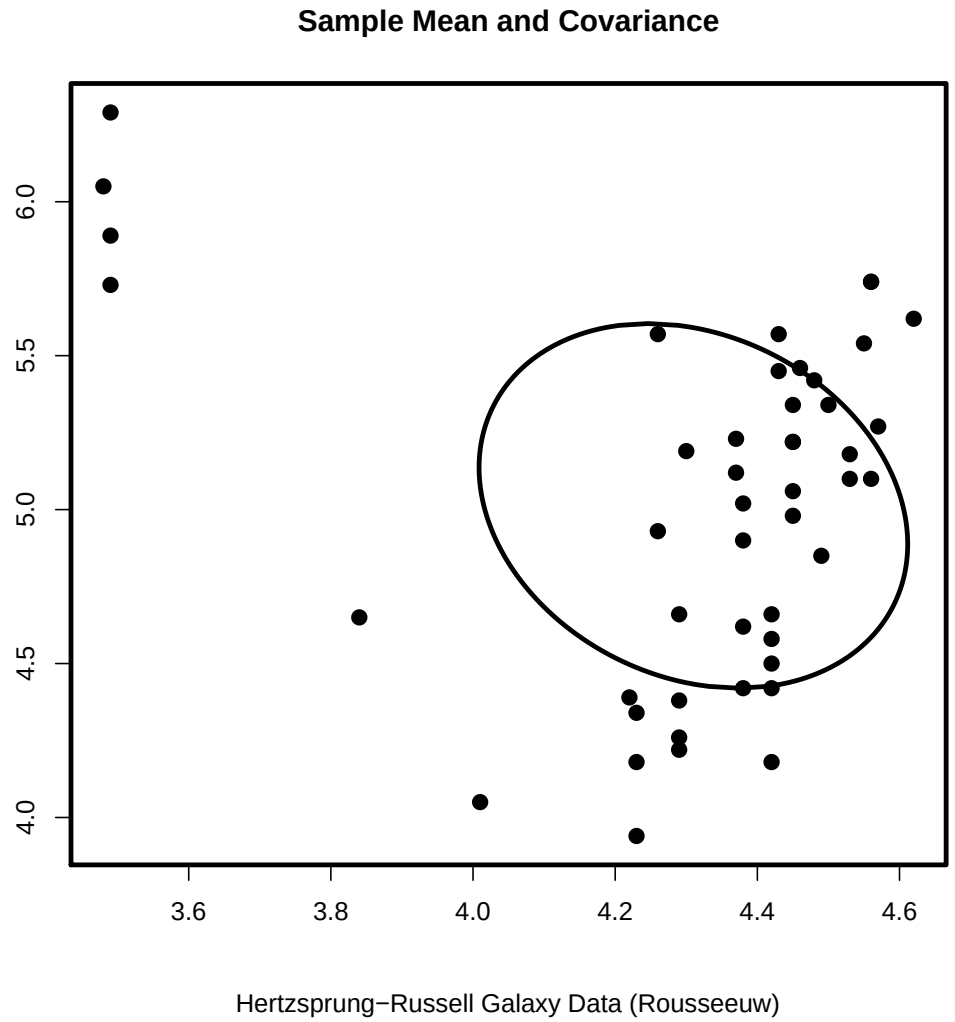
  – **Sample Mean Vector:** $\overline{\mathbf{X}} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i}$

  – **Sample Variance-Covarinance Matrix**

$$\mathbf{S_n} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{X_i} - \overline{\mathbf{X}})(\mathbf{X_i} - \overline{\mathbf{X}})^{\mathbf{T}} = \{\mathbf{s_{ij}}\}$$

  where s_{ii} = sample variance, s_{ij} = sample covariance.

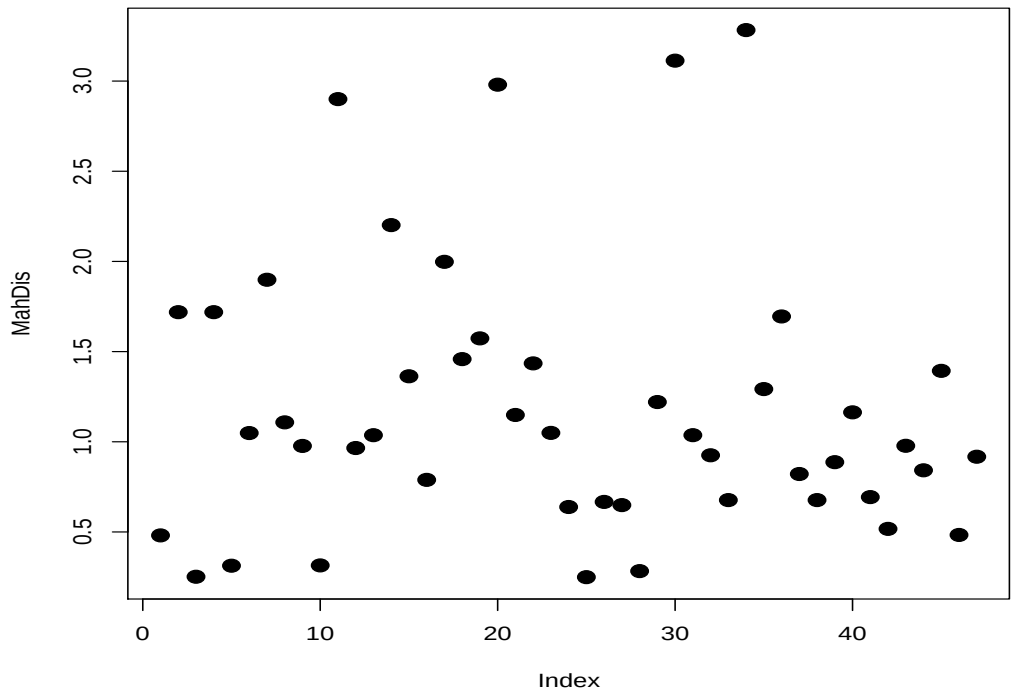
- **Maximum likelihood estimates under normality**



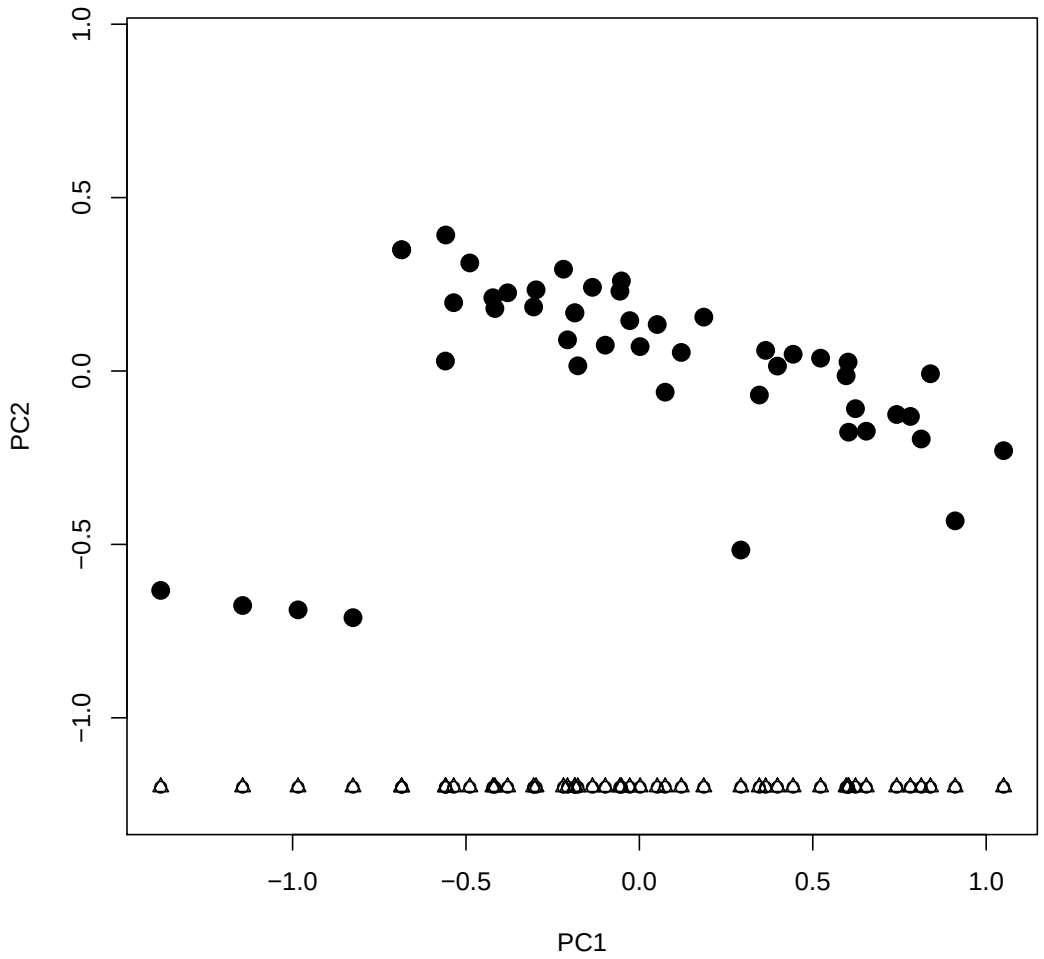
Visualization $\Rightarrow$ Ellipse containing half the data.

$$(\mathbf{X} - \overline{\mathbf{X}})^{\mathbf{T}} \mathbf{S}_{\mathbf{n}}^{-1} (\mathbf{X} - \overline{\mathbf{X}}) \leq \mathbf{c}$$

Mahalanobis distances: $\mathbf{d_i^2 = (X_i - \overline{X})^T S_n^{-1} (X_i - \overline{X})}$



Mahalanobis angles: $\boldsymbol{\theta_{i,j} = Cos^{-1}\left\{(X_i - \overline{X})^T S_n^{-1} (X_j - \overline{X})/(d_i d_j)\right\}}$



Principal Components: Do not always reveal outliers.

ENGINEERING APPLICATIONS

- **Noisy signal arrays, radar clutter.**
  Test for signal, i.e. detector, based on Mahalanobis angle between signal and observed data.

- **Hyperspectral Data.**
  Adaptive cosine estimator = Mahalanobis angle.

- **BIG DATA**

- **Can not clean data via observations. Need automatic methods.**

ROBUST MULTIVARIATE ESTIMATES

Location and Scatter (Pseudo-Covariance)

- **Trimmed versions: complicated**

- **Weighted mean and covariance matrices**

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} w_i \, X_i}{\sum_{i=1}^{n} w_i} \qquad \widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} w_i \, (X_i - \widehat{\mu})(X_i - \widehat{\mu})^T$$

where $w_i = u(d_{i,o})$ **and** $d_{i,o}^2 = (X_i - \widehat{\mu}_o)^T \widehat{\Sigma}_o^{-1} (X_i - \widehat{\mu}_o)$

and with $\widehat{\mu}_o$ **and** $\widehat{\Sigma}_o$ **being initial estimates, e.g.**

$\rightarrow$ ***Sample mean vector and covariance matrix.***

$\rightarrow$ ***High breakdown point estimates.***

MULTIVARIATE M-ESTIMATES

- **MLE type for elliptically symmetric distributions**
- **Adaptively weighted mean and covariances**

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} w_i \, X_i}{\sum_{i=1}^{n} w_i} \qquad \widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} w_i \, (X_i - \widehat{\mu})(X_i - \widehat{\mu})^T$$

- $w_i = u(d_i)$ **and** $d_i^2 = (X_i - \widehat{\mu})^T \widehat{\Sigma}^{-1} (X_i - \widehat{\mu})$
- **Implicit equations**

PROPERTIES OF M-ESTIMATES

- **Root-n consistent and asymptotically normal.**

- **Can be tuned to have desirable properties:**
  - **high efficiency over a broad class of distributions**
  - **smooth and bounded influence function**

- **Computationally simple (IRLS)**

- **Breakdown point:** $\epsilon^* \leq 1/(d+1)$

HIGH BREAKDOWN POINT ESTIMATES

- ***MVE: Minimum Volume Ellipsoid*** $\Rightarrow$ $\epsilon^* = 0.5$

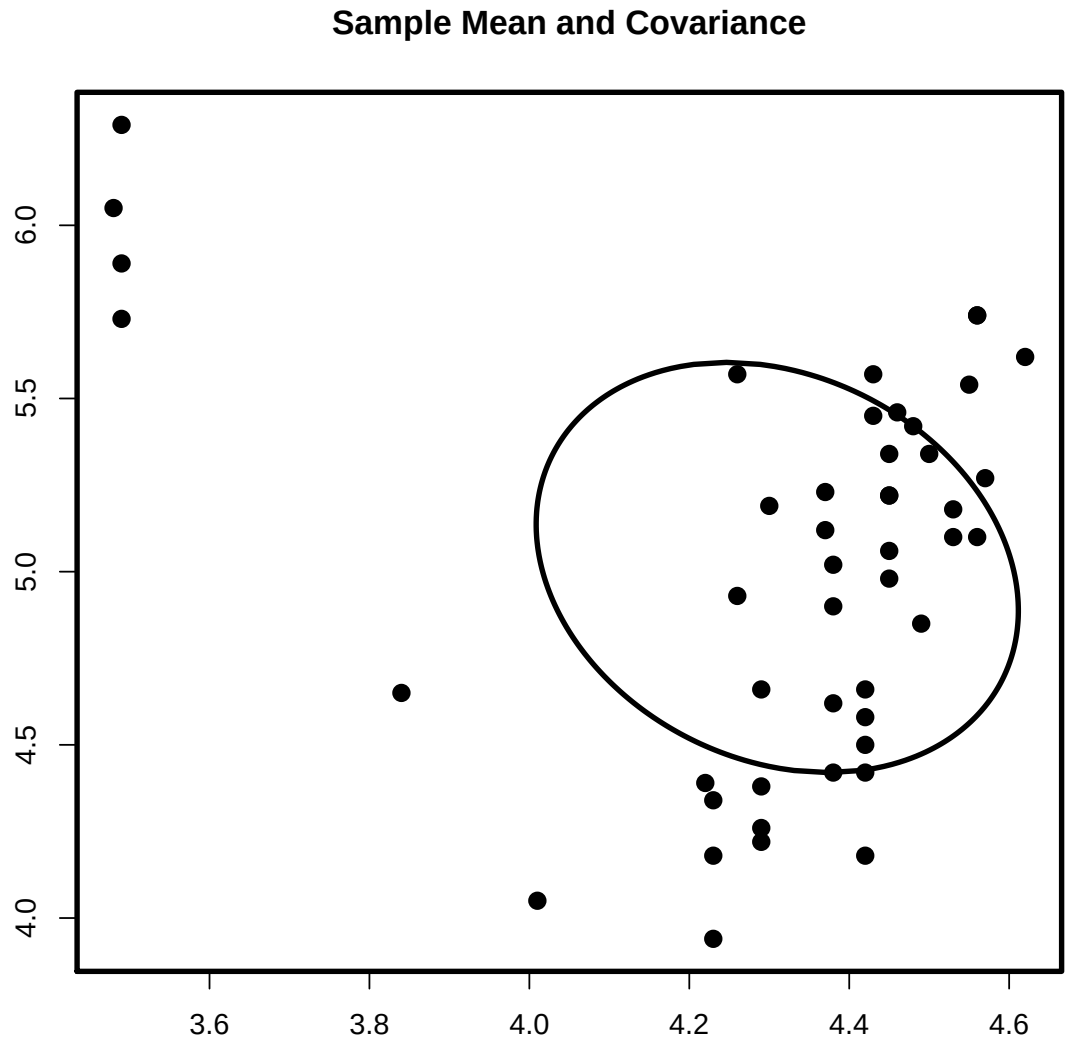
  Center and scatter matrix for smallest ellipse covering at least half of the data.

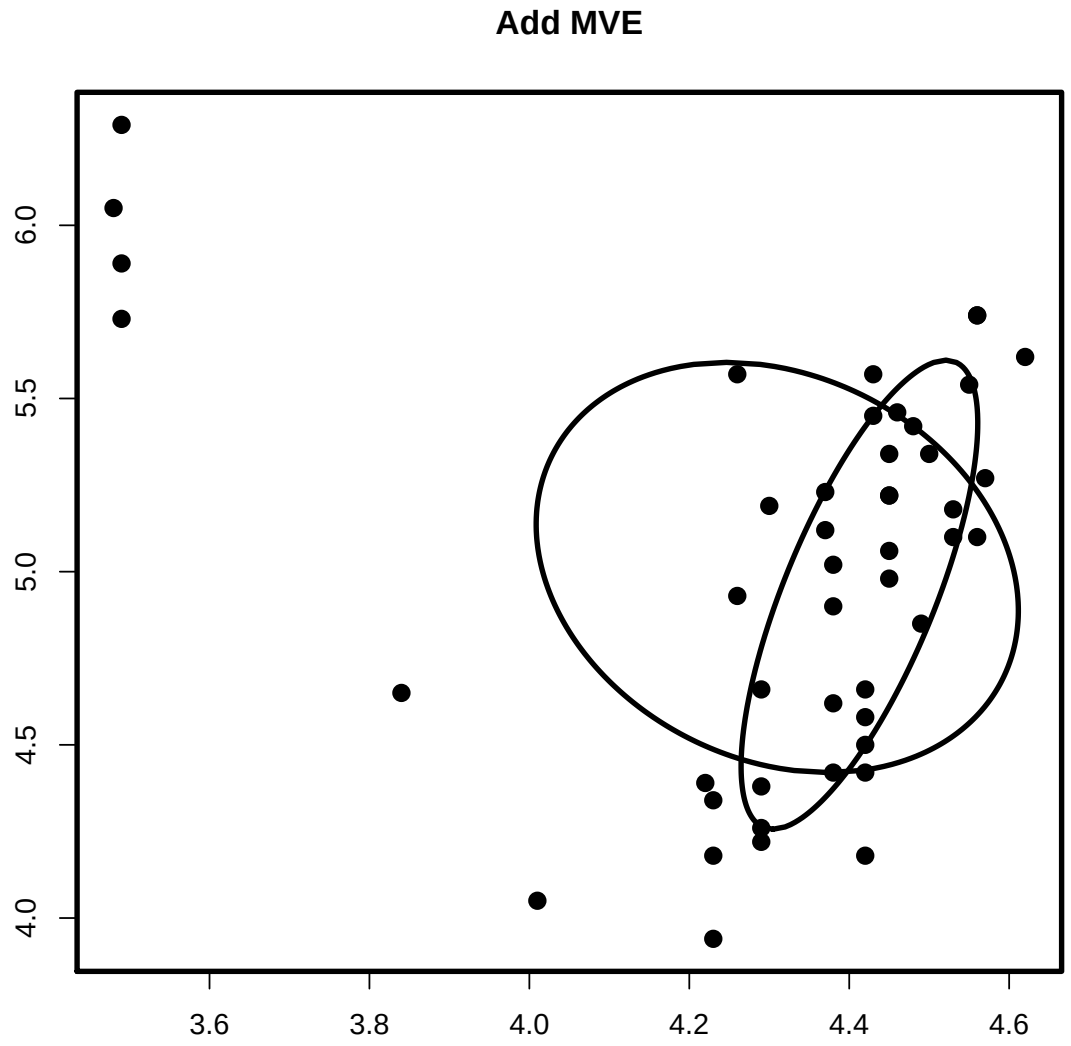
- ***MCD, S-estimates, MM-estimates, CM-estimates*** $\Rightarrow$ $\epsilon^* = 0.5$

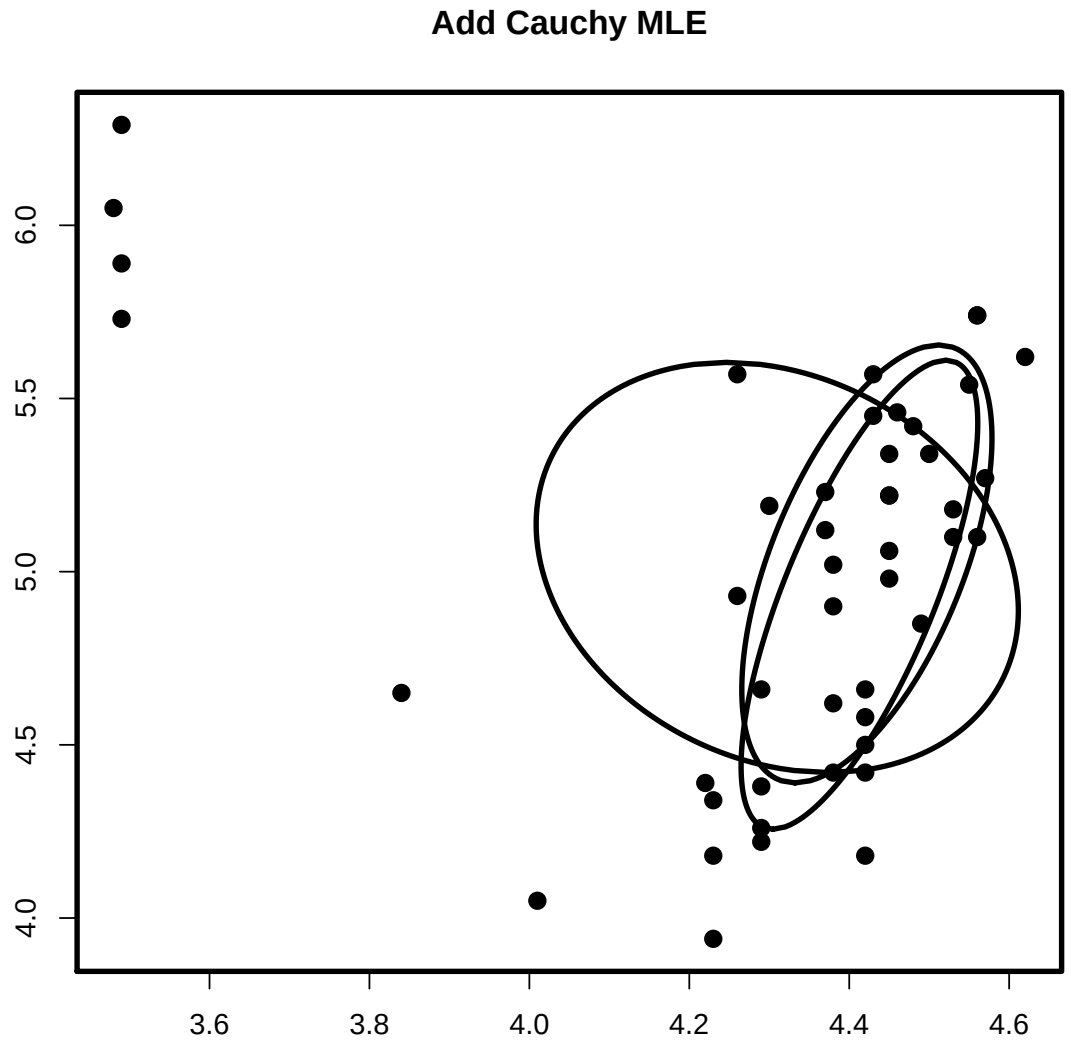
- **Reweighted versions based on high breakdown start.**

- **Computationally intensive:**

  Approximate/probabilistic algorithms.
  Elemental Subset Approach.


Sample Mean and Covariance
6.0
5.5
5.0
4.5
4.0
3.6
3.8
4.0
4.2
4.4
4.6
Hertzsprung−Russell Galaxy Data (Rousseeuw)


Add MVE
6.0
5.5
5.0
4.5
4.0
3.6
3.8
4.0
4.2
4.4
4.6


Add Cauchy MLE
6.0
5.5
5.0
4.5
4.0
3.6
3.8
4.0
4.2
4.4
4.6

ELLIPTICAL DISTRIBUTIONS

and

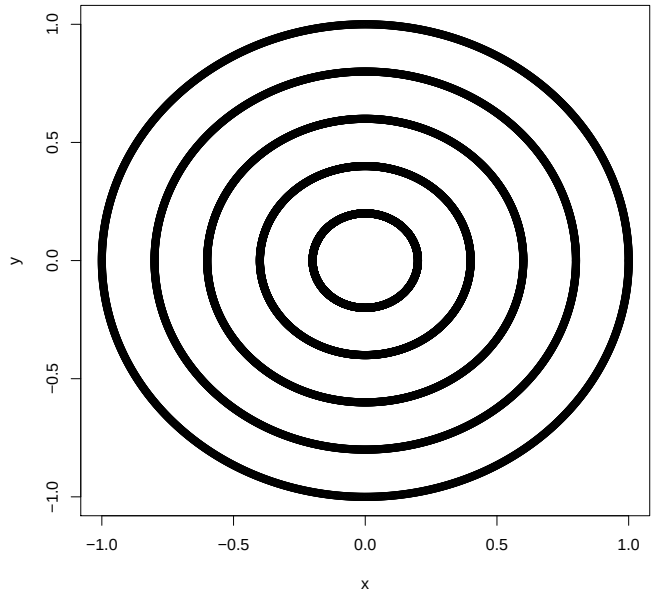
AFFINE EQUIVARIANCE

SPHERICALLY SYMMETRIC DISTRIBUTIONS

$$\boldsymbol{Z = RU} \quad \text{where} \quad \boldsymbol{R \perp U},\ \boldsymbol{R > 0},$$

and $\boldsymbol{U} \sim_{\boldsymbol{d}}$ Uniform on the unit sphere.

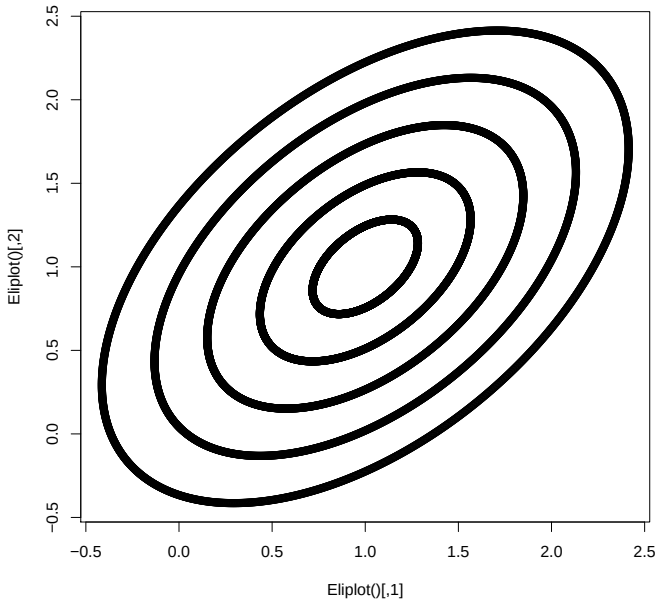
Density: $\boldsymbol{f(z)} = \mathrm{g}(\boldsymbol{z'z})$



ELLIPTICAL SYMMETRIC DISTRIBUTIONS

$$\boldsymbol{X} = \boldsymbol{AZ} + \boldsymbol{\mu} \sim \boldsymbol{\mathcal{E}}(\boldsymbol{\mu}, \boldsymbol{\Gamma}; \boldsymbol{g}), \text{ where } \boldsymbol{\Gamma} = \boldsymbol{AA'}$$

$$\Rightarrow \boldsymbol{f}(\boldsymbol{x}) = |\boldsymbol{\Gamma}|^{-1/2} \mathrm{g}\left((\mathrm{x} - \boldsymbol{\mu})' \boldsymbol{\Gamma}^{-1} (\mathrm{x} - \boldsymbol{\mu})\right)$$



AFFINE EQUIVARIANCE

- Parameters of elliptical distributions:

$$\boldsymbol{X} \sim \mathcal{E}(\mu, \Gamma; g) \Rightarrow \boldsymbol{X}^* = \boldsymbol{B}\mathbf{X} + \boldsymbol{b} \sim \mathcal{E}(\mu^*, \Gamma^*; g)$$
$$\text{where} \quad \mu^* = \boldsymbol{B}\mu + \boldsymbol{b} \quad \text{and} \quad \Gamma^* = \boldsymbol{B}\Gamma\boldsymbol{B}'$$

- Sample version:

$$\mathbf{X}_i \rightarrow \boldsymbol{X}_i^* = \boldsymbol{B}\mathbf{X} + \boldsymbol{b},\ i = 1, \ldots n$$
$$\Rightarrow \quad \widehat{\mu} \rightarrow \widehat{\mu}^* = \boldsymbol{B}\widehat{\mu} + \boldsymbol{b} \quad \text{and} \quad \widehat{V} \rightarrow \widehat{V}^* = \boldsymbol{B}\widehat{V}\boldsymbol{B}'$$

- Examples
  - Mean vector and Covariance matrix
  - M-estimates
  - MVE

ELLIPTICAL SYMMETRIC DISTRIBUTIONS

$$f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Gamma}) = |\boldsymbol{\Gamma}|^{-1/2} g\left((\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Gamma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

- If second moments exist, shape matrix $\boldsymbol{\Gamma} \propto$ covariance matrix.
- For any affine equivariant scatter functional: $\boldsymbol{\Sigma}(\boldsymbol{F}) \propto \boldsymbol{\Gamma}$
- *That is, any affine equivariant scatter statistic estimates a matrix proportional to the shape matrix* $\boldsymbol{\Gamma}$.

NON-ELLIPTICAL DISTRIBUTIONS

- *Difference scatter matrices estimate different population quantities.*
- Analogy: For non-symmetric distributions:
  population mean $\neq$ population median.

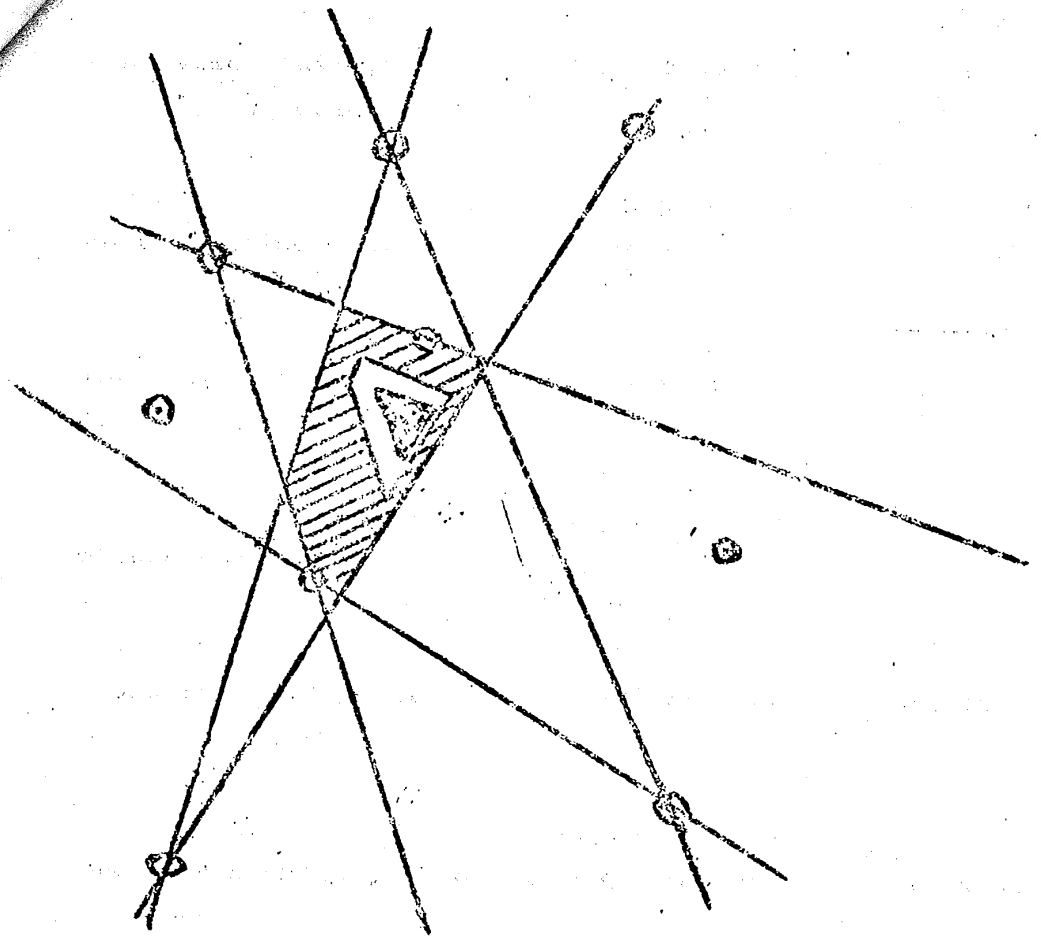
OTHER TOPICS

Projection-based multivariate approaches

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

Exhibit 7 of T6

Depths 3, 3.5, 2.67, and 3.75 relative to the plane for the same nine points



Shaded ring: outside: depth = 3
inside: depth = 3.5

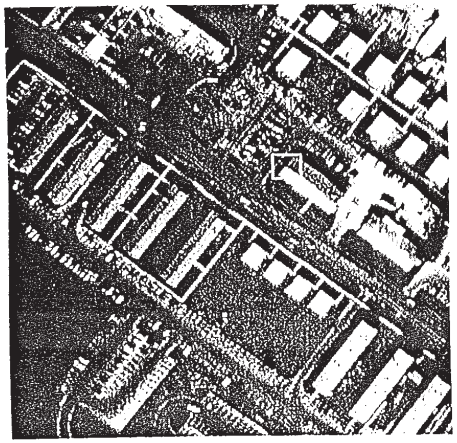
Solid ring: outside: depth = 2.67
inside: depth = 3.75

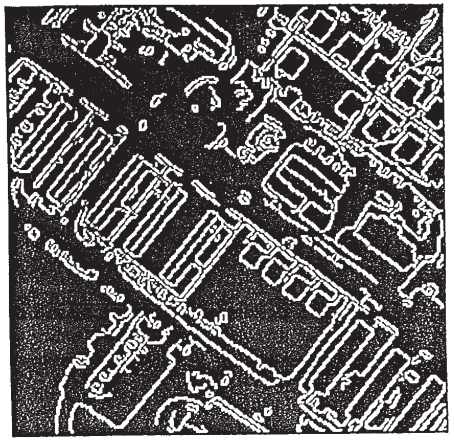
TUKEY'S DEPTH

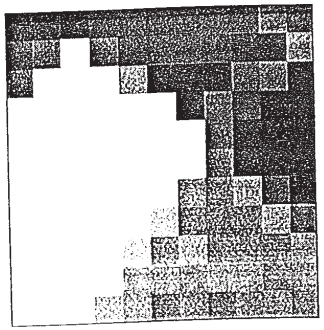
PART 5

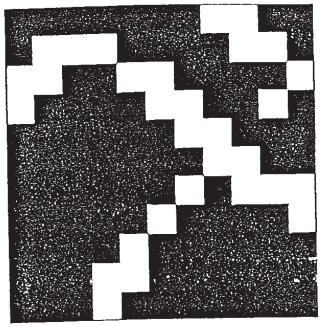
ROBUSTNESS AND COMPUTER VISION

- **Independent development of robust statistics within the computer vision/image understanding community.**









55	55	29	52	64	38	51	58	54	124	90
74	101	255	122	61	55	41	36	27	120	133
102	255	255	255	170	0	8	3	95	125	18
255	255	255	255	255	240	0	105	65	17	0
255	255	255	255	255	255	255	92	9	0	0
255	255	255	255	255	255	255	131	30	0	0
255	255	255	255	255	255	176	152	173	63	0
255	255	255	255	255	214	155	151	148	174	120
255	255	255	255	214	166	190	159	153	147	158
255	255	255	255	169	177	166	172	166	154	157
255	255	255	206	192	162	171	156	165	174	154

Dec. 18, 1962 P. V. C. HOUGH 3,069,654

METHOD AND MEANS FOR RECOGNIZING COMPLEX PATTERNS

Filed March 25, 1960 2 Sheets-Sheet 1

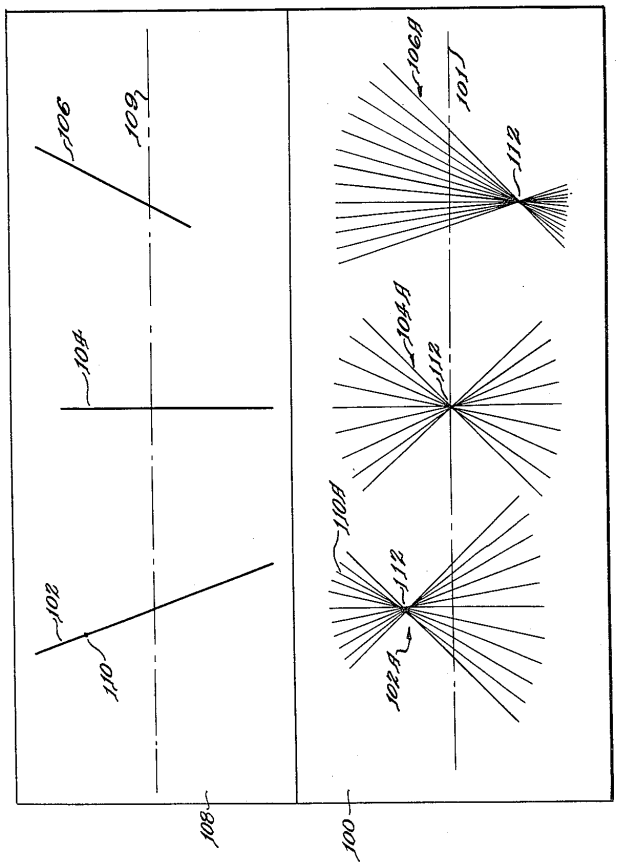


FIG-1

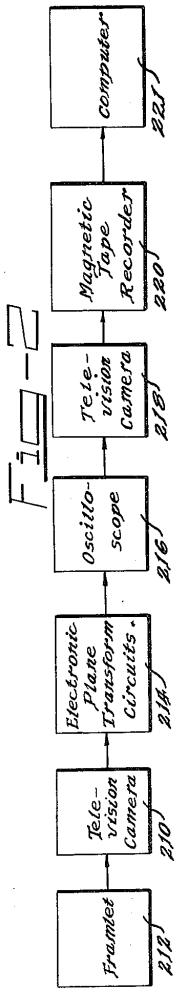


FIG-2

INVENTOR.
Paul V. C. Hough
BY
Roland G. Anderson
Attorney

HOUGH TRANSFORM

(From Maitre, 1985)

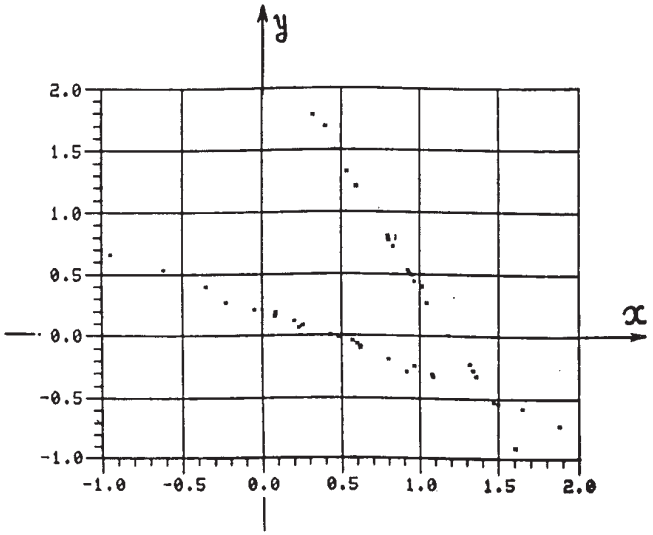


Fig. 3 *a*. — Points de l'espace image issus de deux droites bruitées.

Fig. 3 a. — Image points from two noisy lines.

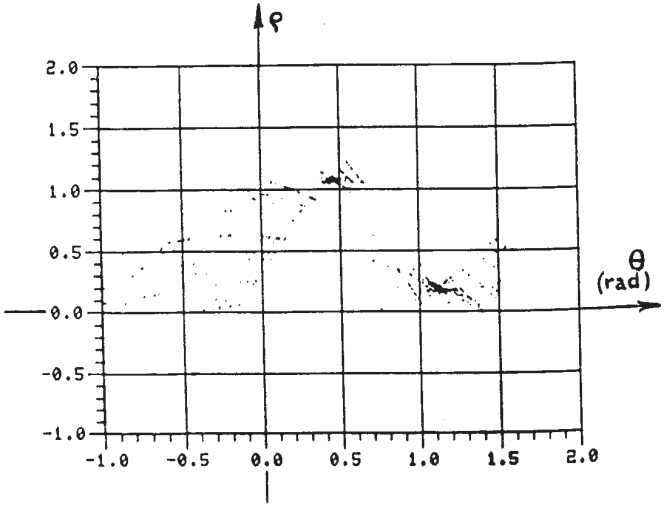


Fig. 3 *b*. — La TH de *m* à 1 associée à la droite en coordonnées normales appliquée à l'ensemble *&* de la figure 3 *a*.

Fig. 3 b. — m-to 1 HT to detect lines in normal coordinates applied to the set & of Fig. 3 a.

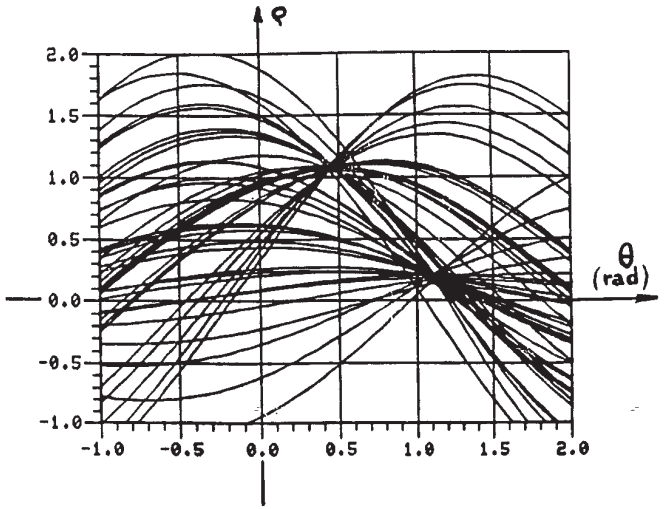


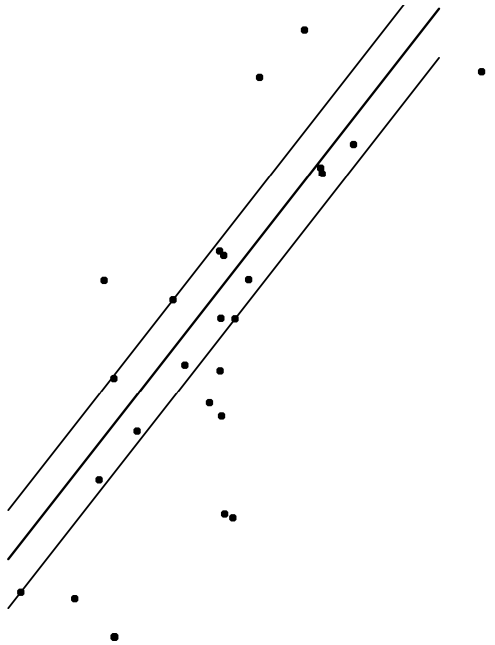
Fig. 3 *c*. — La TH de 1 à *m* appliquée au même ensemble.

Fig. 3 c. — The same with 1 to m HT.

- **Hough transform:** $\rho = xcos(\theta) + ysin(\theta)$
- $y = a + bx$ **where** $a = \rho/sin(\theta)$ **and** $b = -cos(\theta)/sin(\theta)$
- **That is, intersection refers to the line connecting two points.**
- **Hough: Estimation in Data Space to Clustering in Feature Space** ***Find centers of the clusters***
- **Terminology:**
  - **Feature Space = Parameter Space**
  - **Accumulator = Elemental Fit**
- **Computation: RANSAC (Random sample consensus)**
  - **Randomly choose a subset from the Accumulator. (Random elemental fits.)**
  - **Check to see how many data points are within a fixed neighborhood are in a neighborhood.**

Alternative Formulation of Hough Transform/RANSAC

$\Sigma\, \boldsymbol{\rho}(\boldsymbol{r}_i)$ should be small, where $\boldsymbol{\rho}(\boldsymbol{r}) = \begin{cases} \mathbf{1}, & \mid \boldsymbol{r} \mid > \boldsymbol{R} \\ \mathbf{0}, & \mid \boldsymbol{r} \mid \leq \boldsymbol{R} \end{cases}$



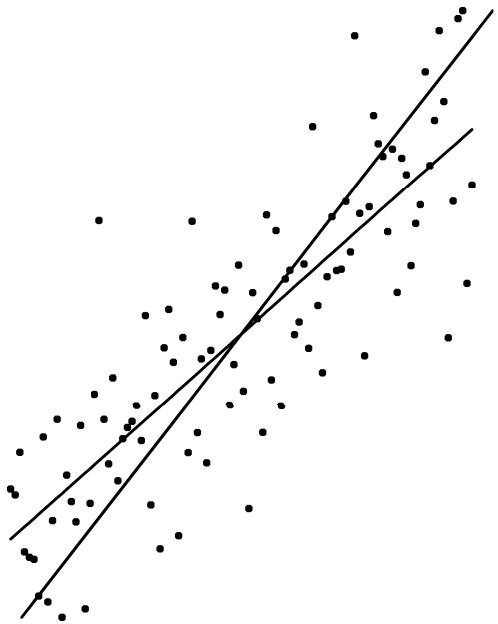
That is, a redescending M-estimate of regression with known scale.
Note: Relationship to LMS.
Vision: Need e.g. 90% breakdown point, i.e. tolerate 90% bad data

Defintion of Residuals?

Hough transform approach does not distinguish between:

- Regression line for regressing $\mathbf{Y}$ on $\mathbf{X}$.
- Regression line for regressing $\mathbf{X}$ on $\mathbf{Y}$.
- Orthogonal regression line.

(Note: Small stochastic errors $\Rightarrow$ little difference.)



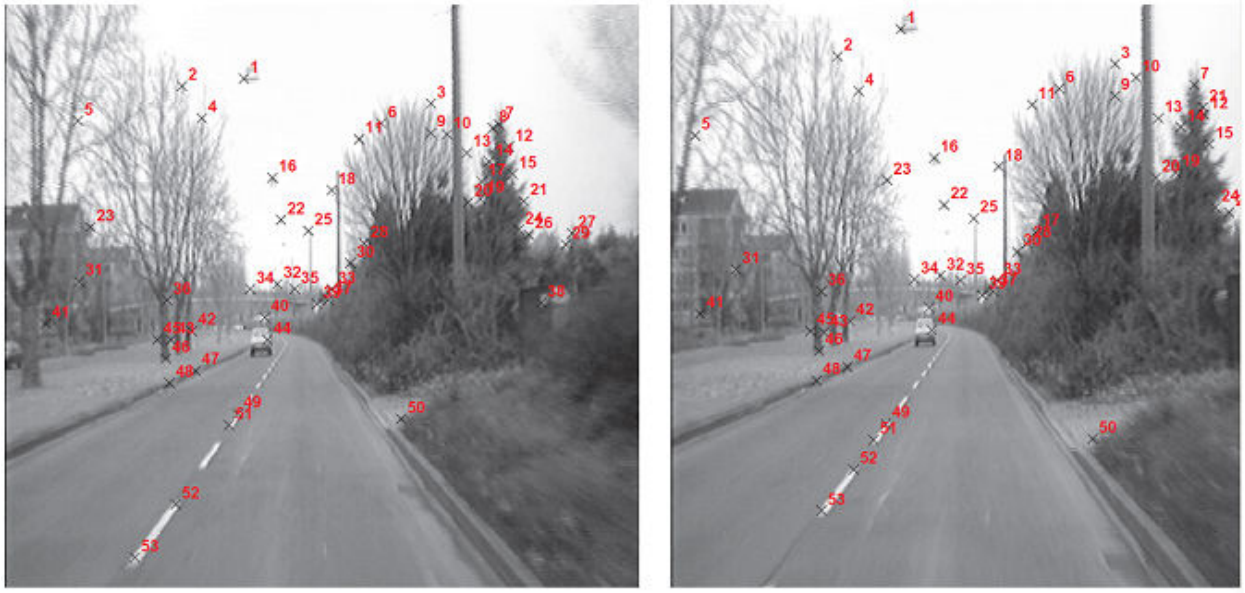
GENERAL PARADIGM

- Line in 2-D: Exact fit to all pairs
- Quadratic in 2-D: Exact fit to all triples

- Conic Sections: Ellipse Fitting

$$(\mathbf{x} - \mu)'\mathbf{A}(\mathbf{x} - \mu) = 1$$

  - Linearizing: Let $\mathbf{x}'_* = (x_1, x_2, x_1^2, x_2^2, x_1x_2, 1)$
  - Ellipse: $\mathbf{a}'\mathbf{x}_* = 0$
  - Exact fit to all subsets of size 5.



Hyperboloid Fitting in 4-D

- Epipolar Geometry: Exact fit to
  - 5 pairs: Calibrated cameras (location, rotation, scale).
  - 8 pairs: Uncalibrated cameras (focal point, image plane).
  - Hyperboloid surface in 4-D.
- Applications: 2D $\rightarrow$ 3D, Motion.
- **OUTLIERS = MISMATCHES**