

Robust Statistics

Partial Overview, General Principles

Werner A. Stahel, ETH Zurich

Fall Semester 2012

1 Introduction

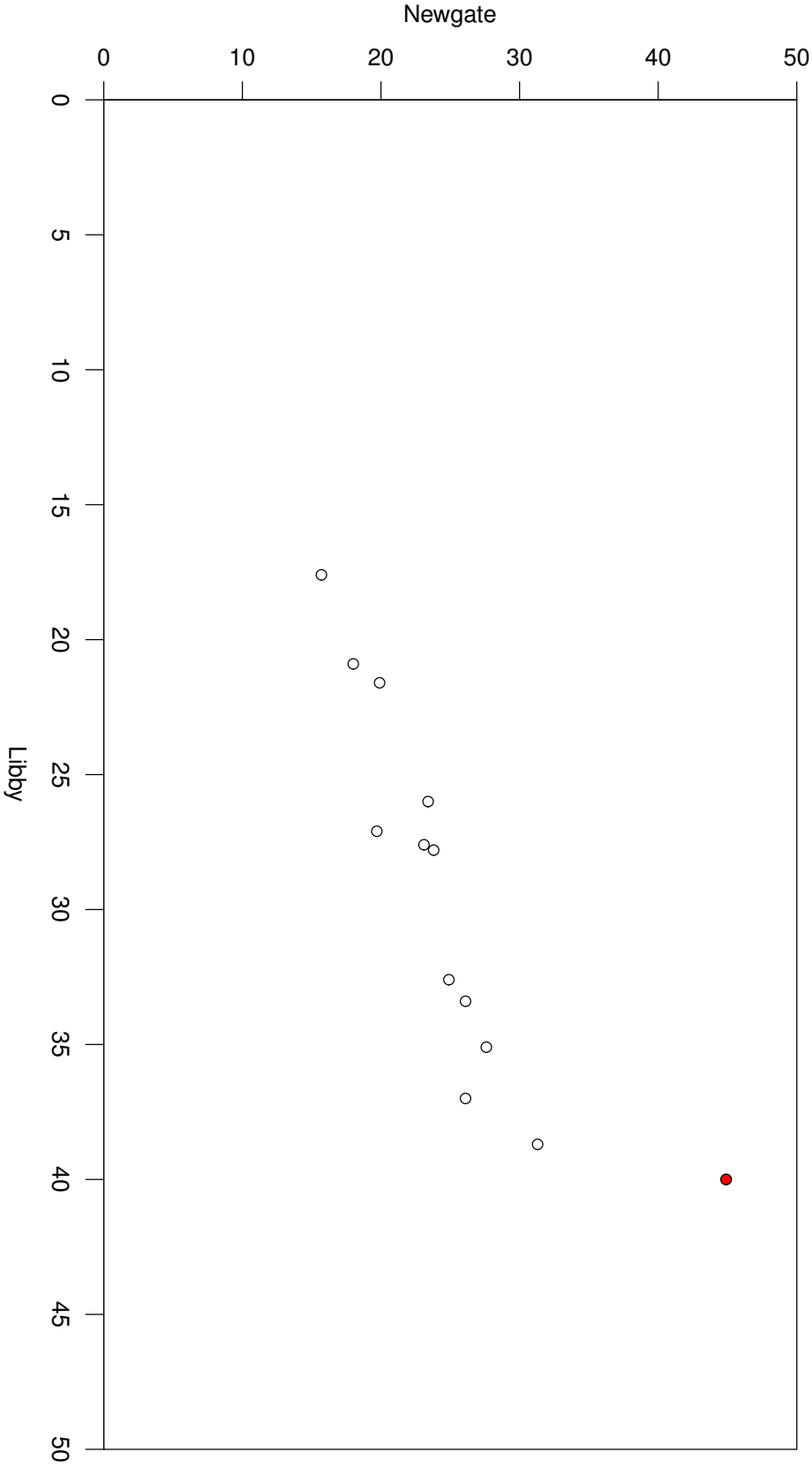
1.1 What is Robust Statistics?

Example

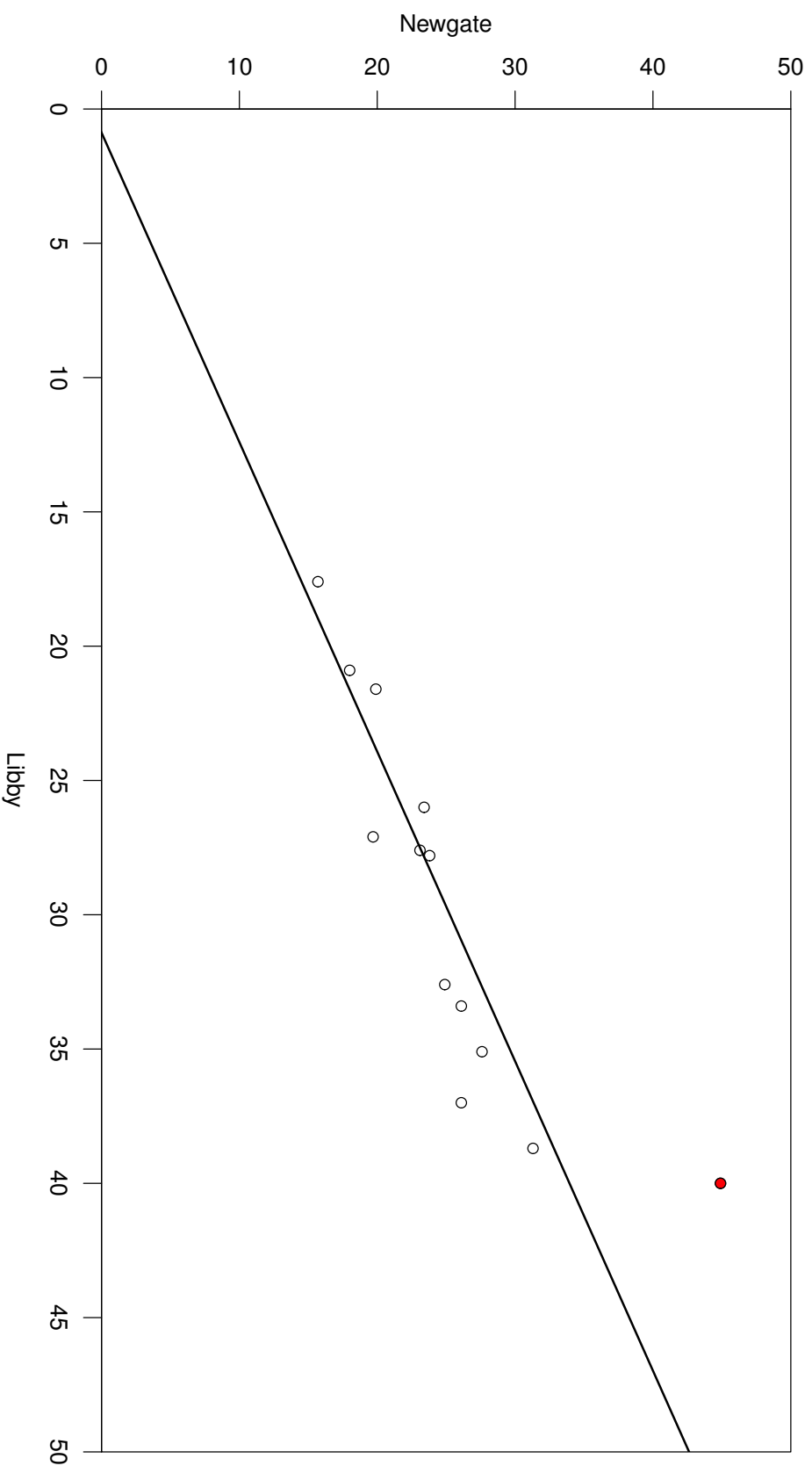
Regression: Water flow of a river at 2 locations.

Simple regression, 13 observations

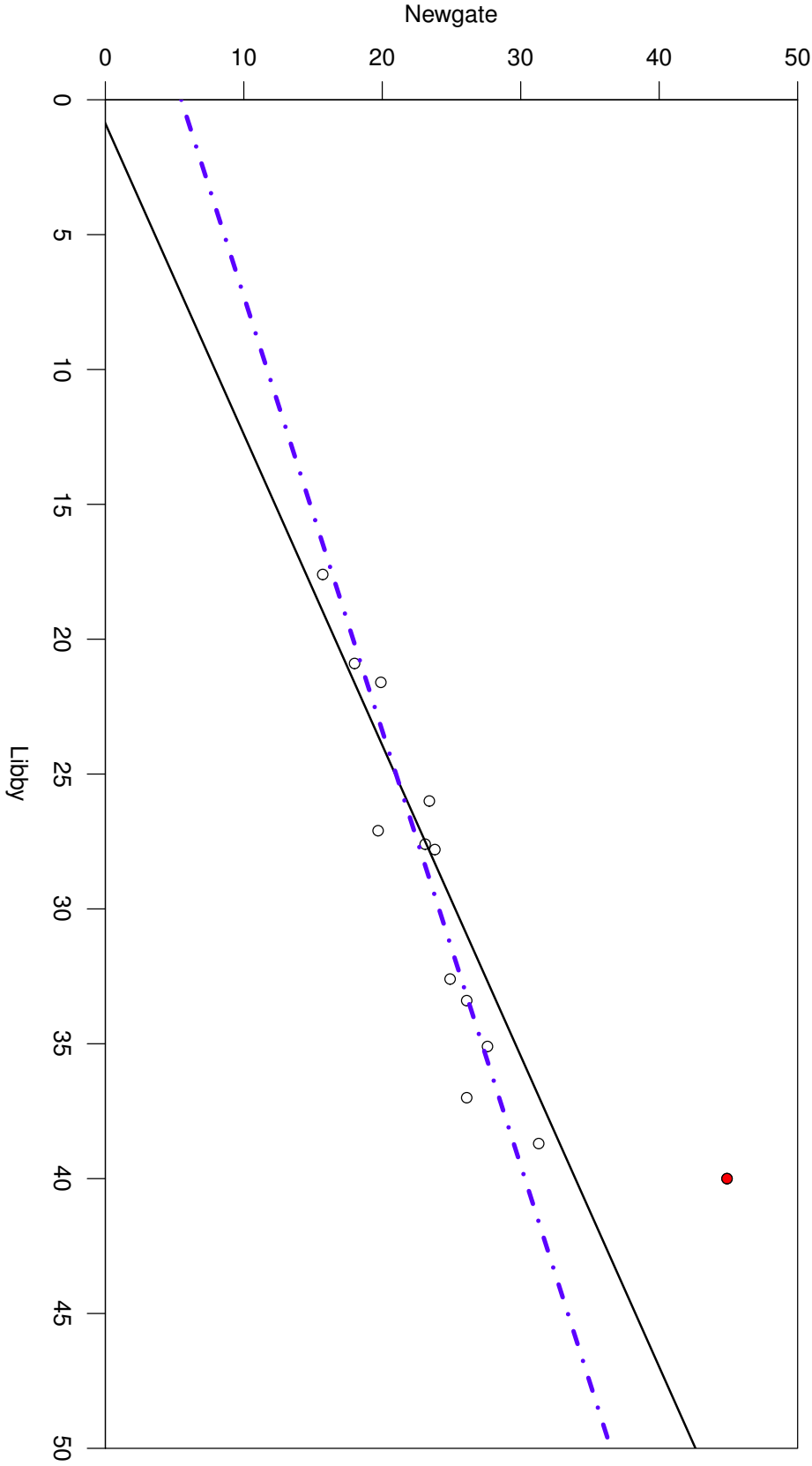
Water Flow



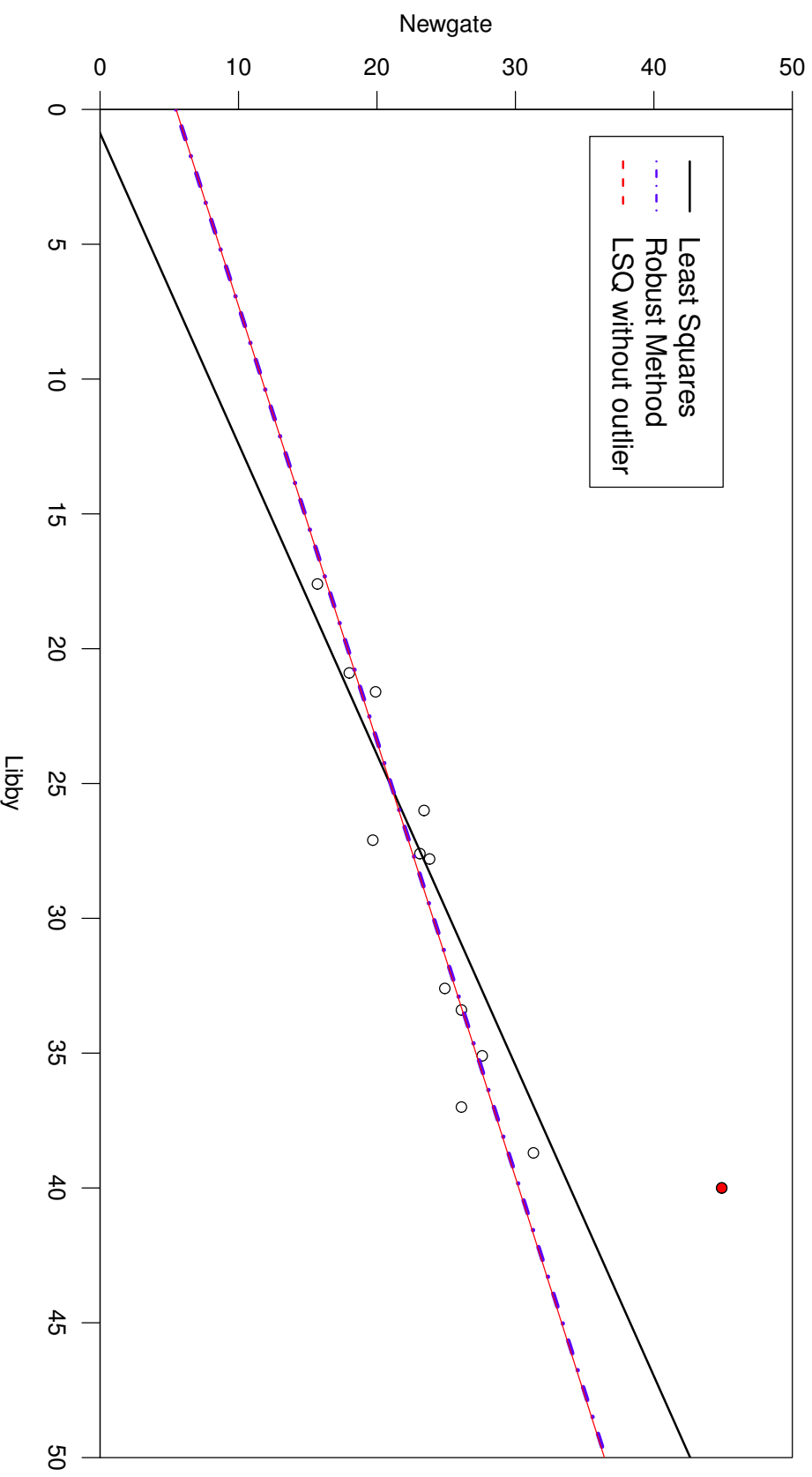
Regression Line (Least Squares)



Robustly Estimated Regression Line



Regression Line after dropping the outlier, Least Squares



Mathematical Statistics:

- Model defines distribution of variables, parameters to be estimated / tested
- Theory determines **optimal** procedures under the conditions defined by the model
- In practice, conditions are not satisfied, but may “hold approximately”.

Robust statistics is the statistics of approximate parametric models.

1.2 History 1

Outliers

Rejection of outliers: Daniel Bernoulli (1770), astronomers

Trimmed mean: 1821:

Drop (k) largest & smallest obs., average the remaining ones.

Rejection rules: Pierce (1852),:

Test extreme observations (largest & smallest) for being an outlier,

Reject if test is significant, average the remaining ones.

Robustness of Tests

Fisher's exact theory for the normal distr. → "monopoly".
E.S. Pearson (1931): Non-robustness of F-test for variances
Non-parametric tests
Focus on robustness of the level,
sometimes neglecting power (but Wilcoxon test is good!)

Robust estimation

Non-robustness of the mean: Tukey (1960)
Minimax theory: Huber (1964)
Influence Function and Breakdown point: Hampel (1968)
Other approaches, see later

Books

Huber (1981) → Huber, Ronchetti (2009)

Hampel, Ronchetti, Rousseeuw, Stahel (1986)

Maronna, Martin, Yohai (2006)

Rousseeuw, Leroy (1987) Regression

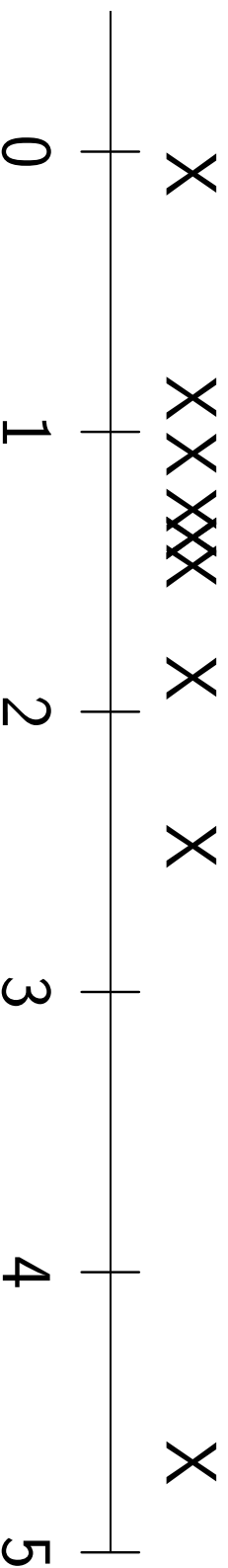
2 Location

2.1 Estimators of Location

- a Example: Prolongation of sleep by a sleeping pill:
Cushny and Peebles, 1905, J. Physiol. 32, 501
„Student”, 1908, Biometrika 6,1).

1.2	2.4	1.3	1.3	0.0	1.0	1.8	0.8	4.6	1.4 .
-----	-----	-----	-----	-----	-----	-----	-----	-----	-------

“How much longer do people sleep?”



b Sorted observations:

0.0 0.8 1.0 1.2 1.3 1.3 1.4 1.8 2.4 4.6

- Average, Mean: 1.45
- Median: 1.3
- **Trimmed Mean** (20%): Drop 2 largest and 2 smallest observations.

Calculate mean of the remaining 6 observations.

→ Average of 1.0, 1.2, 1.3, 1.3, 1.4, 1.8 = 1.33

- Outlier **Rejection Rule**:

If an observation is too far away from the rest, drop it!

Then, calculate the average of the remaining ones.

→ Reject 4.8, get average 1.24

c What do we want?

Advantage of “**vague concepts**” (John Tukey) – like “location”:

Practical question comes before formalization.

- Want to characterize the “typical” observations.
Data Exploration, Exploratory Data Analysis.

- Want Mean (=Average) regardless of distribution:
Insurance needs to cover all cases, cannot disregard “atypical” ones.

Estimate **expected value** $\mathcal{E}\langle X \rangle$, since

$$\bar{X} = \sum_{i=1}^n X_i \approx n \mathcal{E}\langle X \rangle \text{ for large } n.$$

Regression example: Want to characterize the “typical” obs.

→ Predict a new “typical” observation.

2.2 Sensitivity Curve

a Gross Errors

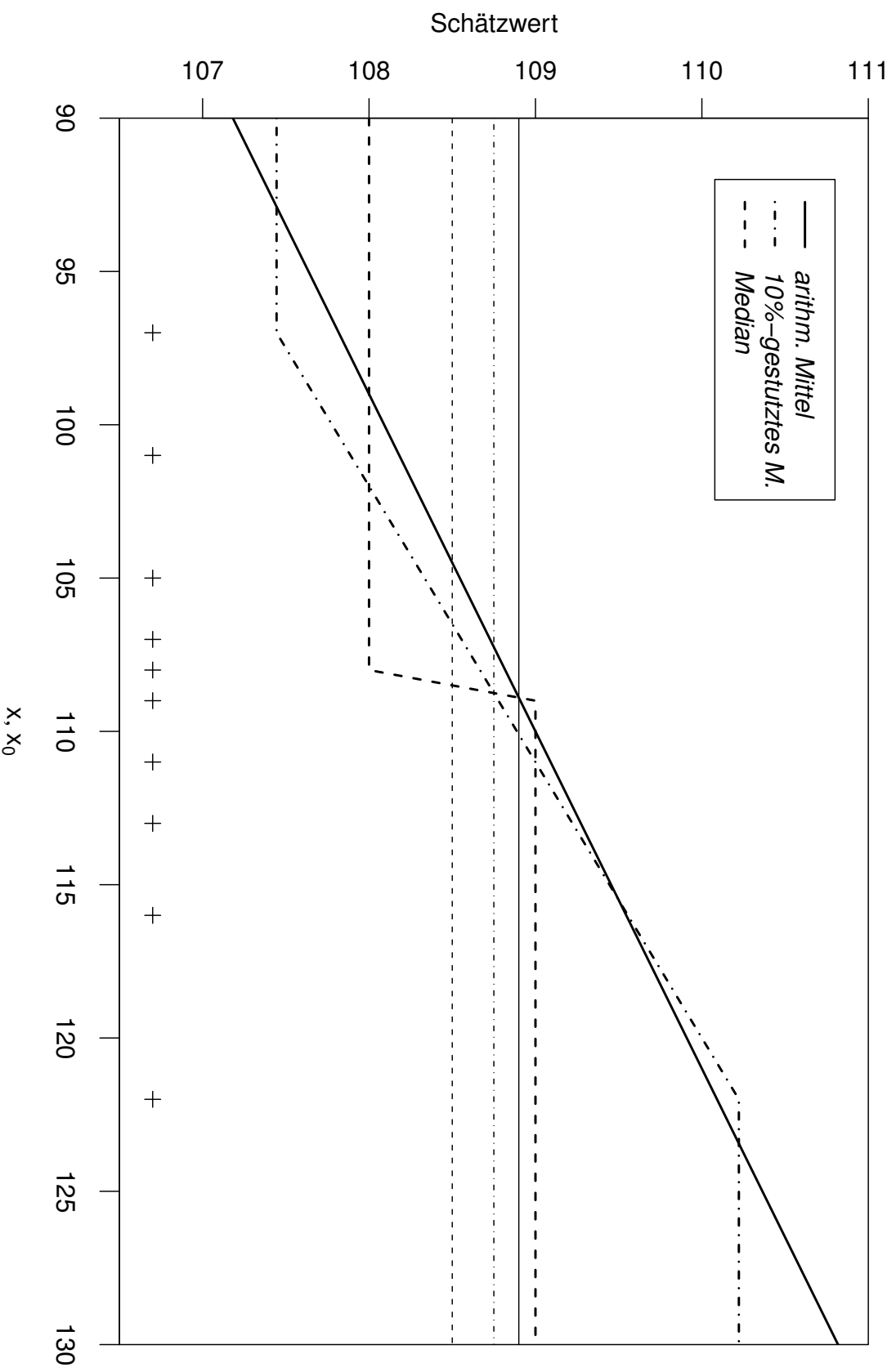
What happens if an **erroneous observation** sneaks in?

“Reasonable Sample” $x_1, \dots, x_n$ plus an arbitrary value x .

Effect on estimator?

$$\begin{aligned} SC \langle x \rangle &= (n+1) \left(\frac{1}{n+1} (x + x_1 + \dots + x_n) - \frac{1}{n} (x_1 + \dots + x_n) \right) \\ &= x + \left(1 - \frac{n+1}{n} \right) (x_1 + x_2 + \dots + x_n) = x - \frac{1}{n} \cdot n\bar{x} \\ &= x - \bar{x} . \end{aligned}$$

(SC depends on the sample $x_1, \dots, x_n$.)



For a robust estimator, the sensitivity curve must be bounded.

Average is not robust!

2.3 Location Model

a Model:

$X_i \sim \mathcal{F}(\mu)$, $i = 1, \dots, n$, independent

$$F(x; \mu) = F(x - \mu; 0) =: F_0(x - \mu)$$

Often, $\mathcal{F}(\mu) = \mathcal{N}(\mu, 1)$, normal distribution.

Assume that the width of the distribution (scale) is known
(= unrealistic simplification, will soon be dropped.)

b **Estimators** $T_n(x_1, \dots, x_n)$

Requirements:

- independent of sequence of observations
- **equivariant:**

$$T_n(x_1 + c, \dots, x_n + c) = T_n(x_1, \dots, x_n) + c$$

^c Known Theory for “known” $\mathcal{F}_0$

- If $\mathcal{F}_0 = \mathcal{N}(\mu, 1)$, average $\bar{X} = (1/n) \sum_{i=1}^n X_i$ is **optimal**:
Smallest variance among all equivariant estimators.
- If $\mathcal{F}_0$ is known, **Maximum Likelihood Estimator** is **optimal**:
Smallest variance among all equivariant estimators –
asymptotically!

3 Maximum Likelihood and M-estimators

3.1 Maximum Likelihood

- a Given a parametric family with density $f_{\underline{\theta}}(x)$
 write as $f(x, \underline{\theta})$

Given the observation x , $f(x, \underline{\theta})$ = likelihood of $\underline{\theta}$.

Same for probabilities $P_{\underline{\theta}}(X = x)$, also written as $f(x, \underline{\theta})$.

Maximum-Likelihood estimator: maximize the likelihood !

$$\hat{\underline{\theta}} = \arg \min \langle f(x, \underline{\theta}) \rangle$$

- Sample $X_1, \dots, X_n \longrightarrow$ joint density $\prod_i f(x_i, \underline{\theta})$.

Maximize this or

$$L(x_1, \dots, x_n; \underline{\theta}) = \sum_i \log \langle f(x_i, \underline{\theta}) \rangle$$

or minimize

$$D(\underline{\theta}) = -2 \sum_i \log \langle f(x_i, \underline{\theta}) \rangle = \sum_i \rho(x_i, \underline{\theta})$$

D : **Deviance**.

ρ -function: “deviation” of the observation x_i from model $\langle \underline{\theta} \rangle$.

c **Example Normal distribution**

For normal distribution with given variance:

$$\rho \langle x, \mu \rangle = ((x - \mu) / \sigma)^2 + c$$

→ Least Squares

d **Example logistic distribution**

$$f_{\mu, \sigma} \langle x \rangle = \frac{1}{(e^{z/2} + e^{-z/2})^2}, \quad z = \frac{x - \mu}{\sigma}$$

Location-scale family. Likelihood $-2 \log \langle e^{z/2} + e^{-z/2} \rangle$.

$$\rho \langle x; \mu, \sigma \rangle = 4 \log \left\langle e^{(x-\mu)/2\sigma} + e^{-(x-\mu)/2\sigma} \right\rangle$$

e **General case**

Form derivative for θ_k and set to 0

$$\frac{\partial L}{\partial \theta_k} \langle x_1, \dots, x_n; \underline{\theta} \rangle = \sum_{i=1}^n s_k \langle x_i; \underline{\theta} \rangle ,$$

„Likelihood scores“

$$s_k \langle x; \underline{\theta} \rangle = \frac{\partial}{\partial \theta_k} \log \langle f \langle x_i, \underline{\theta} \rangle \rangle = -\frac{1}{2} \frac{\partial}{\partial \theta_k} \rho \langle x_i, \underline{\theta} \rangle .$$

$$\sum_{i=1}^n \underline{s} \left\langle x_i; \underline{\hat{\theta}} \right\rangle = 0 .$$

Maximum-Likelihood estimator: Solve for $\underline{\hat{\theta}}$!

f **Logistic distribution.** Location-scale: $z = (x - \mu)/\sigma \longrightarrow$

$$\partial z / \partial \mu = -1/\sigma \quad \text{and} \quad \partial z / \partial \sigma = -(x - \mu)/\sigma^2 = -z/\sigma$$

$$\begin{aligned} s_\mu \langle x; \underline{\theta} \rangle &= \frac{1}{\sigma} \frac{e^{z/2} - e^{-z/2}}{e^{z/2} + e^{-z/2}} \\ s_\sigma \langle x; \underline{\theta} \rangle &= z s_\mu \langle x; \underline{\theta} \rangle \end{aligned}$$

Use scores function of the logistic distribution even if you think (or hope) that the observations follow the normal distribution.

3.2 M-estimators

a Maximum likelihood estimators can be written in 2 forms:

- Maximize the log likelihood! For sample of iid observations:

$$\operatorname{argmax}_{\underline{\theta}} \left\langle \sum_i \log \langle f \langle \underline{x}_i \rangle \rangle \right\rangle$$

or minimize sum of deviances

$$\operatorname{argmin}_{\underline{\theta}} \left\langle \sum_i \rho \langle \underline{x}_i; \underline{\theta} \rangle \right\rangle \quad \rho \langle \underline{x}; \underline{\theta} \rangle = -2 \log \langle f_{\underline{\theta}} \langle \underline{x} \rangle \rangle$$

- Solve implicit equation

$$\sum_i \psi \langle \underline{x}_i; \underline{\theta} \rangle = \underline{0}, \quad \psi \langle \underline{x}; \underline{\theta} \rangle = \frac{\partial}{\partial \underline{\theta}} \rho \langle \underline{x}; \underline{\theta} \rangle$$

Note: For numerical purposes, minimization is often easier

BUT: Specialized algorithms are often based on second form.

b **M-estimators:**

Use Maximum likelihood **even when you do not “believe”** that

f is the density governing the observations $\underline{X}_i$.

= Do not require that ρ is $-2 \log \langle f \rangle$ for any distribution (but it will almost always).

c M-estimators also appear in the same 2 flavors.

Important: We do not require that $\underline{\psi}$ is $\partial \rho / \partial \underline{\theta}$ for any ρ
(in case of multidimensional $\underline{\theta}$)

Second type is more flexible!

d **M-estimators for location**

$$\begin{aligned}\rho\langle x, \mu \rangle &= \rho\langle x - \mu, 0 \rangle =: \rho\langle x \rangle \\ \psi\langle x, \mu \rangle &= \psi\langle x - \mu, 0 \rangle =: \psi\langle x \rangle\end{aligned}$$

For the normal distribution with $\sigma = 1$:

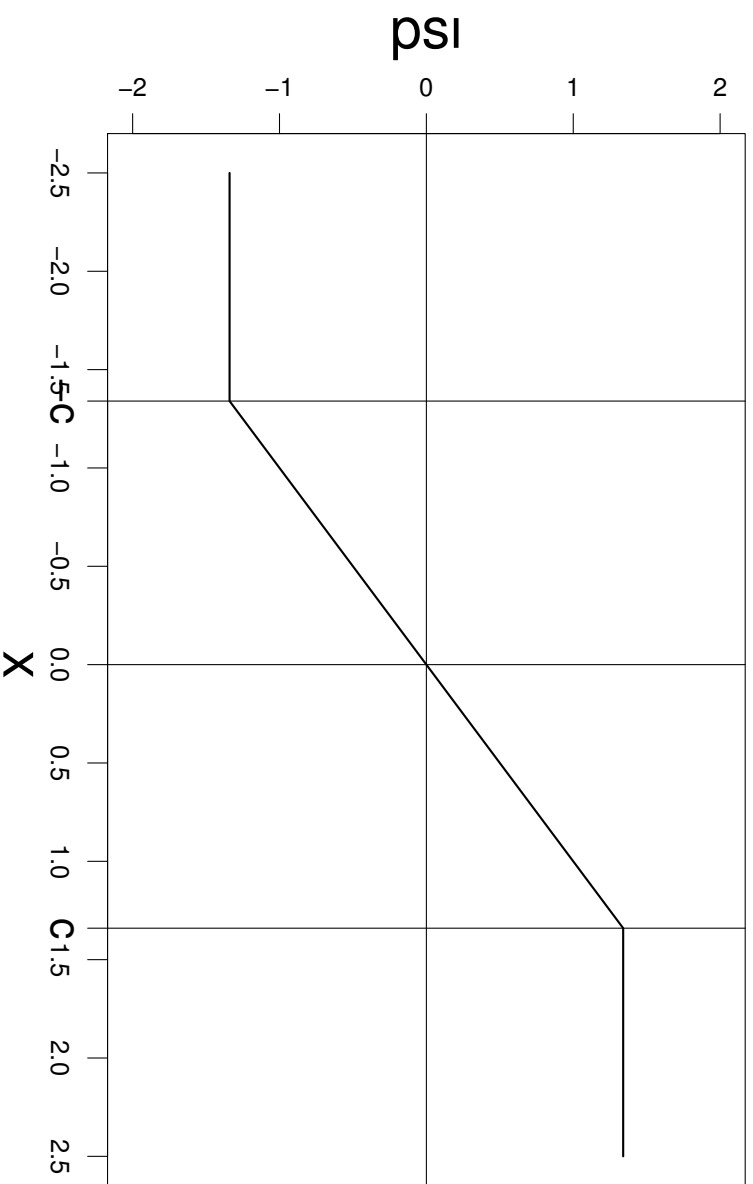
$$\begin{aligned}f_\mu\langle x \rangle &= \text{const } e^{-(x-\mu)^2/2} \\ \rho\langle x \rangle &= x^2 \\ \psi\langle x \rangle &= x \quad (\cdot 2)\end{aligned}$$

e

Choose a ψ function!

Huber:

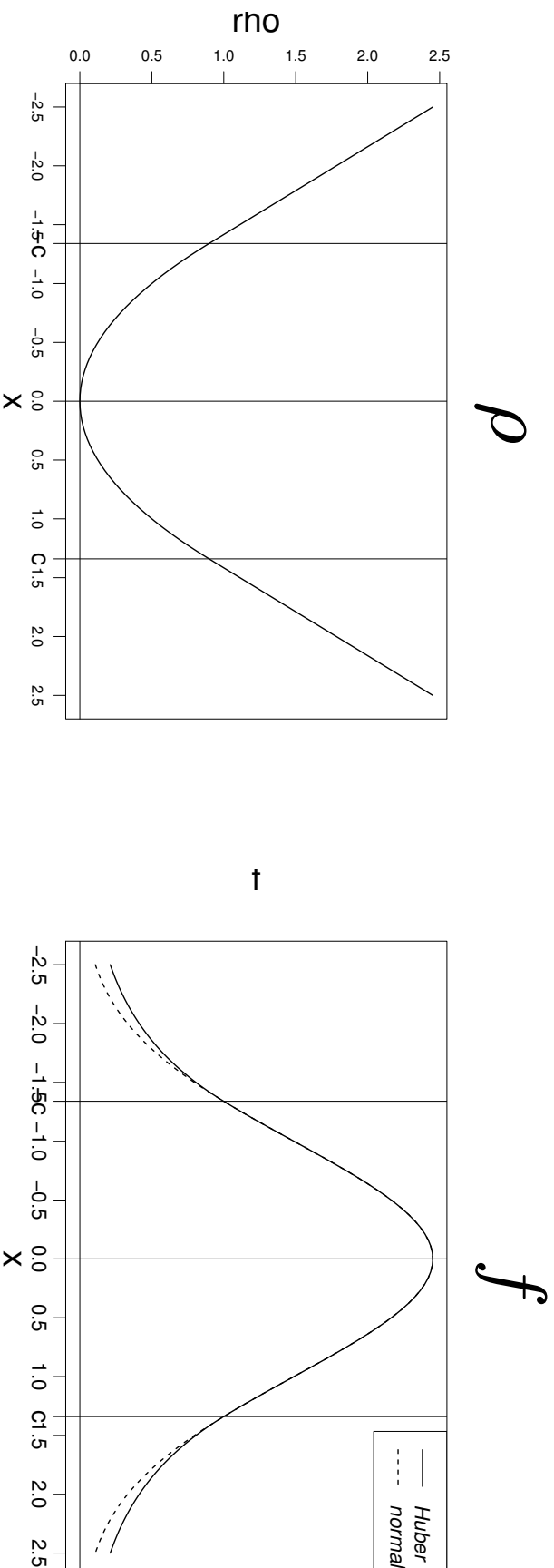
$$\psi\langle x \rangle = \begin{cases} c & \text{for } x > c \\ x & \text{for } -c < x < c \\ -c & \text{for } x < -c \end{cases} = \min\langle c, \max\langle -c, x \rangle \rangle$$



Integrate ($\longrightarrow \rho$) and exponentiate to get corresponding f :

$$\rho\langle x \rangle = \begin{cases} cx - c^2/2 & \text{for } x > c \\ x^2/2 & \text{for } -c < x < c \\ -cx - c^2/2 & \text{for } x < -c \end{cases}$$

$$f\langle x \rangle = \text{const.} \cdot e^{-\rho\langle x \rangle}$$



Start from desired ψ -function and find the distribution such that the M-estimator is the max.li. estimator for it.

f M-estimators as weighted means

Weighted mean: Given weights w_i , use

$$\bar{x}_w = \sum_i w_i x_i / \sum_i w_i$$

Used for observations with different precision $\text{var} \langle X_i \rangle = \sigma^2 w_i$

$\longrightarrow w_i = 1/v_i$ gives optimal estimator for normal data.

M-estimator $\sum_i \psi \langle x_i - \hat{\mu} \rangle = 0$.

Write $\psi \langle x_i - \mu \rangle = w_i (x_i - \mu)$, $w_i = \omega \langle x_i - \mu \rangle$,

$\omega \langle z \rangle = \psi \langle z \rangle / z$. Then,

$$\hat{\mu} = \sum_i w_i x_i / \sum_i w_i$$

weighted mean — where the weights depend on the solution!

Useful for

- **intuitive understanding** – sometimes misleading!
- Iterative algorithm:

Starting value $\hat{\mu}^{(0)}$

Step m : weights $w_i^{(m)} = \omega \langle x_i - \hat{\mu}^{(m-1)} \rangle$

$$\longrightarrow \hat{\mu}^{(m)} = \sum_i w_i^{(m)} x_i / \sum_i w_i^{(m)}$$

4 Asymptotics

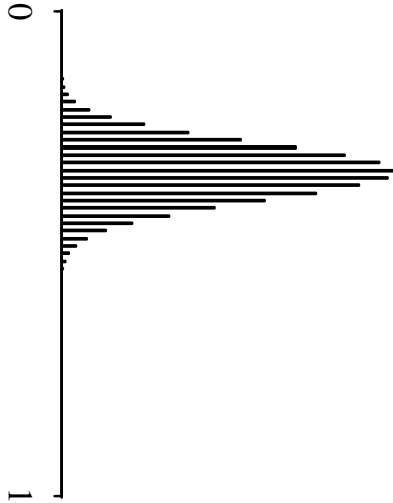
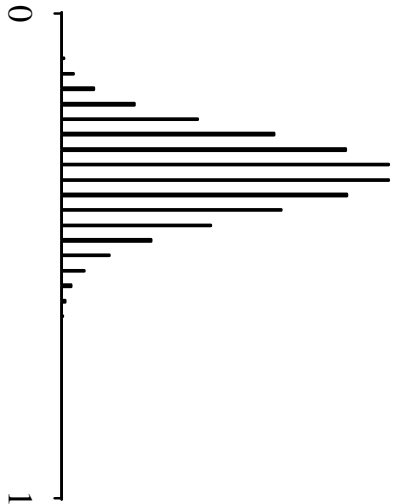
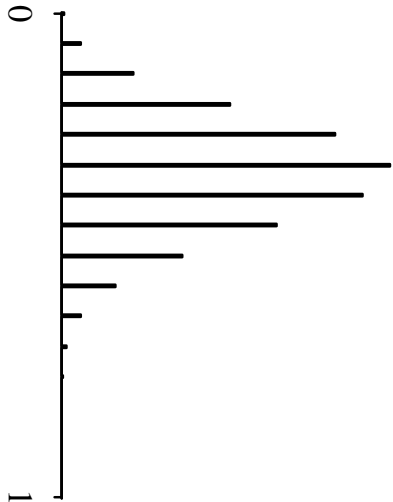
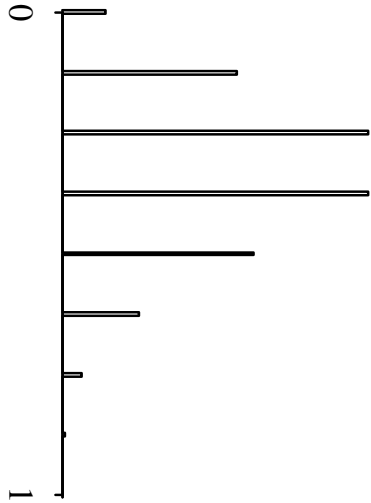
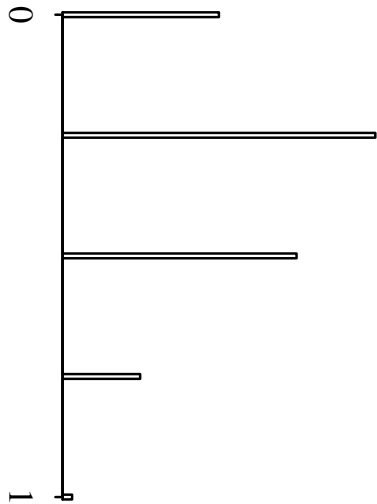
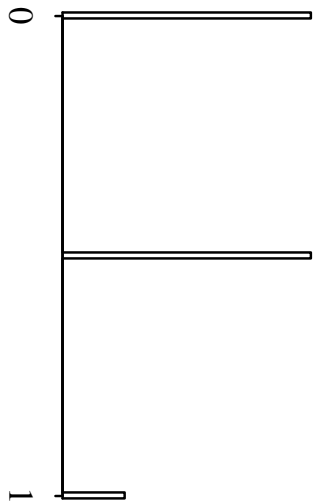
4.1 Consistency

b **Relative frequency** $R_n \rightarrow P\langle A \rangle = \pi = \mathcal{E}\langle R_n \rangle$. More precisely,

$$\lim_{n \rightarrow \infty} P\{|R_n - \pi| > \varepsilon\} = 0$$

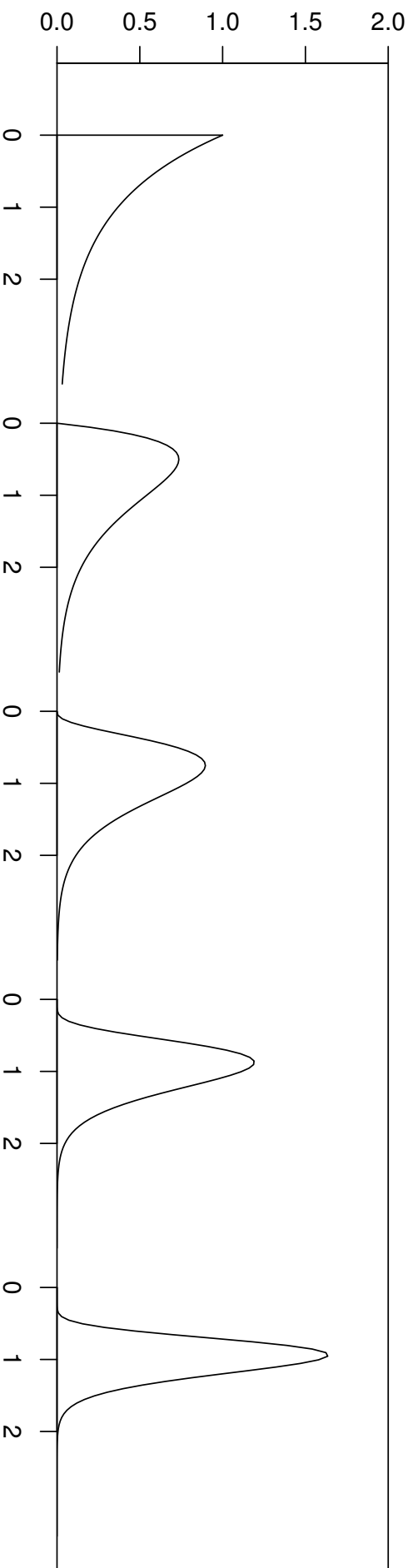
Law of Large Numbers. Jakob Bernoulli (published posth. 1713)

$$\begin{aligned} X \sim \mathcal{B}\langle n, \pi \rangle &\longrightarrow \mathcal{E}\langle X/n \rangle = \mathcal{E}\langle X \rangle / n = n\pi / n = \pi \\ \text{var}\langle X/n \rangle = \text{var}\langle X \rangle / n^2 &= n\pi(1 - \pi) / n^2 = \pi(1 - \pi) / n \\ \longrightarrow 0 \quad \implies \quad X/n &\rightarrow \pi. \end{aligned}$$



c General case: $X_1, X_2, \dots, X_n \dots$ indep. $\longrightarrow$

$$P\{|\bar{X}_n - \mu| > \varepsilon\} \xrightarrow{n \rightarrow \infty} 0 \quad \text{for each } \varepsilon > 0,$$



d **Consistency of the characteristic values.** $\hat{\gamma} \rightarrow \gamma$.

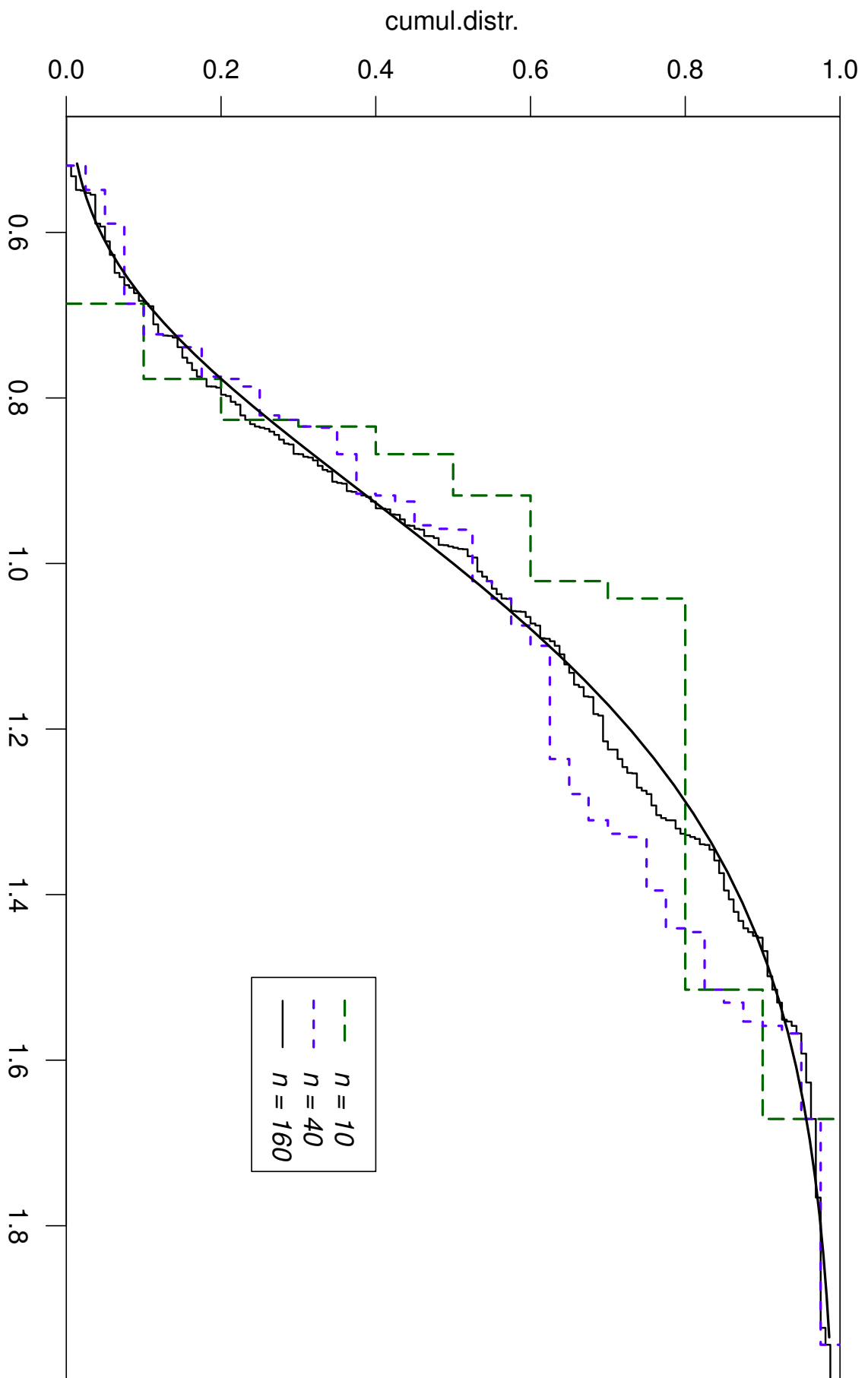
For a derivation, we will need 2 more concepts:
Integrals, Functionals.

First, note that ...

e Empirical cumulative **distribution function** $\longrightarrow$ theoretical,

$$\hat{F}_n \langle x \rangle = \frac{1}{n} \# \langle i : X_i \leq x \rangle \xrightarrow{n \rightarrow \infty} F \langle x \rangle = P \langle X \leq x \rangle .$$

by the law of large numbers.



f **Integrals** $\mathcal{E}\langle X \rangle = \int x dF\langle x \rangle$

For continuous random variables:

$$\mathcal{E}\langle X \rangle = \int x dF\langle x \rangle = \int x f\langle x \rangle dx$$

For discrete ones:

$$\mathcal{E}\langle X \rangle = \int x dF\langle x \rangle = \sum_x x \langle X = x \rangle$$

For empirical distribution functions:

$$\int x d\hat{F}_n\langle x \rangle = \sum_i x_i \frac{1}{n} = \bar{X}$$

Same for functions $g\langle x \rangle$:

$$\int g\langle x \rangle d\hat{F}_n\langle x \rangle = \sum_i g\langle x_i \rangle \frac{1}{n}$$

= mean of $g\langle x_i \rangle$.

g **Functionals.**

$T\langle F \rangle$ theoretical characteristic value, $T\langle \hat{F}_n \rangle$ empirical.

$$T\langle \hat{F}_n \rangle \xrightarrow{n \rightarrow \infty} T\langle F \rangle$$

since $\hat{F}_n \xrightarrow{n \rightarrow \infty} F$.

Needs continuity of the functional.

h **M estimators as functionals.**

$$\underline{T}_\rho \langle F \rangle = \operatorname{argmin}_{\underline{\theta}} \int \rho \langle x, \underline{\theta} \rangle dF \langle x \rangle \quad \text{or}$$

$$\underline{T}_\psi \langle F \rangle = \text{solution of } \int \underline{\psi} \langle x, \underline{\theta} \rangle dF \langle x \rangle = 0 .$$

Estimator: Solution of

$$\int \underline{\psi} \langle x, \underline{\theta} \rangle d\hat{F}_n \langle x \rangle = \frac{1}{n} \sum_i \underline{\psi} \langle x_i, \underline{\theta} \rangle = 0$$

More clearly, write

$$\begin{aligned} \int \underline{\psi} \langle x, \underline{T} \langle \hat{F}_n \rangle \rangle d\hat{F}_n \langle x \rangle &= 0 \\ \int \underline{\psi} \langle x, \underline{T} \langle F \rangle \rangle dF \langle x \rangle &= 0 \end{aligned}$$

This makes clear that $\underline{T} \langle F_n \rangle \rightarrow \underline{T} \langle F \rangle$, since $F_n \rightarrow F$.

i **Example: Location model, arithm. mean**

$$f\langle x, \mu \rangle = f\langle x - \mu, 0 \rangle, \quad f\langle z, 0 \rangle = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

$$\text{log-lik.: } -\frac{1}{2} \log \langle 2\pi \rangle - (x - \mu)^2/2, \quad \text{scores } s\langle x, \mu \rangle = x - \mu$$

$$\longrightarrow \text{estimator } \sum_i (x_i - \hat{\mu}) = 0 \longrightarrow \hat{\mu} = \frac{1}{n} \sum_i x_i$$

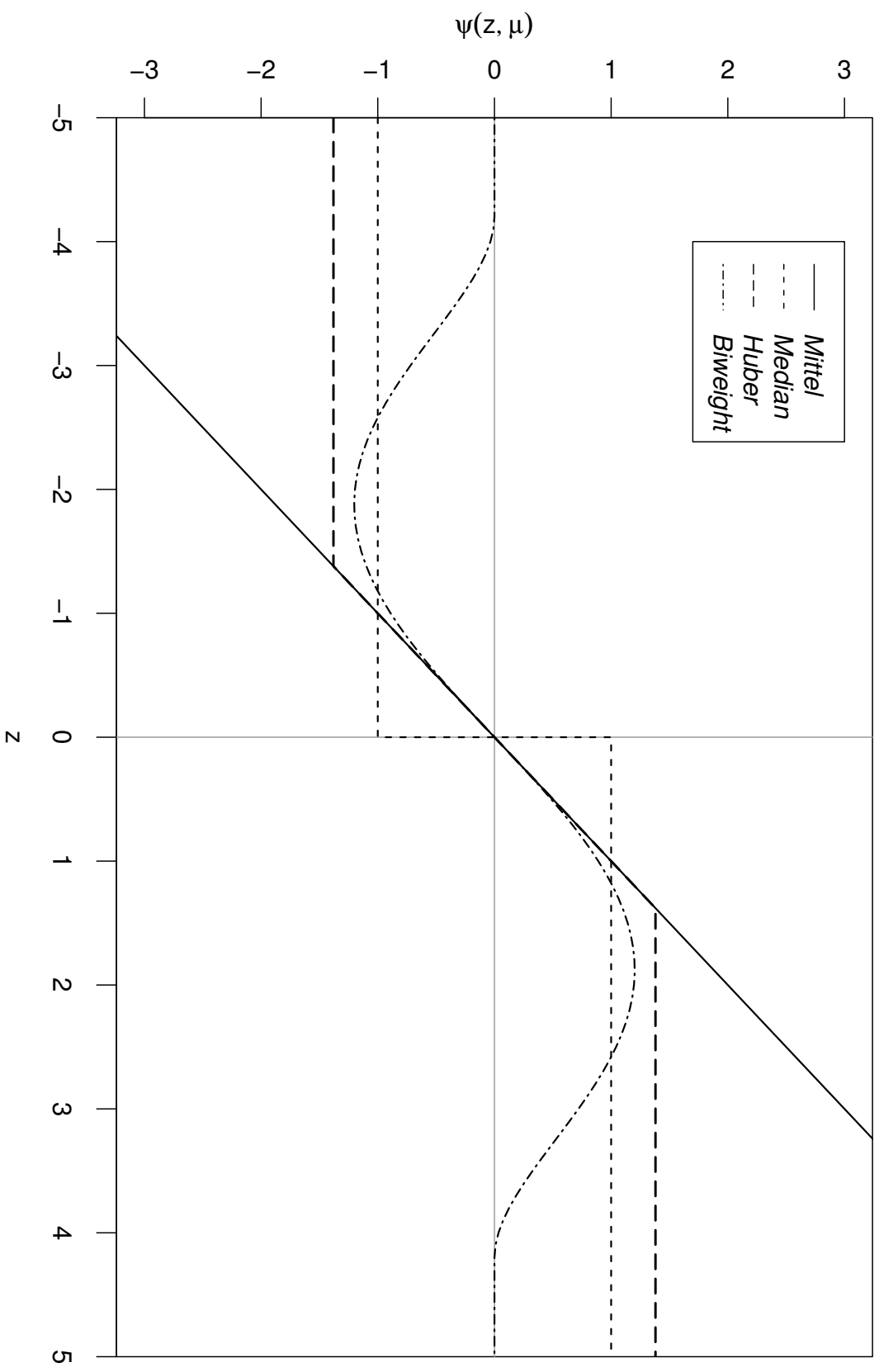
$$\longrightarrow \text{Mean is M estimator with } \psi\langle x, \mu \rangle = x - \mu$$

$$\text{Median is an M estimator, } \psi\langle x, \mu \rangle = \begin{cases} -1 & x - \mu \leq 0 \\ 1 & x - \mu > 0 \end{cases}.$$

Huber estimator: M estimator with

$$\psi\langle x, \mu \rangle = \begin{cases} x - \mu & \text{für } |x - \mu| \leq b \\ -b & \text{für } x - \mu < -b \\ b & \text{für } x - \mu > b \end{cases}$$

b tuning constant.



j Maximum-Likelihood estimator: $X_i \sim F_{\underline{\theta}} \implies \underline{T}_n \xrightarrow{n \rightarrow \infty} \underline{T} \langle F_{\underline{\theta}} \rangle$.

$\underline{T} \langle F_{\underline{\theta}} \rangle$ defined by

$$\int \underline{s} \langle x, \underline{T} \langle F_{\underline{\theta}} \rangle \rangle f \langle x, \underline{\theta} \rangle dx = 0$$

General result: $\underline{T} \langle F_{\underline{\theta}} \rangle = \underline{\theta}$:

Fisher-consistent estimator for the parameter(s) of the family!

Proof: Need to show $\int \underline{s} \langle x, \underline{\theta} \rangle f \langle x, \underline{\theta} \rangle dx = 0$.

$$\begin{aligned} \frac{\partial}{\partial \theta} \log \langle f \langle x, \underline{\theta} \rangle \rangle &= \frac{1}{f \langle x, \underline{\theta} \rangle} \frac{\partial}{\partial \theta} f \langle x, \underline{\theta} \rangle = \underline{s} \langle x, \underline{\theta} \rangle \implies \\ \frac{\partial}{\partial \theta} f \langle x, \underline{\theta} \rangle &= \underline{s} \langle x, \underline{\theta} \rangle f \langle x, \underline{\theta} \rangle \end{aligned}$$

$$\begin{aligned} \int f \langle x, \underline{\theta} \rangle dx &= 1 \\ \Rightarrow \int \frac{\partial}{\partial \theta} f \langle x, \underline{\theta} \rangle dx &= 0 = \int \underline{s} \langle x, \underline{\theta} \rangle f \langle x, \underline{\theta} \rangle dx \end{aligned}$$

k Fisher Consistency

(stated again) Let $\mathcal{F}_{\underline{\theta}}$ be a parametric family with par. $\underline{\theta} = [\theta_1, \dots, \theta_p]$

Sample $X_1, \dots, X_n, X_i \sim \mathcal{F}_{\underline{\theta}}$

Estimator of the parameters:

A function of the sample, $T_k \langle \hat{F}_n \rangle$ designed to estimate par. θ_k

→ If $F = F_{\underline{\theta}}$ then we want $T_k \langle F_{\underline{\theta}} \rangle = \theta_k$

Vector notation: $\underline{T} \langle \underline{\theta} \rangle = \underline{\theta}$

“Fisher consistency”.

| **M estimators:** $T \langle F_{\underline{\theta}} \rangle = \underline{\theta} \iff$

$$\int \underline{\psi} \langle x, \underline{\theta} \rangle dF_{\underline{\theta}} \langle x \rangle = 0$$

Location Model: $F \langle x, \mu \rangle = F_0 \langle x - \mu \rangle$

Fisher consistent for all μ if

$$\int \psi \langle x \rangle dF_0 \langle x \rangle = 0$$

Sufficient: Symmetric f and (anti-) symmetric ψ .

If not fulfilled, it is easily corrected by considering

$$\widetilde{\psi} \langle x \rangle := \psi \langle x \rangle - \int \psi \langle x \rangle dF_0 \langle x \rangle \text{ or}$$

$$\widetilde{\psi} \langle x \rangle := \psi \langle x - T_{\psi} \langle F_0 \rangle \rangle.$$

4.2 Influence Function

b Example of weights of pigs

107 108 111 101 97 113 109 105 116 122.

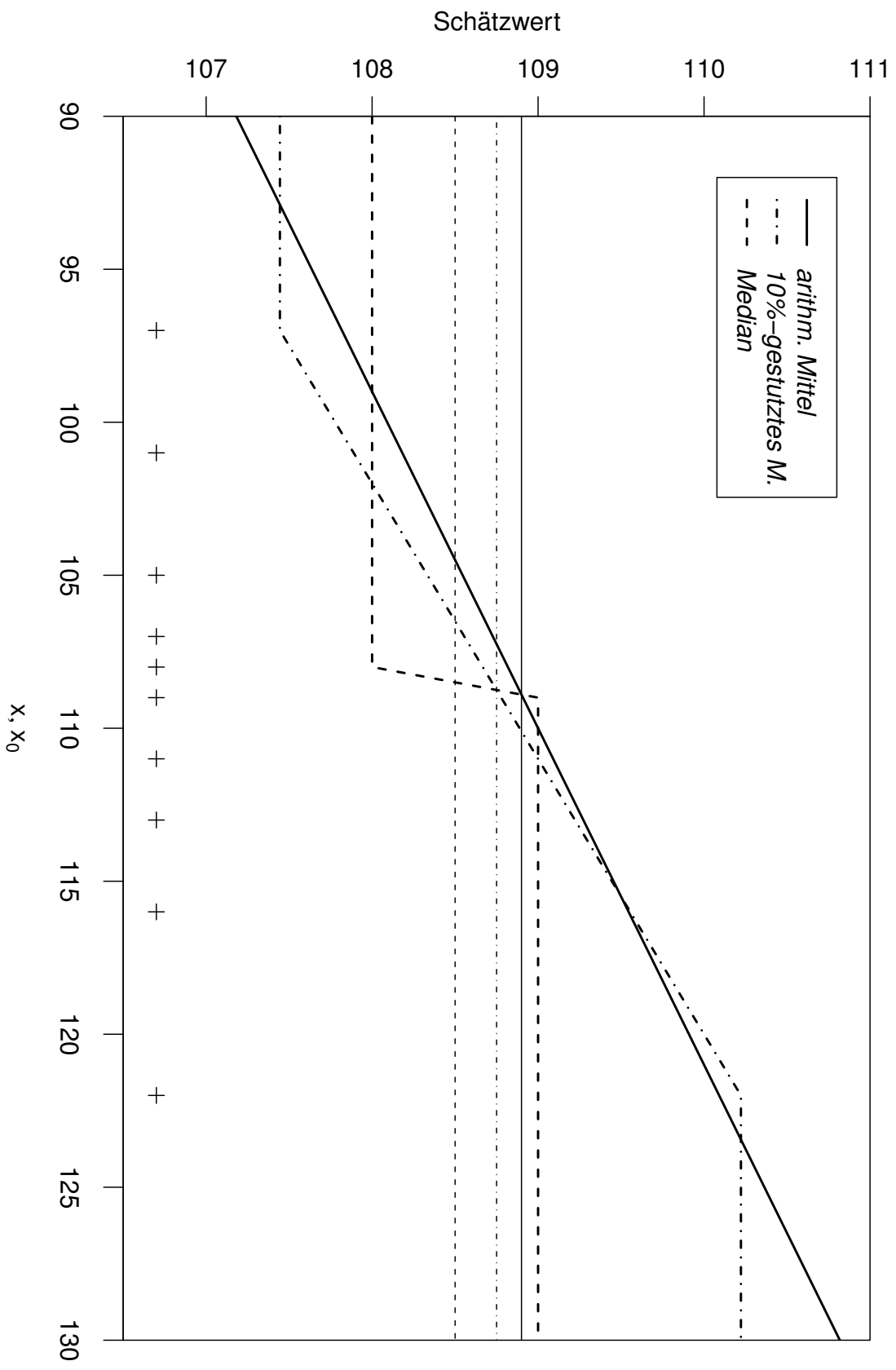
Additional observation $x_0 \longrightarrow$

$$\begin{aligned} & (107 + 108 + \dots + 122 + x_0)/11 \\ &= \frac{n\bar{x} + x_0}{n+1} = \frac{(n+1)\bar{x} - \bar{x}}{n+1} + \frac{x_0}{n+1} = \bar{x} + \frac{1}{n+1}(x_0 - \bar{x}) \\ &= 108.9 + (x_0 - 108.9)/11. \end{aligned}$$

Median $(108 + 109)/2 = 108.5$

$\longrightarrow 108$, if $x_0 \leq 108$

$\longrightarrow 109$, if $x_0 \geq 109$



c **Empirical Influence Function or Sensitivity Curve**

$$SC \langle x_0; T, x_1, \dots, x_n \rangle = n \left(T \langle x_1, \dots, x_n, x_0 \rangle - T \langle x_1, \dots, x_n \rangle \right)$$

d SC depends on the dataset.

Want an “asymptotic version” that depends on the underlying distribution F instead.

$$n \rightarrow \infty \longrightarrow \text{sample} \hat{=} \hat{F}_n \rightarrow F.$$

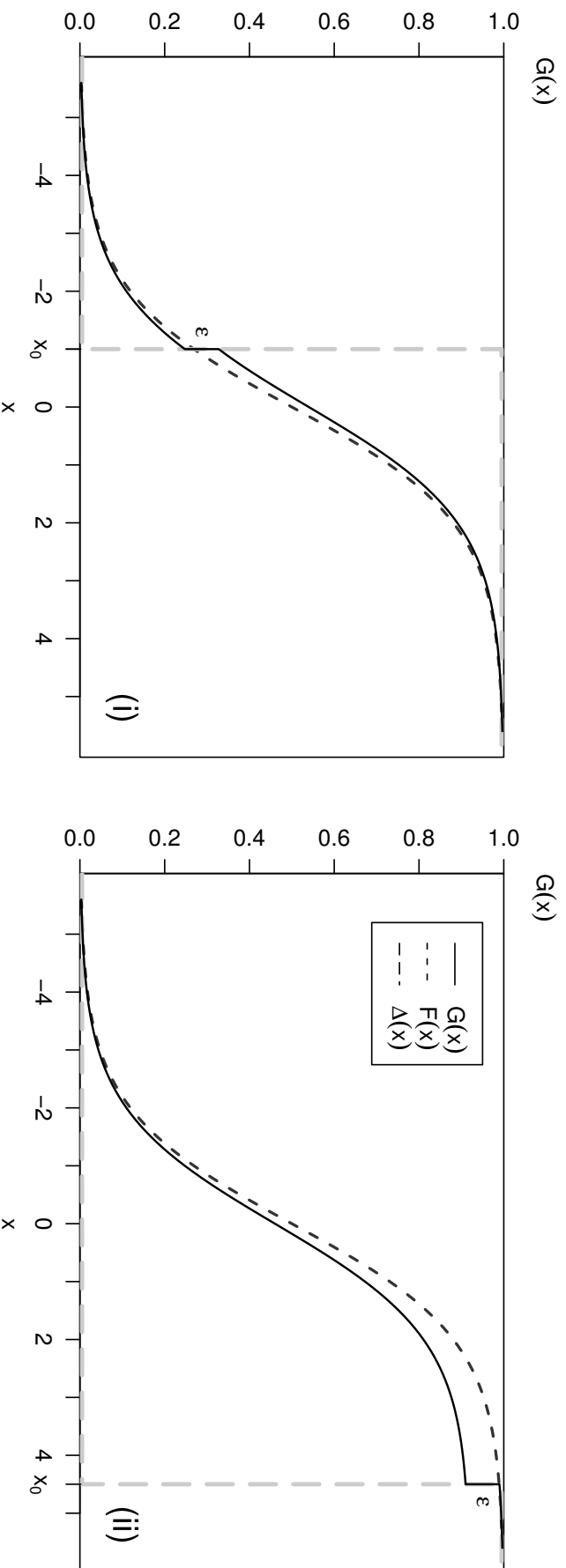
“Additional observation” $\longrightarrow$ small fraction of obs.

e

Gross Error Model

$$G\langle \cdot \rangle = (1 - \varepsilon) F\langle \cdot \rangle + \varepsilon \Delta_{x_0}\langle \cdot \rangle$$

Model for “wrong” observation or “gross error”



More general: $G\langle x \rangle = (1 - \varepsilon) F\langle x \rangle + \varepsilon H\langle x \rangle$

= **Gross Error Model** $\longrightarrow$ Gross Error “neighborhood”

$U_\varepsilon\langle F \rangle = \{(1 - \varepsilon) F\langle \cdot \rangle + \varepsilon H\langle \cdot \rangle \mid H \text{ arbitrary}\}$
 (not a real neighborhood!)

f Influence Function.

$$\text{IF} \langle x; T, F \rangle = \lim_{\varepsilon \rightarrow 0} \frac{T \langle (1 - \varepsilon) F + \varepsilon \Delta_x \rangle - T \langle F \rangle}{\varepsilon}.$$

Limit of the Sensitivity Curve.

g Gross Error Sensitivity

IF measures the “damage” to the “true value” of an estimator caused by a gross error.

Keep the damage bounded!

→ Measure of robustness: The maximum of $|\text{IF}|$:

$$\gamma^* \langle T, F \rangle = \sup_x \langle |\text{IF} \langle x; T, F \rangle| \rangle$$

h IF for $\bar{X}$:

$$\begin{aligned}\mathcal{E}\langle (1-\varepsilon)F + \varepsilon\Delta_x \rangle &= (1-\varepsilon)\mathcal{E}\langle F \rangle + \varepsilon\mathcal{E}\langle \Delta_x \rangle \\ &= (1-\varepsilon)\mathcal{E}\langle F \rangle + \varepsilon x \\ \text{IF}\langle x; \bar{X}, F \rangle &= x - \mathcal{E}\langle F \rangle\end{aligned}$$

* Median $\text{med}\langle F \rangle = F^{-1}\langle 0.5 \rangle$.

If $x > t_{\varepsilon}^{+} = F^{-1}\langle 0.5/(1 - \varepsilon) \rangle$ then $\text{med}\langle (1 - \varepsilon)F + \varepsilon\Delta_x \rangle = t_{\varepsilon}^{+}$

If $x > t_{\varepsilon}^{-} = F^{-1}\langle 1 - 0.5/(1 - \varepsilon) \rangle$ then ... = t_{ε}^{-}

$$\frac{d}{d\varepsilon} F^{-1}\langle 0.5/(1 - \varepsilon) \rangle = \frac{1}{f\langle F^{-1}\langle 0.5/(1 - \varepsilon) \rangle \rangle} \cdot \frac{0.5 \cdot (-1)}{(1 - \varepsilon)^2} \cdot (-1) \\ \rightarrow 1/(2f\langle \mu \rangle), \quad \mu = F^{-1}\langle 0.5 \rangle$$

$$\text{IF}\langle x; \text{med}, F \rangle = \begin{cases} -1/(2f\langle \mu \rangle) & \text{für } x < \text{med}\langle F \rangle \\ 1/(2f\langle \mu \rangle) & \text{für } x > \text{med}\langle F \rangle \end{cases}$$

is not continuous at the median, but jumps

from $-1/(2f\langle \mu \rangle)$ to $1/(2f\langle \mu \rangle)$.

j Influence Function for M estimators.

$$\text{IF} \langle x; F \rangle = \frac{1}{c} \psi \langle x, \theta \rangle \quad \text{with} \quad c = - \int \frac{\partial}{\partial \theta} \psi \langle x, \theta \rangle f \langle x, \theta \rangle dx .$$

Influence function is proportional to ψ .

If T estimates θ , that is, $\int \psi \langle x, \underline{\theta} \rangle f \langle x, \underline{\theta} \rangle dx = 0$, then

$$c = \int \psi \langle x, \underline{\theta} \rangle s \langle x, \underline{\theta} \rangle f \langle x, \underline{\theta} \rangle dx .$$

Special case: Maximum-Likelihood estimator

$$c = - \int \frac{\partial}{\partial \theta} s \langle X, \theta \rangle f \langle x, \underline{\theta} \rangle dx = \int s \langle X, \theta \rangle^2 f \langle x, \underline{\theta} \rangle dx .$$

* Proof: $T \langle G \rangle$ for the distribution $G = (1 - \varepsilon)F + \varepsilon \Delta_x$:

$$\begin{aligned}
 \int \psi \langle x, T \langle G \rangle \rangle dG \langle x \rangle &= (1 - \varepsilon) \int \psi \langle x, T \langle G \rangle \rangle dF \langle x \rangle + \varepsilon \psi \langle x, T \langle G \rangle \rangle \\
 \psi \langle x, T \langle G \rangle \rangle &\approx \psi \langle x, T \langle F \rangle \rangle + \frac{\partial}{\partial \theta} \psi \langle x, T \langle F \rangle \rangle (T \langle G \rangle - T \langle F \rangle) \\
 \int \psi \langle x, T \langle G \rangle \rangle dF \langle x \rangle &\approx \int \psi \langle x, T \langle F \rangle \rangle dF \langle x \rangle \\
 &\quad + (T \langle G \rangle - T \langle F \rangle) \int \frac{\partial}{\partial \theta} \psi \langle x, T \langle F \rangle \rangle dF \langle x \rangle \\
 \int \psi \langle x, T \langle G \rangle \rangle dG \langle x \rangle &\approx \\
 (1 - \varepsilon) (T \langle G \rangle - T \langle F \rangle) \int \frac{\partial}{\partial \theta} \psi \langle x, T \langle F \rangle \rangle dF \langle x \rangle \\
 &\quad + \varepsilon (\psi \langle x, T \langle F \rangle \rangle + (T \langle G \rangle - T \langle F \rangle) \frac{\partial}{\partial \theta} \psi \langle x, T \langle F \rangle \rangle)
 \end{aligned}$$

Neglect term with $\varepsilon(T \langle G \rangle - T \langle F \rangle)!$ Set $= 0$. $\longrightarrow$

$$T \langle G \rangle T \langle F \rangle \approx - \frac{\varepsilon \psi \langle x, T \langle F \rangle \rangle}{(1 - \varepsilon) \int \frac{\partial}{\partial \theta} \psi \langle x, T \langle F \rangle \rangle dF \langle x \rangle}$$

k Linearization.

$$T \left\langle \hat{F}_n \right\rangle \approx T \langle F \rangle + \frac{1}{n} \sum_{i=1}^n \text{IF} \langle X_i; T, F \rangle$$

$$T \langle G \rangle \approx T \langle F \rangle + \int \text{IF} \langle x; T, F \rangle dG \langle x \rangle$$

If the latter holds, it is called first order **von Mises** expansion and IF is the “kernel” of the “von Mises functional” T .

This looks like a first order Taylor expansion.

$$G = F \longrightarrow \int \text{IF} \langle x; T, F \rangle dF \langle x \rangle = 0$$

There are several notions of derivatives for functionals:

- Gâteaux
- von Mises
- Hadamard
- Fréchet

They are increasingly strong properties of functionals.

4.3 Asymptotic Distribution

a Central limit theorem.

The distribution of the standardized average

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}}$$

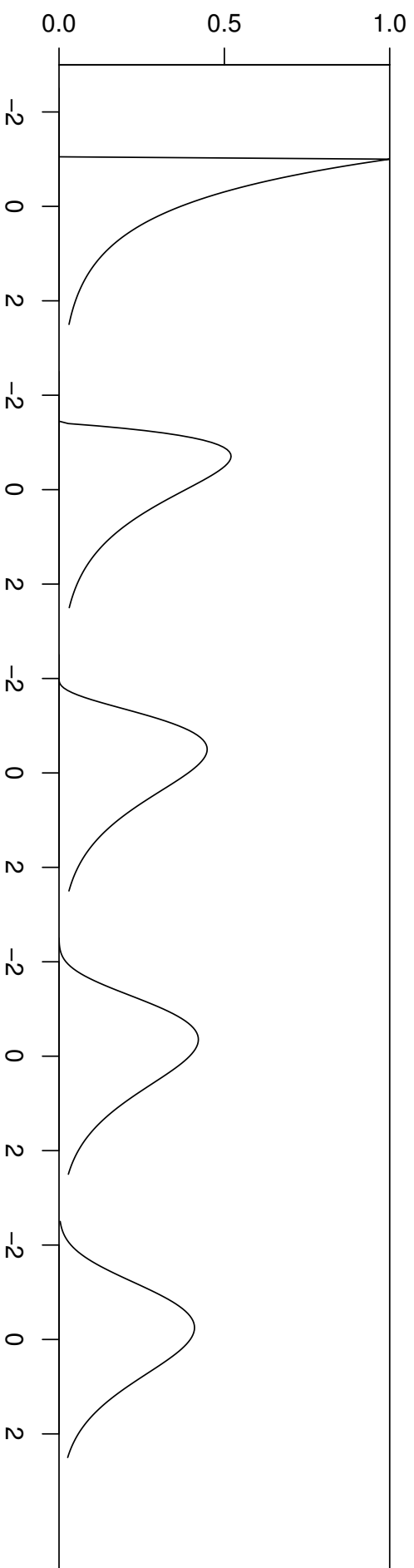
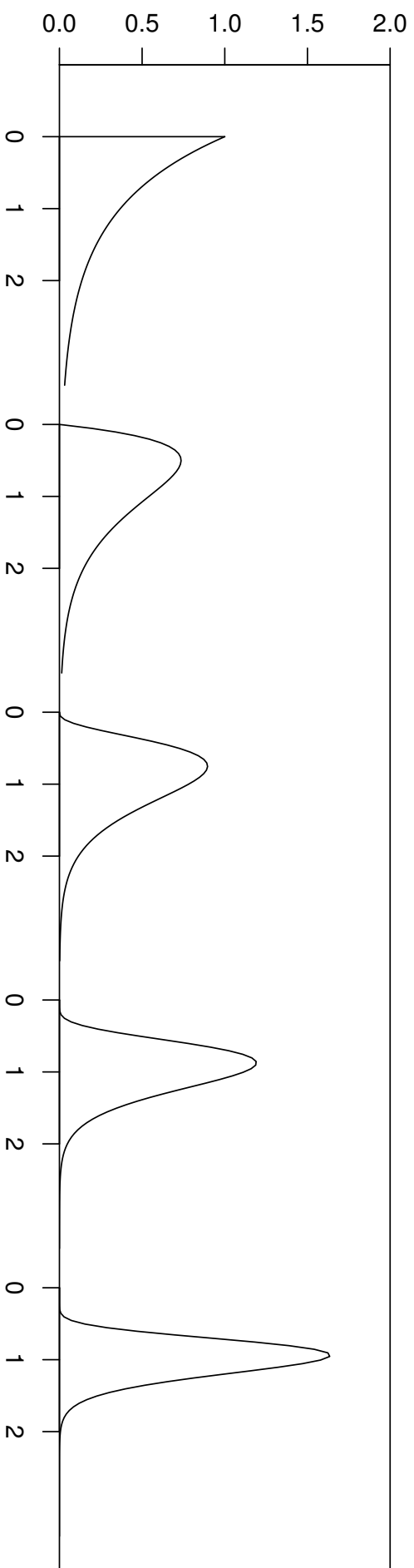
approaches the standard normal distribution as n grows.

$$P \langle Z_n \leq z \rangle \xrightarrow{n \rightarrow \infty} \Phi \langle z \rangle$$

(Φ : cum. distribution function of the standard normal)

... if X_i are independent and equally distributed
and the variance is finite.

b Expressed differently: $\bar{X}_n \approx \sim \mathcal{N} \langle \mu, \sigma^2/n \rangle$



c Generalizations

Many statistics T_n (estimators, test st.) are
 “asymptotically normal”
 for reasonable distributions of the observations

$$X_i \sim F, \text{ indep.}, \quad T_n = T_n \langle X_1, \dots, X_n \rangle$$

$$\implies T_n \approx \sim \mathcal{N} \langle T \langle F \rangle, V \langle T, F \rangle / n \rangle \quad \text{or}$$

$$P \left\langle \frac{(T_n - T \langle F \rangle)}{\sqrt{V \langle T, F \rangle / n}} \leq z \right\rangle \xrightarrow{n \rightarrow \infty} \Phi \langle z \rangle$$

Versions of the CLT even for non-identically distributed and
 dependent X_i .

This is the deeper reason for the dominant role
 of the normal distribution.

d **Central Limit Theorem for functionals**

$$\begin{aligned} T\langle X_1, X_i, \dots, X_n \rangle &\approx T\langle F \rangle + \frac{1}{n} \sum_{i=1}^n \text{IF}\langle X_i; T, F \rangle \\ &\approx \sim \mathcal{N}\langle T\langle F \rangle, V\langle T, F \rangle / n \rangle \end{aligned}$$

$$V\langle T, F \rangle = \text{var}\langle \text{IF}\langle X; F \rangle \rangle$$

We always have $\mathcal{E}\langle \text{IF}\langle X; T, F \rangle \rangle = 0$. $\longrightarrow$

$$V\langle T, F \rangle = \mathcal{E}\langle \text{IF}\langle X; T, F \rangle^2 \rangle.$$

e **Asymptotic variance for M-estimators.**

$$V\langle T, F \rangle = \frac{1}{c^2} \int \psi\langle x, \theta \rangle^2 dF\langle x \rangle$$

$$c = \int \psi\langle x, \theta \rangle s\langle x, \theta \rangle dF_\theta\langle x \rangle$$

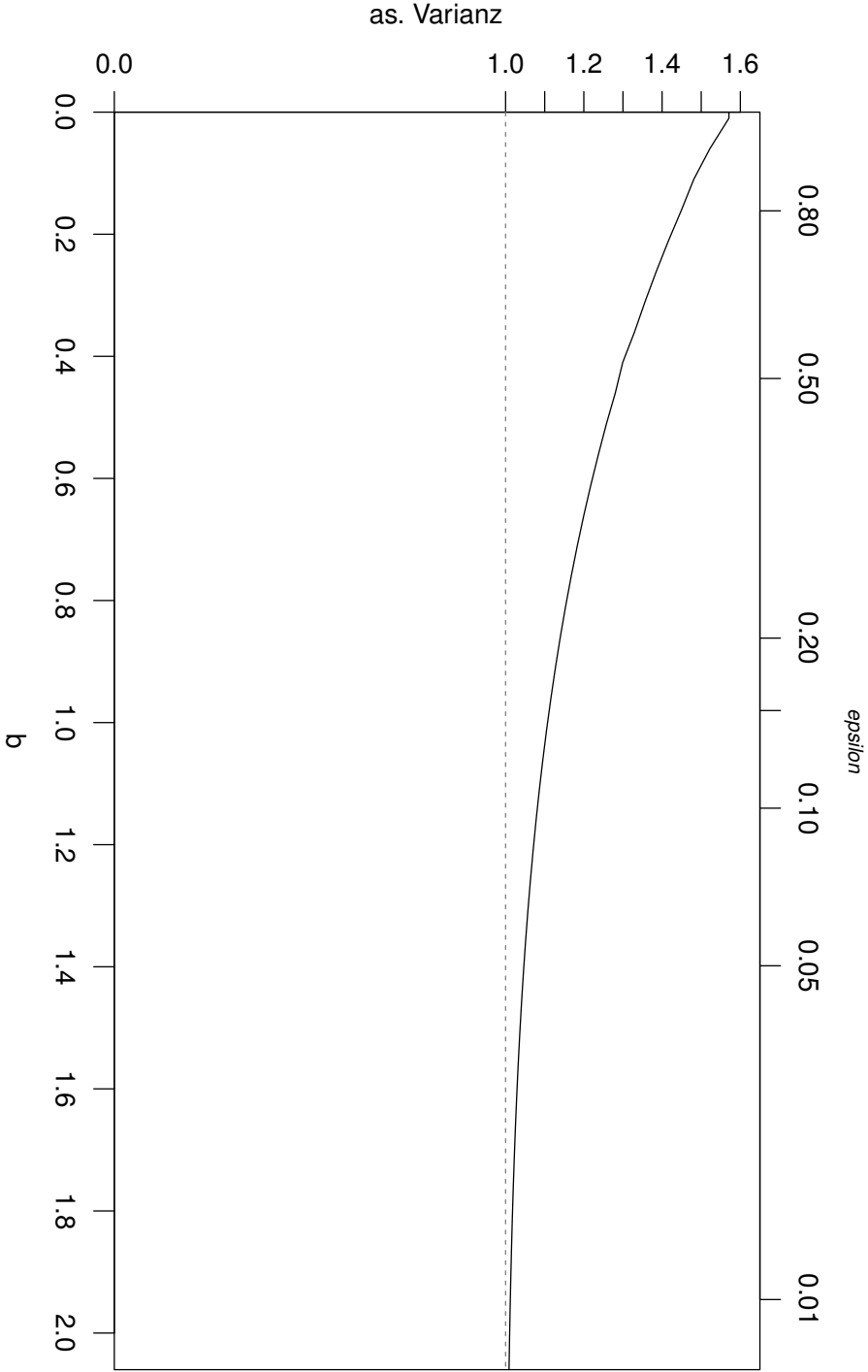
Maximum-Likelihood estimator: $\text{Integral} = c \longrightarrow$

$$V\langle T, F \rangle = 1/c, \quad c = \int s\langle x, \theta \rangle^2 dF_\theta\langle x \rangle.$$

$c =$ **Fisher-Information.**

f **Example Huber estimator** for standard normal distribution

$$V \langle T_H, \Phi \rangle = \int \psi \langle x \rangle^2 d\Phi \langle x \rangle \bigg/ \left(\int \psi' \langle x \rangle d\Phi \langle x \rangle \right)^2$$



g Example Logistic distribution

Maximum-Likelihood estimator for $\mathcal{L}(\mu = 0, \sigma = 1)$

$$\int \frac{(e^{z/2} - e^{-z/2})^2}{(e^{z/2} + e^{-z/2})^4} dz = 0.333$$

= 1 for $\sigma = 1.732$.

h **Maximum likelihood estimator is the best est. asymptotically.**

$V\langle T, F_\theta \rangle \geq 1/I\langle F_\theta \rangle$ for all Fisher consistent estimators T .

- i Precise theories for asymptotic normality:
 - M-estimators: Huber 1964 / 1967
 - Fréchet derivative is sufficient (...) derivative is sufficient
 - ...

j Consider a T_n which has an Influence Function

$\text{IF} \langle x; \{T_n\}, F_\theta \rangle$ at all F_θ and is asymptotically normal.

Then, the M-estimator defined by

$$\psi \langle x, F_\theta \rangle = \text{IF} \langle x; \{T_n\}, F_\theta \rangle$$

is asymptotically equivalent to $\{T_n\}$.

→ **M-estimators are enough for asymptotic considerations!**

k Tests and confidence intervals

based on the asymptotic normal distribution

- Standardized test statistic $T = (\hat{\theta} - \theta_0) / \sqrt{v/n} \approx \Phi$
- Confidence interval for θ : $\hat{\theta} \pm 1.96 \cdot \sqrt{v/n}$.

More on robust tests later.

l Sums of squares

→ chisquared distribution!

$$T = \sum_k (N_k - \mu_k)^2 / \mu_k, \quad \mu_k = \mathcal{E} \langle N_k \rangle$$

$$T \approx \sim \chi^2_\nu$$

Contradiction to the Central Limit Theorem?

5 Optimality Results

5.1 Huber's Minimax Theorem

^a Gross Error Model

Remember:

- “Robustness is the theory of approximate parametric models.”

$$U_\varepsilon \langle F \rangle = \{ (1 - \varepsilon) F \langle \cdot \rangle + \varepsilon H \langle \cdot \rangle \mid H \text{ arbitrary} \}$$

- M-estimators in the form $\int \psi \langle x, T \langle F \rangle \rangle dF \langle x \rangle = 0$ are a very flexible class of estimators.

Every asympt. normal est. is equivalent to an M-est. r.

b The Game

- Nature chooses a distribution F from a set $\mathcal{F}$ of distributions – e.g. $\mathcal{F} = U_\varepsilon \langle F \rangle$.
- The statistician chooses a function $\psi \in \Psi$
- The payoff to the statistician is

$$K \langle \psi, F \rangle = \frac{\int (\psi' \langle x \rangle dF \langle x \rangle)^2}{\int (\psi \langle x \rangle)^2 dF \langle x \rangle}$$

“usually” = $1/V \langle T, F \rangle$.

Statistician chooses minimax strategy:

Maximize the minimal payoff = **minimize the maximal variance**

Assume that Nature gives you the worst distribution.

Choose the best estimator for this distribution.

Make sure that it is at least as good for other distributions.

Need saddlepoint $K \langle \psi_0, F_0 \rangle$ such that

$$K \langle \psi_0, F_0 \rangle \leq K \langle \psi_0, F \rangle \quad \forall F \in \mathcal{F} \text{ and}$$

$$K \langle \psi_0, F_0 \rangle \geq K \langle \psi, F_0 \rangle \quad \forall \psi \in \Psi$$

c **Huber's Theorem (1964, Ann. Math. Statistics)**

Assumptions: $\mathcal{F}$ is convex, has densities,

finite Fisher Information $I \langle F \rangle = \int (f' \langle x \rangle / f \langle x \rangle)^2 f \langle x \rangle dx$

Then,

- (i) If there is $F_0 \in \mathcal{F}$ such that $I \langle F_0 \rangle \leq I \langle F \rangle \quad \forall F \in \mathcal{F}$
 and $\psi_0 \langle \cdot \rangle := f'_0 \langle \cdot \rangle / f \langle \cdot \rangle \in \Psi$, then
 $[\psi_0, F_0]$ is the saddlepoint of the game,

$$K \langle \psi, F_0 \rangle \leq K \langle \psi_0, F_0 \rangle = I \langle F_0 \rangle \leq K \langle \psi_0, F \rangle$$

$$\forall F \in \mathcal{F} \text{ and } \forall \psi \in \Psi.$$

(ii) Conversely, if $[\psi_0, F_0]$ is a saddlepoint, and

$\exists c : c \cdot f'_0/f_0 \in \Psi$, then

$I \langle F_0 \rangle \leq I \langle F \rangle \quad \forall F \in \mathcal{F}$, F_0 is unique

and $\psi_0 = c \cdot f'_0/f_0$.

(iii) ...

d Location Model

Assume: G has convex support, $-\log \langle g \rangle$ is convex.

$T \langle F \rangle$ M-estimator given by ψ : $\int \psi \langle x - T \langle F \rangle \rangle dF \langle x \rangle = 0$

$$\text{asvar}_F \langle T \rangle = \frac{\int \psi \langle x - T \langle F \rangle \rangle^2 dF \langle x \rangle}{\left(\int \psi' \langle x - T \langle F \rangle \rangle dF \langle x \rangle \right)^2}$$

“Neighborhood” $U_\varepsilon \langle F \rangle = \{(1 - \varepsilon)F + \varepsilon H\}$

Loss = asvar. “Saddlepoint” $[\psi_0, \mathcal{F}_0]$:

$\text{asvar}_F \langle T \rangle$ minimax only among dist.s F for which $E_F \langle \psi_0 \rangle = 0$

Solution defines the problem!

→ Restriction to symmetric F and H is natural

... but “non-robust” !

ψ_0 is the bounded version of the max.li. scores

$$\psi_0 \langle x \rangle = \min \langle s \langle x \rangle, \max \langle s \langle x \rangle, -b \rangle, b \rangle$$

$$s \langle x \rangle = -f' \langle x \rangle / f \langle x \rangle .$$

“Tuning constant” b is a function of allowed contamination ε

$$(1 - \varepsilon)^{-1} = \int_{-b \leq f \langle x \rangle \leq b} f \langle x \rangle dx + (f \langle x_0 \rangle + f \langle x_1 \rangle) / b$$

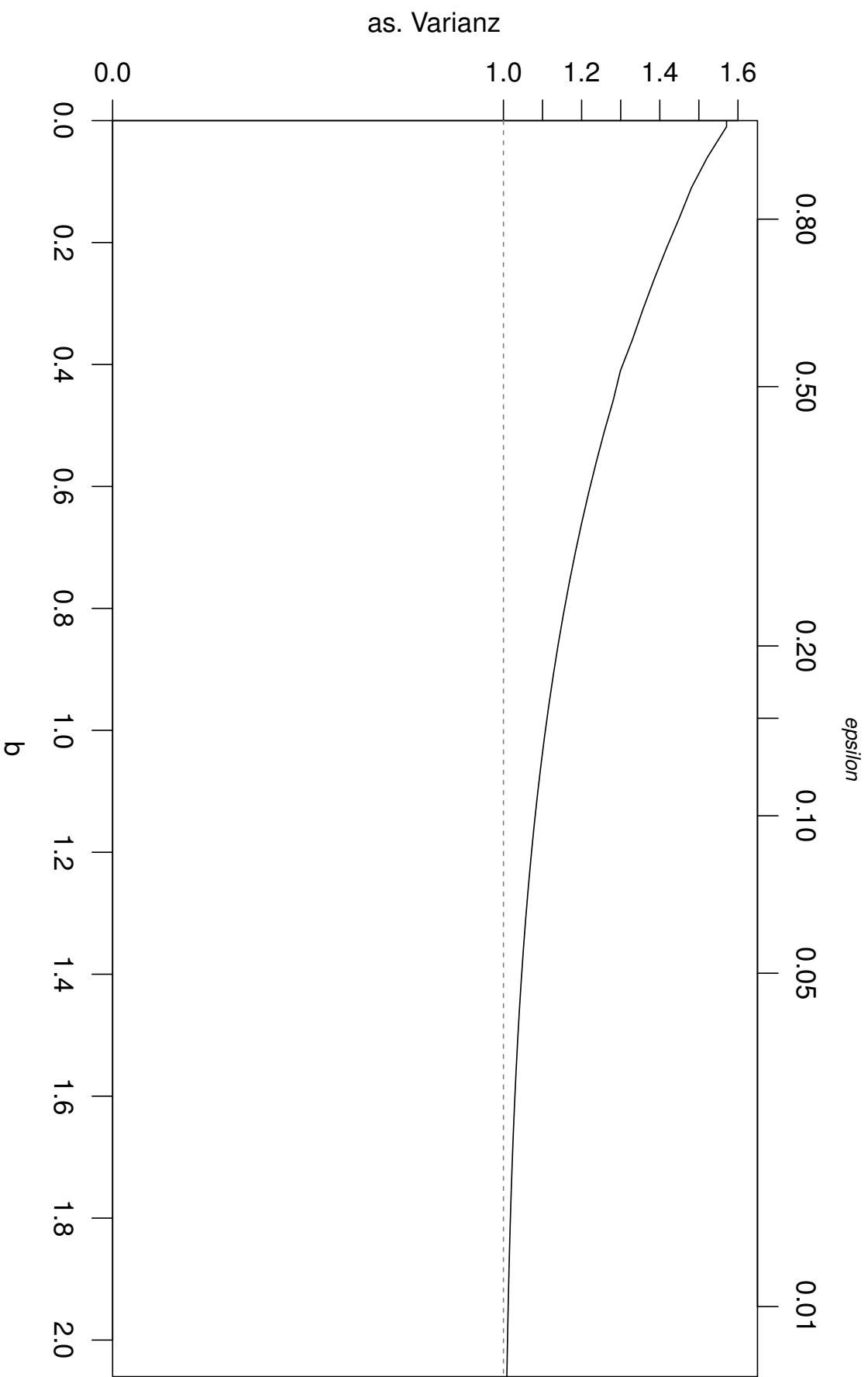
e **Result for the normal location model**

When $\mathcal{F} = U_\varepsilon \langle \mathcal{N} \langle \mu, 1 \rangle \rangle$, then the saddlepoint is given by the Huber function with a tuning constant b that is a function of ε , and

F_0 is the corresponding distribution, see (3.2.e)

Tuning constants and $\text{asvar}_\Phi \langle T \rangle$:

ε	b	asvar_Φ
0.4417	0.5	1.263
0.1428	1.0	1.107
0.0498	1.4	1.047
0.0064	2.0	1.010



f Consequences

- Exact treatment of the robustness question:
“Behavior under the assumption that X_i follow approximately the distribution F ”

- $1/K$ measures the asymptotic variance. What about bias?
Bias dominates for large n .

Asymptotic bias: $T\langle G \rangle - T\langle F \rangle$, $G = \varepsilon F + (1 - \varepsilon)H$.

- Minimax Result for Bias is also treated in Huber (1964)!

Result: ...

- Trick for minimax variance result:
restriction to symmetric distributions !
→ misunderstanding: Robust methods are for symmetric distributions only.
- Asymmetric contamination allowed → “unavoidable bias”
→ Robust estimator defines quantity to be estimated for contaminated distributions
Still want to be (Fisher) consistent = asympt. unbiased at the “nominal model” F_θ .

5.2 Hampel's Optimal Estimators

^a Asymptotic bias

... in gross error "neighborhood":

$$G = (1 - \varepsilon)F + \varepsilon H = F + \varepsilon(H - F)$$

$$\begin{aligned} T\langle G \rangle &= T\langle F \rangle + \varepsilon \int \text{IF}\langle x; T, F \rangle d(H - F)\langle x \rangle \\ &= T\langle F \rangle + \varepsilon \int \text{IF}\langle x; T, F \rangle dH\langle x \rangle - 0 \end{aligned}$$

$$|T\langle G \rangle - T\langle F \rangle| \leq \varepsilon \text{GES}\langle T, F \rangle$$

$$\text{GES}\langle T, F \rangle = \gamma^*_{\langle T, F \rangle} = \sup_x \langle \|\text{IF}\langle x; T, F \rangle\| \rangle$$

Robust estimator $\longrightarrow$ bounded GES.

b Optimality problem

Among all (M-)estimators with $\text{GES} \leq \text{a constant } \gamma$,
find the one that minimizes

$$\text{asvar} \langle T, F \rangle = \int \text{IF} \langle x; T, F \rangle^2 dF \langle x \rangle$$

Search in the class of M-estimators $\longrightarrow$ class of ψ -functions.

= Hampel's optimality problem

$\longrightarrow$ Variational problem

- c Hampel's optimality problem as approximate minimax problem:

$$\begin{aligned}
 \text{asvar}_G \langle T \rangle &\approx \text{asvar} \left\langle \left[T \langle F \rangle + \frac{1}{n} \sum_i \text{IF} \langle X_i; T, F \rangle \right] \right\rangle \\
 &= \int \text{IF} \langle x; T, G \rangle^2 dG \langle x \rangle \approx \int \text{IF} \langle x; T, F \rangle^2 dG \langle x \rangle \\
 &= (1 - \varepsilon) \text{asvar}_F \langle T \rangle + \varepsilon \int \text{IF} \langle x; T, F \rangle^2 dH \langle x \rangle \\
 &\leq (1 - \varepsilon) \text{asvar}_F \langle T \rangle + \varepsilon \gamma^* \langle T, F \rangle
 \end{aligned}$$

d **Hampel's Lemma 5**

Model and assumptions:

- Parametric model $\{F_\theta \mid \theta \in \Theta\}$, Θ open, convex $\subset \mathcal{R}$
- Density $f_\theta > 0$ on set of possible observations x .
- Fix θ_0 . Denote $F_0 := f_{\theta_0}$.
 $s \langle x, \theta_0 \rangle = \partial \langle \log \langle f_\theta \langle x \rangle \rangle \rangle / \partial \theta|_{\theta_0}$ exists, and
- $\int s \langle x, \theta_0 \rangle dF_0 \langle x \rangle = 0$
- Fisher Information $J \langle F_0 \rangle = \int s \langle x, \theta_0 \rangle^2 dF_0 \langle x \rangle$ exists

Lemma 5

Choose $b > 0$.

(a) There exists a constant a such that

$$\widetilde{\psi}\langle x \rangle := \min \langle \max \langle s \langle x, \theta_0 \rangle - a, -b \rangle, b \rangle$$

satisfies $\int \widetilde{\psi} \langle x \rangle \, dF_0 \langle x \rangle = 0$ and

$$d := \int \widetilde{\psi} \langle x \rangle \, s \langle x, \theta_0 \rangle \, dF_0 \langle x \rangle > 0.$$

(Lemma 5, continued) This $\tilde{\psi}$ minimizes

$$\int \psi \langle x \rangle^2 dF_0 \langle x \rangle / \left(\int \psi \langle x \rangle s \langle x, \theta_0 \rangle dF_0 \langle x \rangle \right)^2$$

among all mappings ψ that satisfy

$$\int \psi \langle x \rangle dF_0 \langle x \rangle = 0$$

$$\int \psi \langle x \rangle s \langle x, \theta_0 \rangle dF_0 \langle x \rangle > 0 \quad \text{and}$$

$$\sup_x |\psi \langle x \rangle / \int \psi \langle x \rangle s \langle x, \theta_0 \rangle dF_0 \langle x \rangle| \leq \gamma := b/d$$

Any other solution of this extremal problem coincides with a nonzero multiple of $\tilde{\psi}$ almost everywhere (with respect to F_0).

e

Interpretation

- Like Huber's Theorem, the "Lemma" is about functions. No probabilities, estimators...
Interpretation for statistical purposes comes now.
- Optimality is determined for a fixed θ_0 .
Only makes sense if true for "all" θ
(at least a in neighborhood of θ_0).
- $\int \psi \langle x \rangle dF_0 \langle x \rangle = 0$ means Fisher consistency.
(We need such a condition, since otherwise, the non-random "estimator" $\hat{\theta} \langle F \rangle = \theta$ is optimal under all aspects ...)

- The quantity minimized is $\text{asvar} \langle T \rangle$,
where T is the M-estimator defined by ψ .
- The expression $\sup_x \dots$ is the GES of the estimator
and b/d is the GES of the constructed estimator.
- We would prefer to start with a given bound γ .
But then, the construction is more difficult, because
 γ needs to be large enough for a solution to exist.

f Proof

- a exists since

$\int \min \langle s \langle x, \theta_0 \rangle - a, \max \langle s \langle x, \theta_0 \rangle - a, -b \rangle, b \rangle dF_0 \langle x \rangle$
 is continuous in a and $\rightarrow b > 0$ for $a \rightarrow -\infty$
 and $\rightarrow -b < 0$ for $a \rightarrow \infty$.

- $d > 0$ technical, see Hampel et al. (1986, p.118)

- Optimality: Let ψ be a function satisfying the conditons. Then, we can replace ψ by $d \cdot \psi / \int \psi \langle x \rangle s \langle x, \theta_0 \rangle dF_0 \langle x \rangle$, that is, we can assume $\int \psi \langle x \rangle s \langle x, \theta_0 \rangle dF_0 \langle x \rangle = d$ and need to minimize $\int \psi \langle x \rangle^2 dF_0 \langle x \rangle$. Denote $s \langle x, \theta_0 \rangle - a =: \tilde{s} \langle x \rangle$. Note that

$$\begin{aligned} & \int (\tilde{s} \langle x \rangle - \psi \langle x \rangle)^2 dF_0 \langle x \rangle \\ &= \int \tilde{s} \langle x \rangle^2 dF_0 \langle x \rangle - 2d + \int \psi \langle x \rangle^2 dF_0 \langle x \rangle \end{aligned}$$

Since the first 2 terms on the r.h.s. do not depend on ψ , we minimize the l.h.s.

The integrand is minimized pointwise by $\tilde{\psi}$.

- This also proves uniqueness
(up to values on sets with 0 probability)

g **Approximately minimal MSE**

Mean Squared Error

$$\begin{aligned} \text{MSE} &= \text{bias}^2 + \text{var} \approx (T \langle G \rangle)^2 + n^{-1} \cdot \text{as.var} \langle T, G \rangle \\ &\leq \text{GES} \langle T, F \rangle^2 + n^{-1} \cdot (\text{as.var} \langle T, F \rangle + \dots) \end{aligned}$$

For given ε and n , we can choose b of the Huber est.

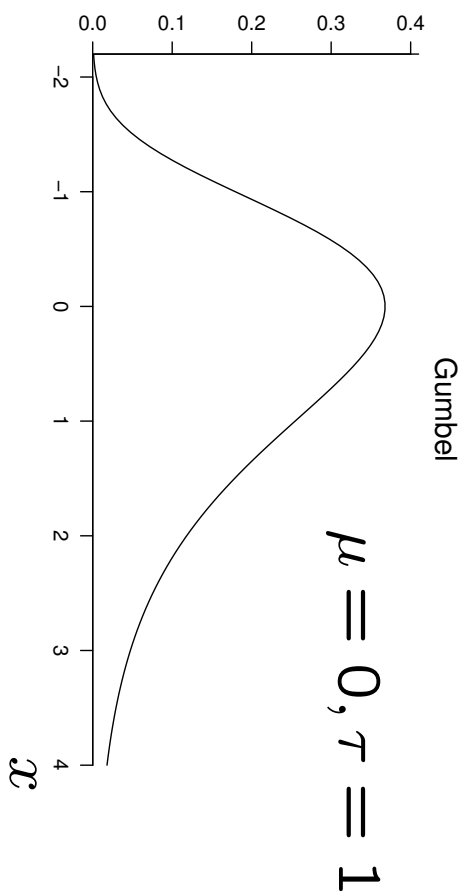
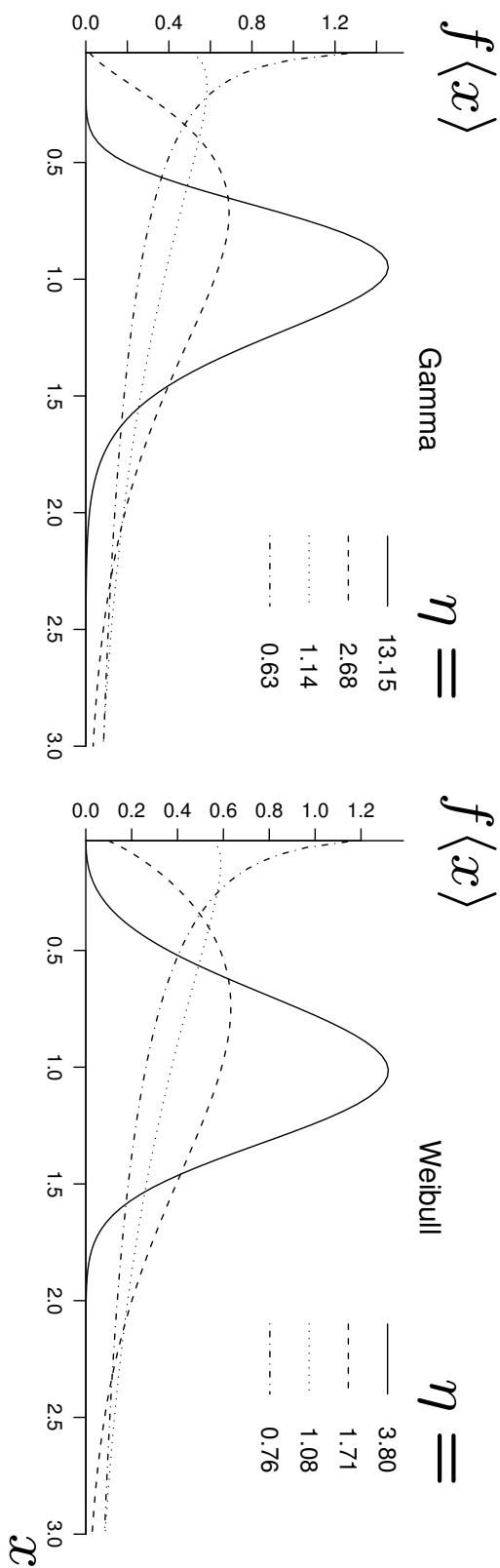
such that this approximate MSE is minimized.

→ “Optimal” choice of the tuning constant, depending on n .

More precise treatment needs “Change of Variance Function” and respective “sensitivity”, see Hampel et al (1986, Ch. 2.7).

h **Example: Log-Weibull distribution**

Weibull: Often used in reliability studies
Failure time or survival time $Y \sim \mathcal{W}$



Failure time or survival time $Y \sim \mathcal{W}$

Then, $X = -\log \langle Y \rangle \sim$ Gumbel or “extreme value” distribution

Cum.d.f. $F_{\mu,\tau} \langle x \rangle = 1 - \exp \langle -e^z \rangle$, $z = \frac{x-\mu}{\tau}$

Density $f_{\mu,\tau} \langle x \rangle = \tau^{-1} e^z \exp \langle -e^z \rangle$

Location-Scale family. Consider fixed scale $\tau = 1$.

Optimal estimators for μ ? $\longrightarrow$ **exercise!**

$\log \langle f_{\mu,1} \langle x \rangle \rangle = z - e^z$.

Scores: $s \langle x, \mu \rangle = e^z - 1$.

Max.li.: $\sum_i e^{x_i - \hat{\mu}} - n = 0 \longrightarrow \exp \langle \hat{\mu} \rangle = \frac{1}{n} \sum_i \exp \langle x_i \rangle$.

Optimal estimators for μ ?

Given b , determine a such that (general formula!)

$$\begin{aligned} & \int \min \langle \max \langle s \langle x, \theta_0 \rangle - a, -b \rangle, b \rangle dF_0 \langle x \rangle \\ &= b (1 - F_0 \langle c_1 \langle a, b \rangle \rangle - F_0 \langle c_0 \langle a, b \rangle \rangle) \\ &\quad - (f_0 \langle c_1 \langle a, b \rangle \rangle - f_0 \langle c_0 \langle a, b \rangle \rangle) \\ &\quad - a (F_0 \langle c_1 \langle a, b \rangle \rangle - F_0 \langle c_0 \langle a, b \rangle \rangle) \\ &= 0 \end{aligned}$$

$$c_0 \langle a, b \rangle = s^* \langle a - b \rangle, \quad c_1 \langle a, b \rangle = s^* \langle a + b \rangle$$

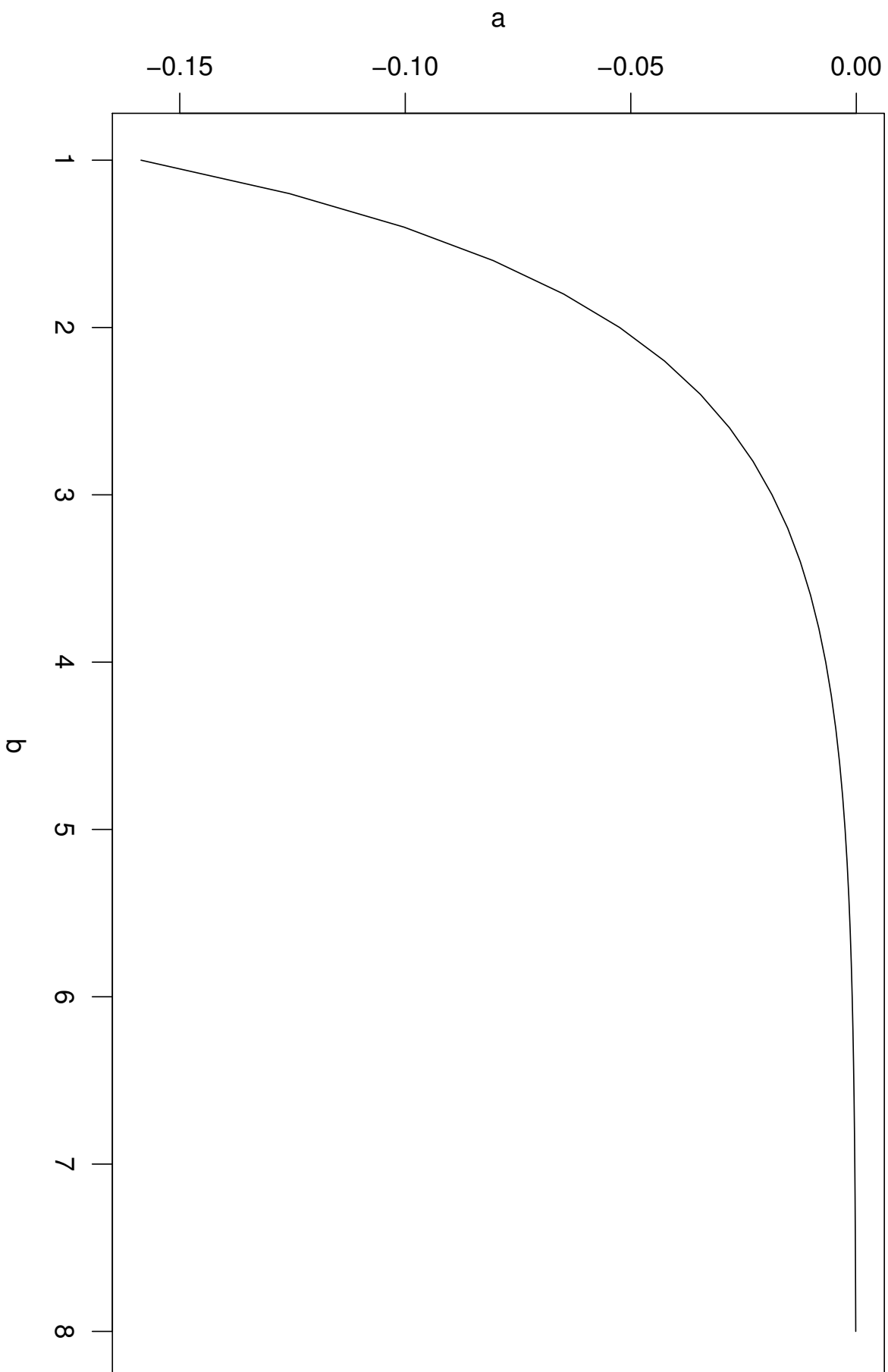
where s^* is the inverse function of $s \langle \cdot, \theta_0 \rangle$.

This defines the function $a \langle b \rangle$.

For Gumbel, $s^* \langle c \rangle = \log \langle 1 + c \rangle$.

Function $a\langle b \rangle$:

```
f.ho.b2a <- function(b, distr="gumbel")
{ ## Purpose:  Hampel optimality: calculate a from b
  ## Arguments: b: chosen bound
  if (distr=="gumbel") {
    fd <- function(x) exp(x-exp(x))
    fp <- function(x) 1-exp(-exp(x))
    fsi <- function(x) log(max(1e-20,1+x))
  } else stop("not programmed for this distr.")
  ## function for uniroot
  ff <- function(a,b) {
    lc0 <- fsi(a-b) ;   lc1 <- fsi(a+b)
    b*(1-fp(lc1)-fp(lc0)) - (fd(lc1)-fd(lc0)) -
      a*(fp(lc1)-fp(lc0))
  }
  rr <- uniroot(ff, c(-1,1) , b=b )
  rr$root
}
```



Estimating function

```
## function that estimates mu from data x
f.estgumbel0 <- function(x, b) {
  la <- f.ho.b2a(b)
  lf.psi <- function(x, a, b) pmin(pmax(exp(x)-1-a, -b), b)
  ff <- function(mu, a, b, x) sum( lf.psi(x-mu, a, b) )
  rr <- uniroot(ff, c(-10,10), a=1a, b=b, x=x)
  rr$root
}
```

Simulation:

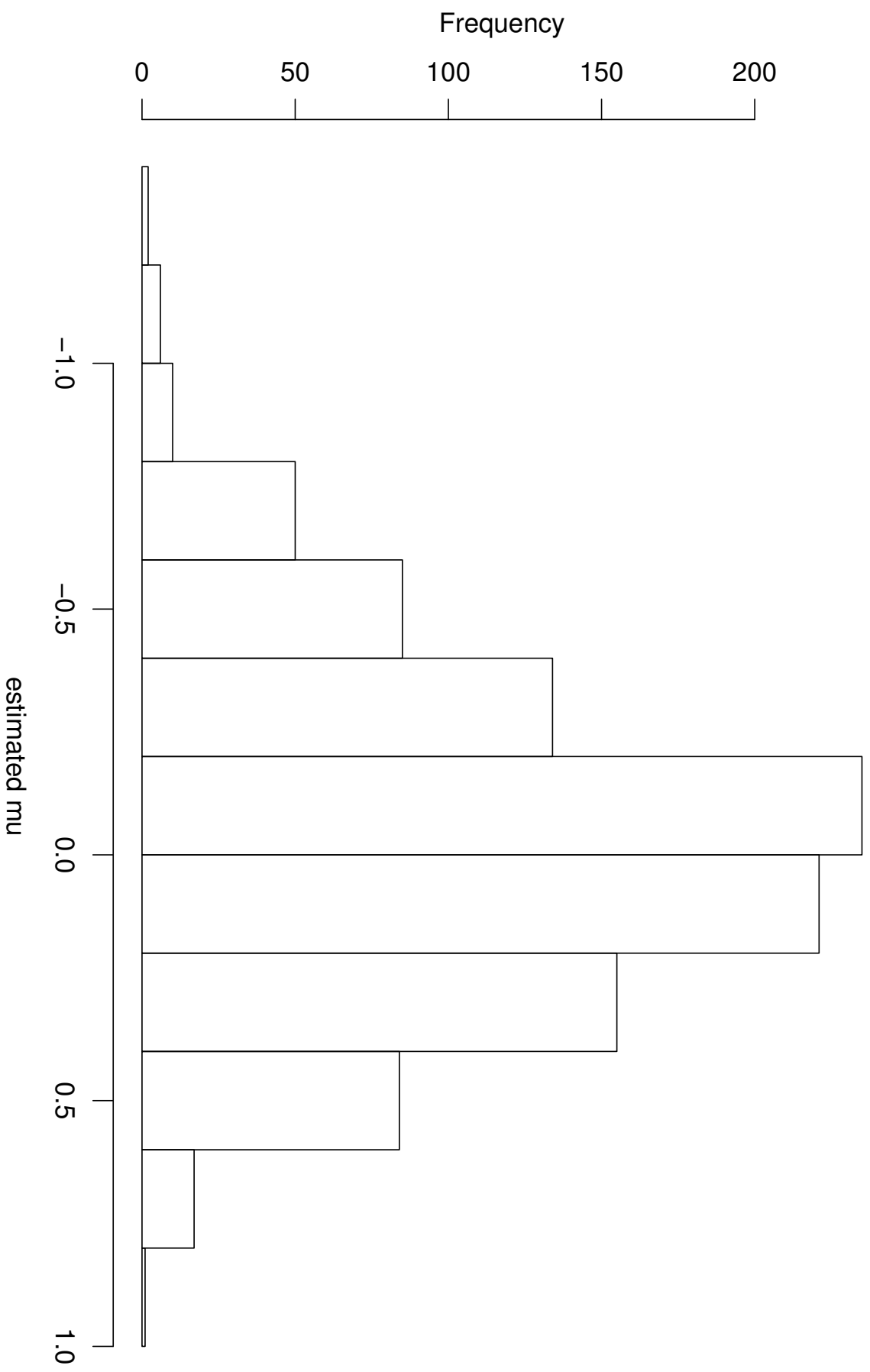
```
## Gumbel quantiles and random numbers
f.qgumbel <- function(p) log(-log(1-p))
f.rgumbel <- function(n) f.qgumbel(runif(n))

## simulate distribution of the estimator
f.singumbel <- function(n=20, nrep=100, b=2) {
  lx <- matrix(f.rgumbel(n*nrep), n)
  apply(lx, 2, f.estgumbel0, b=b)
}

r.simest <- f.singumbel(n=t.n, 1000, b=2)
c(n=t.n, expec=mean(r.simest), sd=lsd <- sd(r.simest),
  se=lsd/sqrt(length(r.simest)-1))
```

n	expec	sd	se
10.0000	-0.0345	0.3508	0.0111

Simulated distr. of opt. M-est., $b = 0.5$, $n = 10$



6 Breakdown and Maximal Bias

6.1 Asymmetric Contamination, Definitions

a More realistic contamination

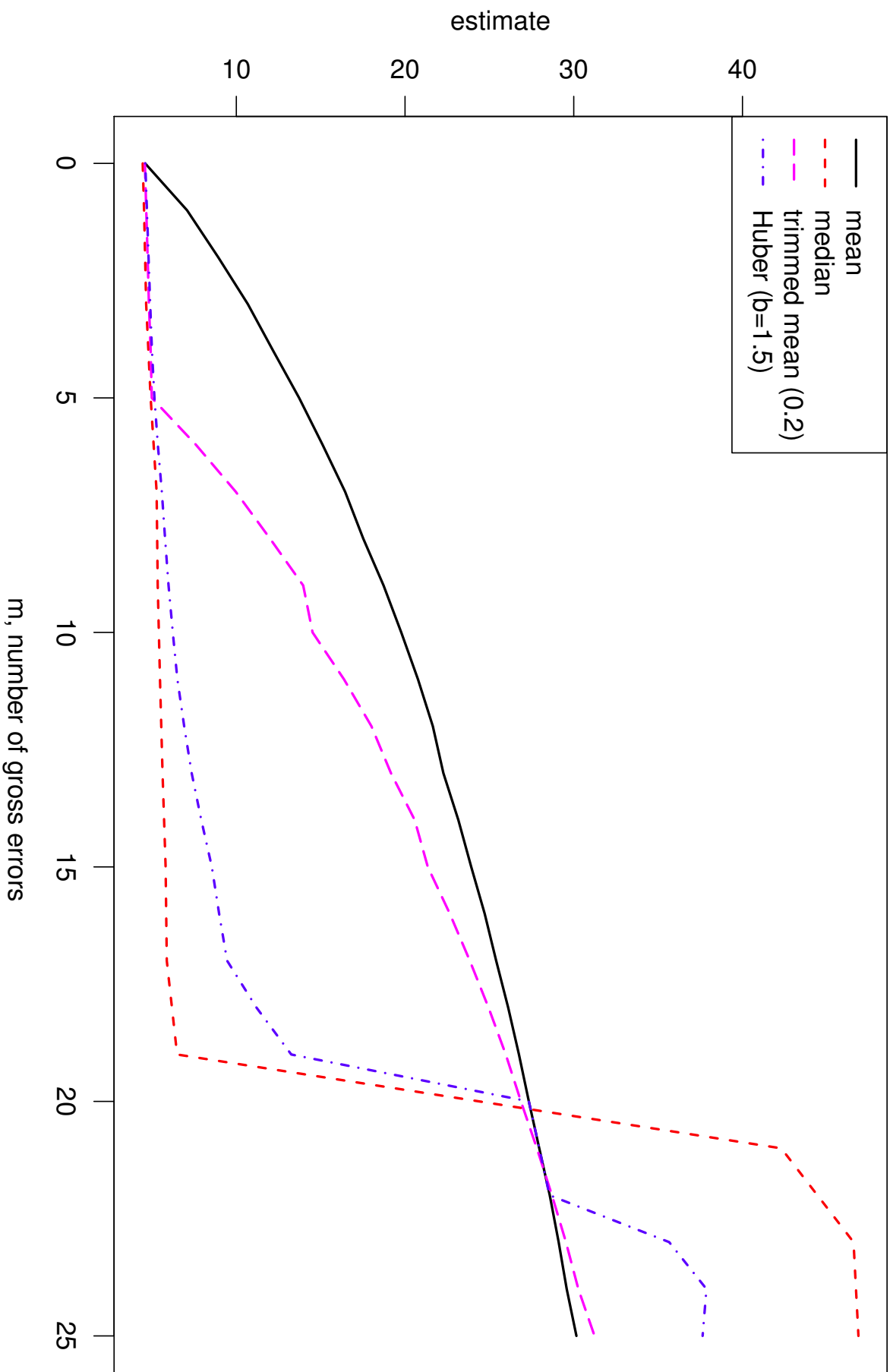
- Asymmetric contamination
- More than “infinitesimal”!

Example: $n = 20$ observations, $X_i \sim \mathcal{N}\langle 5, 1 \rangle$.

Contamination: add m observations $Y_j \sim \mathcal{N}\langle 50, 3^2 \rangle$.

What happens to $T\langle X_1, \dots, X_{20}, Y_1, \dots, Y_m \rangle$?

20 obs. from $N(5, 1^2)$ + m obs. from $N(50, 3^2)$



b Asymptotic version

$n \rightarrow \infty$, $m = \varepsilon(n + m)$, $\mathcal{N} \langle 20, 5^2 \rangle$ replaced by arbitrary H
 $T \langle X_1, \dots, X_n, Y_1, \dots, Y_m \rangle \rightarrow T \langle (1 - \varepsilon)F + \varepsilon H \rangle$ (as before)

Maximal Bias

$$\text{MB} \langle \varepsilon; T, F \rangle = \sup \langle |T \langle (1 - \varepsilon)F + \varepsilon H \rangle - T \langle F \rangle| | H \text{ arbitrary} \rangle$$

c **What did we want to estimate?**

Clear for symmetric distributions

General: we estimate the functional at the central model F

Estimation of the parameter θ in a param.model $[F_\theta]$:
possibly need corrected estimator:

Let $g\langle\theta\rangle = T\langle F_\theta\rangle$. Replace T by $\tilde{T}\langle F\rangle = g^{-1}\langle T\langle F\rangle\rangle$.

Correction avoids asymptotic bias for “central model” F , but
under contamination bias is unavoidable

MB curve shows worst case.

d Breakdown

For large ε , $\text{MB}\langle \varepsilon; T, F \rangle = \infty$. Unavoidable!

Consider, for location model $\mathcal{N}\langle \mu, 1 \rangle$, contamination by $\mathcal{N}\langle \tilde{\mu}, 1 \rangle$

For $\varepsilon > 0.5$, $T\langle F \rangle$ must be nearer to $\tilde{\mu}$ than to μ .

Worst case is then $\tilde{\mu} \rightarrow \infty \implies T\langle F \rangle \rightarrow \infty$

Breakdown point:

$$\varepsilon^*\langle T, F \rangle := \sup\langle \varepsilon | \text{MB}\langle \varepsilon; T, F \rangle < \infty \rangle$$

Breakdown point:

$$\varepsilon^* \langle T, F \rangle := \sup \langle \varepsilon | \text{MB} \langle \varepsilon; T, F \rangle < \infty \rangle$$

General definition:

- Θ an open set of parameters θ . $T \langle F \rangle \in \Theta$.
No “Breakdown”: $T \langle G \rangle$ is in a compact subset of Θ
for all G in neighborhood
- Contamination = gross error model
replaced by “Prokhorov neighborhood”

e

Relation to the Gross Error Sensitivity

If the linearization

$$T \langle (1 - \varepsilon)F + \varepsilon H \rangle \approx T \langle F \rangle + \varepsilon \int \text{IF} \langle x; T, F \rangle \, dH \langle x \rangle$$

is valid, we get

$$\begin{aligned} |T \langle (1 - \varepsilon)F + \varepsilon H \rangle - T \langle F \rangle| & \\ & \approx \leq \varepsilon \int |\text{IF} \langle x; T, F \rangle| \, dH \langle x \rangle \leq \varepsilon \max(|\text{IF}|) \end{aligned}$$

$$\text{MB} \langle \varepsilon; T, F \rangle \approx \varepsilon \gamma^* \langle T, F \rangle$$

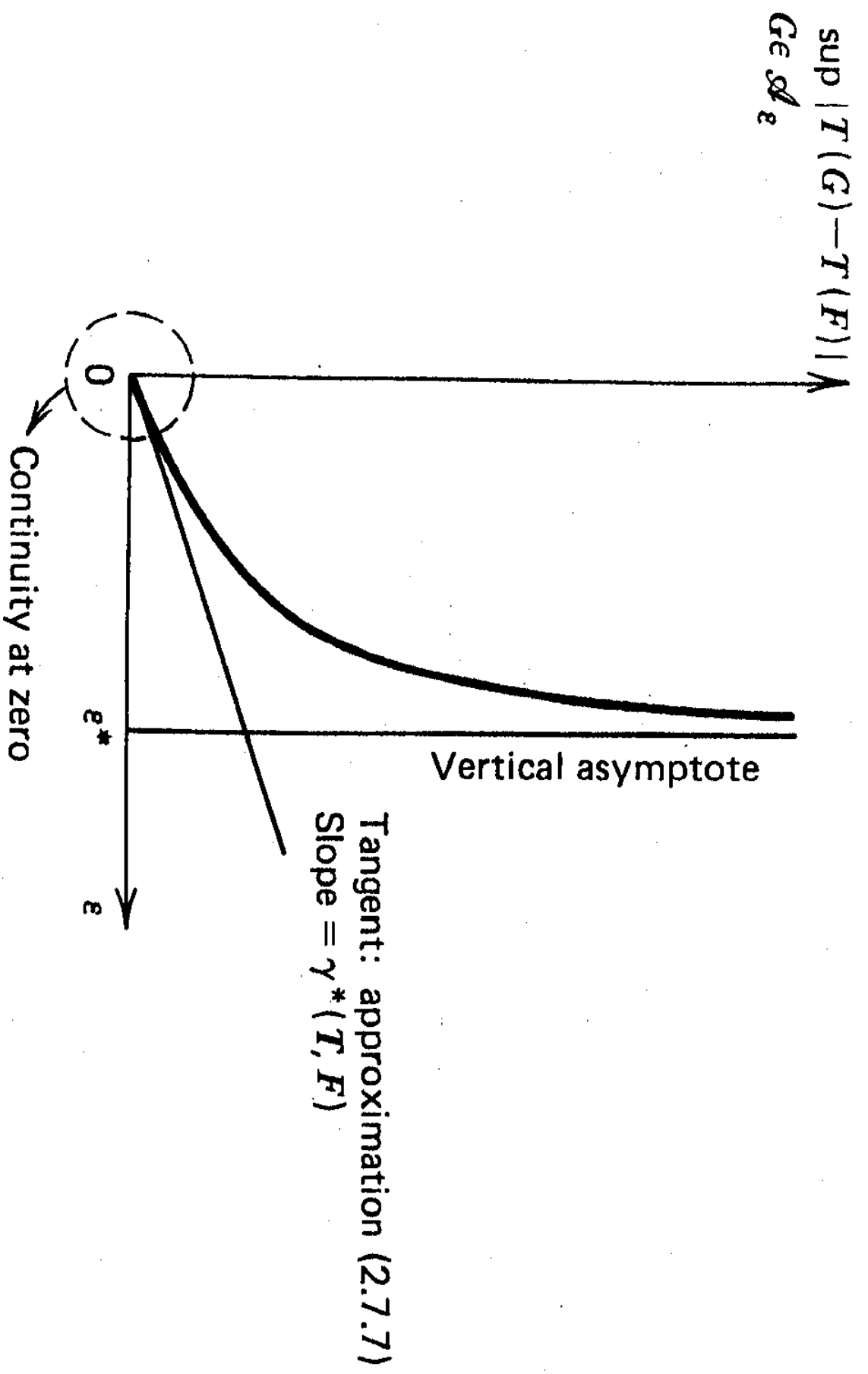


Figure 1. Plot of $\sup\{|T(G) - T(F)|; G \in \mathcal{A}_\varepsilon\}$ as a function of ε .

f Back from ∞ . $\longrightarrow$ **Finite sample Breakdown Point.**

Sample $x_1, \dots, x_n$.

“Corrupt” the sample:

Replace $m = \varepsilon n$ observations by arbitrary values $\longrightarrow$
 $z_1, \dots, z_n$.

Finite sample breakdown point:

$$\varepsilon_n^* \langle T; x_1, \dots, x_n \rangle := \max \left\langle \frac{m}{n} \mid \sup \langle \|T \langle z_1, \dots, z_n \rangle \| \rangle < \infty \right\rangle$$

sup over all possible corrupted samples

Often, finite sample and asymptotic breakdown points coincide.

6.2 Bias and Variance

a **Minimax as. Bias**

Huber (1964): Minimax solution is the median.

b **Combining Bias and Variance**

$\text{MSE} = \text{bias}^2 + \text{variance}$

$$\text{MSE}_n \approx \text{asBias}^2 \langle F \rangle + \text{asvar} \langle F \rangle / n$$

As $n \rightarrow \infty$, bias becomes more important than variance.

Huber (1964), Part 7:

For $\varepsilon = 0.1 \rightarrow c = 1.14$, $\text{max.bias} = 0.17$, $n_r = 9$,
 for $\varepsilon = 0.01 \rightarrow c = 1.95$, $\text{max.bias} = 0.021$, $n_r = 600$,
 for $\varepsilon = 0.001 \rightarrow c = 2.64$, $\text{max.bias} = 0.0027$, $n_r = 35000$.
 n_r : bias is half the standard deviation.

7 Scale Estimation

7.1 M-estimators of Scale

a The Scale Model

$$X = \sigma Z, \quad Z \sim F_1$$

F_1 : Distribution with scale parameter 1. Defines σ .

b Log Scale

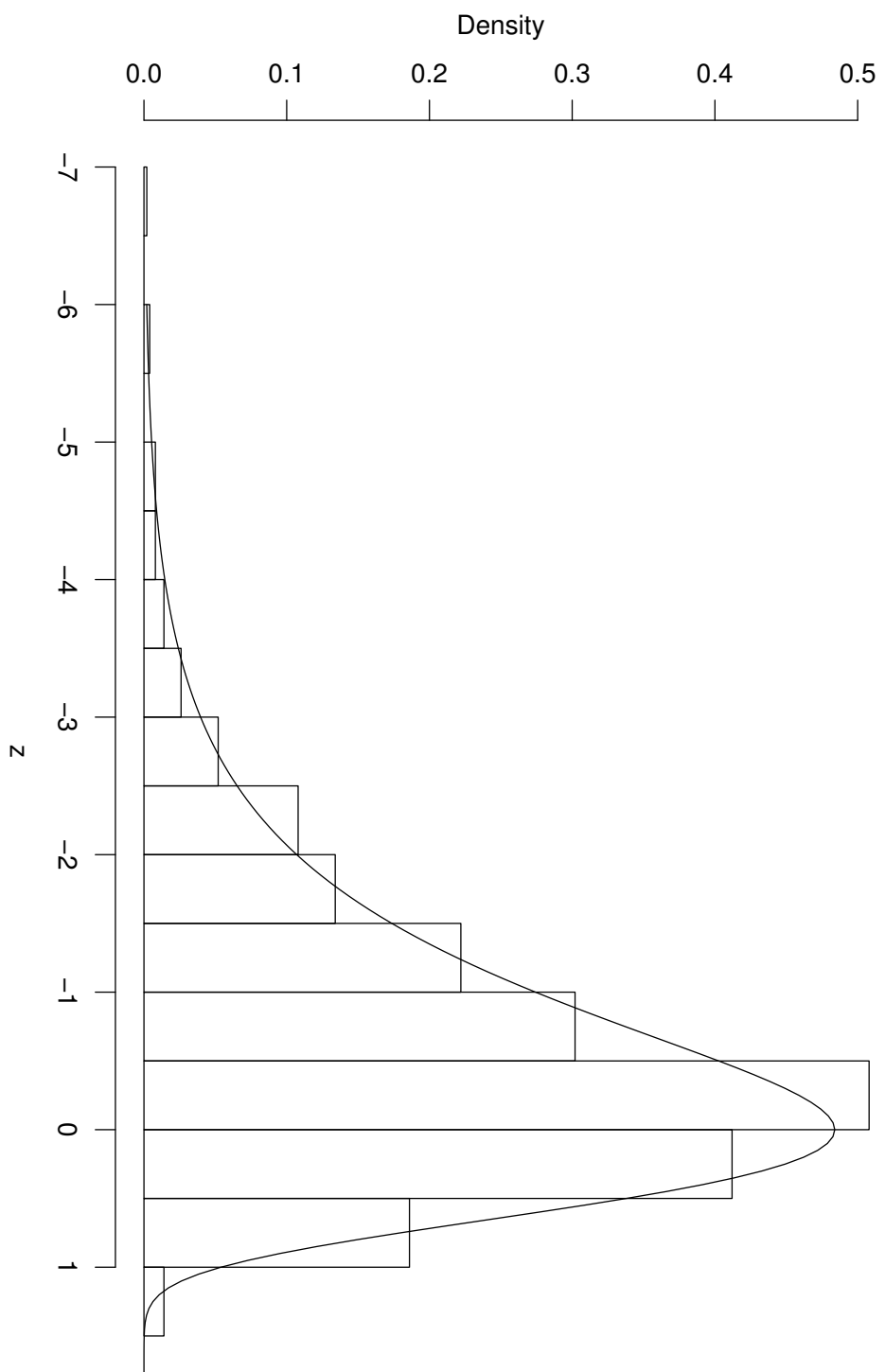
Transform the observation and the parameter!

$$\log \langle ||X|| \rangle = \tilde{\mu} + \log \langle ||Z|| \rangle, \quad \tilde{\mu} = \log \langle \sigma \rangle, \quad \log \langle ||Z|| \rangle \sim \tilde{F}_0$$

→ Location model. $\tilde{F}_0$ long-tailed to the left.

$Z \sim \mathcal{N} \langle 0, 1 \rangle \rightarrow$ density of $Y = \log \langle ||Z|| \rangle$:

$$\tilde{f}_0 \langle y \rangle = c \exp \langle -e^y / 2 \rangle e^y, \quad c = (2\pi)^{-1/2}$$



c **M-Estimators**

(back to scale, unlogged) Max.li. for σ , $F_1 = \Phi = \mathcal{N}\langle 0, 1 \rangle$:

$$\sum_i \left((x_i / \hat{\sigma})^2 - 1 \right) = 0 \quad \hat{\sigma}^2 = \frac{1}{n} \sum_i x_i^2$$

Cut large terms!

$$\sum_i \left(\psi \langle x_i / \hat{\sigma} \rangle^2 - \kappa \right) = 0$$

ψ : Huber's ψ .

General:

$$\sum_i \chi \langle x_i / \hat{\sigma} \rangle = 0, \quad \int \chi \langle x / \hat{\sigma} \rangle \langle F \rangle \, dF \langle x \rangle = 0$$

κ ?

We want Fisher consistency! $\hat{\sigma} \langle F_{\sigma} \rangle = \sigma$

$$\int \chi \langle x / \hat{\sigma} \rangle \, dF_{\hat{\sigma}} \langle x \rangle = \int \chi \langle z \rangle \, dF_1 \langle z \rangle = 0$$

$$\int (\psi^2 \langle z \rangle - \kappa) \, dF_1 \langle z \rangle = 0$$

$$\longrightarrow \quad \kappa = \int \psi^2 \langle z \rangle \, dF_1 \langle z \rangle$$

... usually for $F_1 = \Phi$.

d Influence Function

General result for M-estimators:

$$\text{IF} \langle x; F \rangle = \frac{1}{c} \psi \langle x, \theta \rangle \quad c = - \int \frac{\partial}{\partial \theta} \psi \langle x, \theta \rangle f_{\theta} \langle x \rangle dx$$

$$\psi \langle x, \theta \rangle = \chi \langle x/\sigma \rangle \quad \longrightarrow \quad \frac{\partial}{\partial \theta} \psi \langle x, \theta \rangle = -\chi' \langle x/\sigma \rangle \quad x/\sigma^2$$

$$\text{IF} \langle z; F_1 \rangle = \frac{1}{c} \chi \langle z \rangle \quad c = \int \chi' \langle z \rangle f_1 \langle z \rangle dz$$

$$\text{IF} \langle x; F_{\sigma} \rangle = \sigma \text{IF} \langle x/\sigma; F_1 \rangle$$

e

Weighted Scale

New!

$$\chi \langle z \rangle = \omega^2 \langle z \rangle (z^2 - \kappa)$$

κ such that

$$\int \omega^2 \langle z \rangle (z^2 - \kappa) dF_1 \langle z \rangle = 0$$

Advantage: Allows for easy definition of “redescenders”:

If $\omega \langle z \rangle \rightarrow 0$ for $z \rightarrow \infty$, then outliers have vanishing influence
(since $IF \propto \chi$)

f **Log Scale**

Again: Transform the observation and the parameter!

$$\log \langle ||X|| \rangle = \mu + \log \langle ||Z|| \rangle, \quad \mu = \log \langle \sigma \rangle, \quad \log \langle ||Z|| \rangle \sim \tilde{F}_0$$

Location model. $\tilde{F}_0$ long-tailed to the left.

7.2 Other Scale Estimators

a Inter-Quartile Range

$$\text{IQR} = \gamma(q_{0.75} - q_{0.25}) = \gamma(F^{-1}\langle 0.75 \rangle - F^{-1}\langle 0.25 \rangle)$$

γ to make it Fisher-consistent,

$$\gamma^{-1} = \Phi^{-1}\langle 0.75 \rangle - \Phi^{-1}\langle 0.25 \rangle = 1.35.$$

b MAD

Median Absolute Deviation

$$\text{MAD} = \gamma \text{median} \langle ||X - \text{median} \langle X \rangle|| \rangle$$

γ to make it Fisher-consistent,

$$\gamma^{-1} = \Phi^{-1}\langle 0.75 \rangle = 0.6745. \quad \gamma = 1.483.$$

c Qn

Rousseeuw & Croux (1993). Basic idea: “Elemental subsets”.
 Minimal set of obs. that allows to “estimate” a scale:
 Pairs of observations.

$\hat{\sigma}_{ij} = |x_j - x_i|$. Ask for “robust consensus”.

$$\hat{\sigma} = \gamma \operatorname{median}_{i < j} \langle |x_j - x_i| \rangle$$

Breakdown point = 0.29.

→ Use a lower quantile to get breakdown point 0.5!

Let $h = [n/2] + 1$,

$$Q_n = \gamma(|x_j - x_i|)_{[k]}, \quad k = \binom{h}{2} \approx \binom{n}{2} / 4$$

$$\gamma^{-1} = \sqrt{2} \Phi^{-1} \langle 5/8 \rangle, \quad \gamma = 2.222.$$

7.3 Location and Scale

a The model

$$X = \mu + \sigma Z, \quad Z \sim F_0$$

$$F\langle x; \mu, \sigma \rangle = F\langle (x - \mu)/\sigma; 0, 1 \rangle = F_0\langle (x - \mu)/\sigma \rangle$$

$$\begin{aligned} f\langle x; \mu, \sigma \rangle &= \frac{1}{\sigma} f\langle (x - \mu)/\sigma; 0, 1 \rangle = \frac{1}{\sigma} f_0\langle (x - \mu)/\sigma \rangle \\ &= \frac{1}{\sigma} \exp\langle -\rho\langle (x - \mu)/\sigma \rangle \rangle \end{aligned}$$

$$\text{Normal distribution } F_0 = \Phi: \rho\langle z \rangle = z^2/2$$

b Two parameters

→ M-estimators: 2 equations, simultaneous

- Max.li. for long-tailed distribution: Minimize

$$\sum_i (-\log \langle \sigma \rangle - \rho \langle (x_i - \mu) / \sigma \rangle)$$

or solve 2 scores equations. Let $\psi = \rho'$.

$$\begin{aligned} \psi \langle x; \mu, \sigma \rangle &= s_\mu \langle x; \mu, \sigma \rangle = \frac{1}{c\sigma} \psi \langle z \rangle & z &= (x - \mu) / \sigma \\ \chi \langle x; \mu, \sigma \rangle &= s_\sigma \langle x; \mu, \sigma \rangle = \frac{1}{c\sigma} \psi \langle z \rangle \cdot z - \frac{1}{\sigma} \end{aligned}$$

$$\text{Solve } \sum_i \psi \langle (x_i - \mu) / \sigma \rangle = 0, \quad \sum_i \chi \langle (x_i - \mu) / \sigma \rangle = 0$$

If ψ is monotone, $\psi \langle z \rangle \cdot z \rightarrow \infty$ for $z \rightarrow \infty$

→ not robust

If $\psi \langle z \rangle \cdot z$ bounded → ψ “redescending”

→ may lead to non-unique solutions or inefficient est. of μ

- Proposal 2 (Huber, 1964)

$$\sum_i \psi \langle (x_i - \mu) / \sigma \rangle = 0$$

$$\sum_i \psi^2 \langle (x_i - \mu) / \sigma \rangle = n \int \psi^2 \langle z \rangle d\Phi \langle z \rangle$$

No longer Max.li. for any distribution.

→ “ ψ style” of M-estimators is more flexible!

- Use weighted scale! $\chi \langle z \rangle = \omega^2 \langle z \rangle (z^2 - \kappa)$

- c Experience of Princeton simulation study (1970):
Need for scale estimation can seriously affect
the robustness properties of the location estimator

→ If location is of primary interest, and
scale is a “nuisance” parameter, then
estimate scale as robustly as possible → MAD

When it comes to **testing**, e.g. $\mu = \mu_0$, a scale est. is needed.

Should be efficient for achieving a good power!

- Better scale estimator needed. Weighted scale.
- later! (when we treat regression)

8 Historical Remarks and Alternative Approaches

8.1 Overview

(This section, 2 slides, has been included in the Introduction.)
See Hampel et al (1986): Introduction by F. Hampel

Outliers

Rejection of outliers: Daniel Bernoulli (1770), astronomers

Trimmed mean: 1821

Rejection rules: Pierce (1852),:

Test extreme observations (largest & smallest) for
being an outlier,

Reject if test is significant, average the remaining ones.

Robustness of Tests

Fisher's exact theory for the normal distr. → "monopoly".
E.S. Pearson (1931): Non-robustness of F-test for variances
→ Focus on robustness of the level,
sometimes neglecting power

Non-parametric tests: Power of Wilcoxon test is very good!

Robust estimation

Non-robustness of the mean: Tukey (1960)

Minimax theory: Huber (1964)

Influence Function and Breakdown point: Hampel (1968)

8.2 Other Approaches and Results

a Rejection rules

see above

b L-estimators

Prototype: Trimmed mean

Weighted average, weight based on rank:

$$T_n = \sum_i a_i X_{[i]}, \quad X_{[1]}, \dots, X_{[n]} \text{ ordered}, \quad \sum_i a_i = 1$$

Functional: Define a_i by a function $h : [0, 1] \rightarrow \mathbb{R}$

$$a_i = \int_{[(i-1)/n, i/n]} h\langle z \rangle dz / \int_{[0,1]} h\langle z \rangle dz$$

$$T\langle F \rangle = \frac{\int x h\langle F\langle x \rangle \rangle dF\langle x \rangle}{\int h\langle z \rangle dz}$$

c **R-estimators**

Prototype:

Hodges-Lehmann estimator, derived from Signed Rank Test

Test statistic $\longrightarrow$ p-value. Choose μ such that p-value=1.

$\longrightarrow \hat{\mu} = \text{med}_{h \leq i} \langle (X_h + X_i)/2 \rangle$, median of “Walsh averages”

General form: For a given skew symmetric function $h \langle \cdot \rangle$,

$T \langle F \rangle$ is solution of

$$\int h \langle \tfrac{1}{2} F \langle y \rangle + \tfrac{1}{2} (1 - F \langle 2t - y \rangle) \rangle dF \langle y \rangle = 0$$

for t . The integral gets 0 when the sample $X_1, \dots, X_n$ and its mirror image about t , $2t - X_1, \dots, 2t - X_n$ have their ranks “mixed” as well as possible.

d **Adaptive estimators**

For $n \rightarrow \infty$, the distribution can be estimated from the data
→ asymptotically fully efficient estimators!

Need large samples.

e **Robustness towards dependence**

long range dependence: Hampel (1987), Beran (1992), ...

f **Other Aspects of Robustness**

Robustified Likelihood Ratio Test, Confidence Int.: Huber (1968)

Capacities: Huber & Strassen (1973)

g

Further Developments

Princeton Robustness Year 1970:

Simulation for many location and scale estimators

First results on regression: Huber (1971)

Covariance Matrix: Maronna (1976)

Time Series: Martin (1979), Künsch (1981)

Robust tests (infinitesimal): Ronchetti (1980)

Combinatorial estimators: Rousseeuw and his students

h

Main Research Groups

- Tukey, Bell Labs & Princeton
→ St. Morgenthaler (Lausanne), David Tyler (Rudgers)
- Huber and Hampel in Zurich → Alfio Marazzi (Lausanne)
Ronchetti, Rousseeuw, Stahel, Künsch, Mächler (ETH)
M-estimation, Hampel's Approach "based on the Influence Function"
- Elvezio Ronchetti, Geneva → Maria-Pia Victoria-Feser,
Eva Cantoni, Stephane Heritier
- Peter Rousseeuw, Belgium → Mia Hubert, C. Croux

- Argentina: Victor Yohai, Ricardo Maronna (& Doug Martin, Seattle)
High breakdown est., Minimax Bias, sound theory!
→ Book
- Penn State: Hettmansperger → Shrader & McKean (Aus)
R-estimators
- Oja, Finland (connected to Berne): Multivariate combinatorial, nonparametric
- Chicago: Portnoy, Koenker, X. He
Regression Quantiles, sound theory
- ...

i **Books**

Huber (1981 (2009))

Hampel, Ronchetti, Rousseeuw, Stahel (1986)

Rousseeuw & Leroy (1987): Regression

Maronna, Martin, Yohai (2006)

Conferences

Robustness Year in Princeton (1970)

Minneapolis (1989)

L_1 norm (Neuchatel, 1987 ff)

Int. Conference on Robust Statistics (ICORS)

9 Multidimensional Parameter

9.1 Generalization of Basic Notions

- a We have treated mostly a single parameter (location)
sometimes in presence of a nuisance parameter (scale)

Most notions generalize easily to the multidimensional case.

Since we will deal with regression in an example and later on,

I will call the random variable Y ,

and allow it to be multivariate: $\underline{Y} = [Y^{(1)}, \dots, Y^{(j)}, \dots, Y^{(m)}]^T$.

b **Model**

Parametric family $F_{\underline{\theta}}$ with density $f(\underline{y}; \underline{\theta})$

$\underline{y}$: multivariate, $\underline{\theta}$: multidimensional parameter

Sample: $\underline{Y}_1, \dots, \underline{Y}_n$ i.i.d. $\sim F_{\underline{\theta}}(\underline{y})$ (should be $\mathcal{F}$)

c **Multivariate Location**

“Example”! Multivariate Normal Distr. with given Cov.mtx.

$$\underline{Y}_i \sim \mathcal{N}_m(\underline{\mu}, \mathbf{I})$$

$$f(\underline{y}; \underline{\mu}) = (2\pi)^{m/2} \exp\left(-\frac{1}{2}(\underline{y} - \underline{\mu})^T(\underline{y} - \underline{\mu})\right)$$

(Covariance matrix: Next Chapter)

- d Will also look at non-iid. observations, $Y_i \sim \mathcal{F}_{i,\underline{\theta}}$,
e.g. (simple) regression,

$$Y_i \sim \mathcal{N}(\eta_i, \sigma^2) \quad \eta_i = \beta_0 + \beta_1 x_i$$

- x_i not random, but fixed
(sometimes treated as random later on).

e

Influence Function, Asymptotic Distribution

Functional $\underline{T}\langle F \rangle$,

Influence function IF $\langle \underline{y}; \underline{T}, F \rangle = \lim_{\varepsilon \rightarrow 0} \frac{\underline{T}\langle (1-\varepsilon)F + \varepsilon \Delta_{\underline{y}} \rangle - \underline{T}\langle F \rangle}{\varepsilon}$

Asymptotic distr.

$$\begin{aligned} \underline{T}\langle \hat{F}_n \rangle &\approx \mathcal{N}\langle \underline{T}\langle F \rangle, V\langle \underline{T}, F \rangle / n \rangle \\ V\langle \underline{T}, F \rangle &= \int \text{IF}\langle \underline{y}; \underline{T}, F \rangle \text{IF}\langle \underline{y}; \underline{T}, F \rangle^T dF\langle \underline{y} \rangle \end{aligned}$$

f Maximum Likelihood

Minimize $\sum_i \rho \langle \underline{y}_i, \underline{\theta} \rangle, \quad \rho \langle \underline{y}, \underline{\theta} \rangle = -\log \langle f \langle \underline{y}; \underline{\theta} \rangle \rangle$

(“Efficient”) **Scores:**

$$s_k \langle \underline{y}, \underline{\theta} \rangle = \frac{\partial}{\partial \theta_k} \log \langle f \langle \underline{y}_i; \underline{\theta} \rangle \rangle = -\frac{\partial}{\partial \theta_k} \rho \langle \underline{y}_i, \underline{\theta} \rangle$$

$$\longrightarrow \text{vector } \underline{s} \langle \underline{y}, \underline{\theta} \rangle = -\frac{\partial}{\partial \underline{\theta}} \rho \langle \underline{y}_i, \underline{\theta} \rangle.$$

Estimator:

$$\sum_{i=1}^n \underline{s} \left\langle \underline{y}_i, \hat{\underline{\theta}} \right\rangle = \underline{0}$$

g Multivariate Location: Minimize $\sum_i \frac{1}{2} (\underline{y}_i - \underline{\mu})^T (\underline{y}_i - \underline{\mu})$.

Scores $\underline{s} \langle \underline{y}, \underline{\mu} \rangle = \underline{y} - \underline{\mu}$

Estimator: $\sum_i (\underline{y}_i - \underline{\hat{\mu}}) = 0 \implies \underline{\hat{\mu}} = \frac{1}{n} \sum_i \underline{y}_i$.

h **M-estimators**

Minimize

$$\operatorname{argmin}_{\underline{\theta}} \left\langle \sum_i \rho \langle \underline{y}_i, \underline{\theta} \rangle \right\rangle \quad \text{MLE: } \rho \langle \underline{y}, \underline{\theta} \rangle = -\log \langle f \langle \underline{y}, \underline{\theta} \rangle \rangle$$

or solve implicit equation

$$\sum_i \underline{\psi} \langle \underline{y}_i, \underline{\theta} \rangle = \underline{0}, \quad \underline{\psi} \langle \underline{y}, \underline{\theta} \rangle = \frac{\partial}{\partial \underline{\theta}} \rho \langle \underline{y}, \underline{\theta} \rangle$$

Important: We do not require that $\underline{\psi}$ is $\partial \rho / \partial \underline{\theta}$ for any ρ
(in case of multidimensional $\underline{\theta}$)

Second type is more flexible!

i Multivariate Location: Bound the scores!

→ **Multivariate Huber Function:** $h_c \langle \underline{x} \rangle = \underline{x} \min \langle c / \|\underline{x}\|, 1 \rangle$

→ Estimator $\sum_i \underline{h}_c \langle \underline{y}_i - \hat{\underline{\mu}} \rangle = 0$

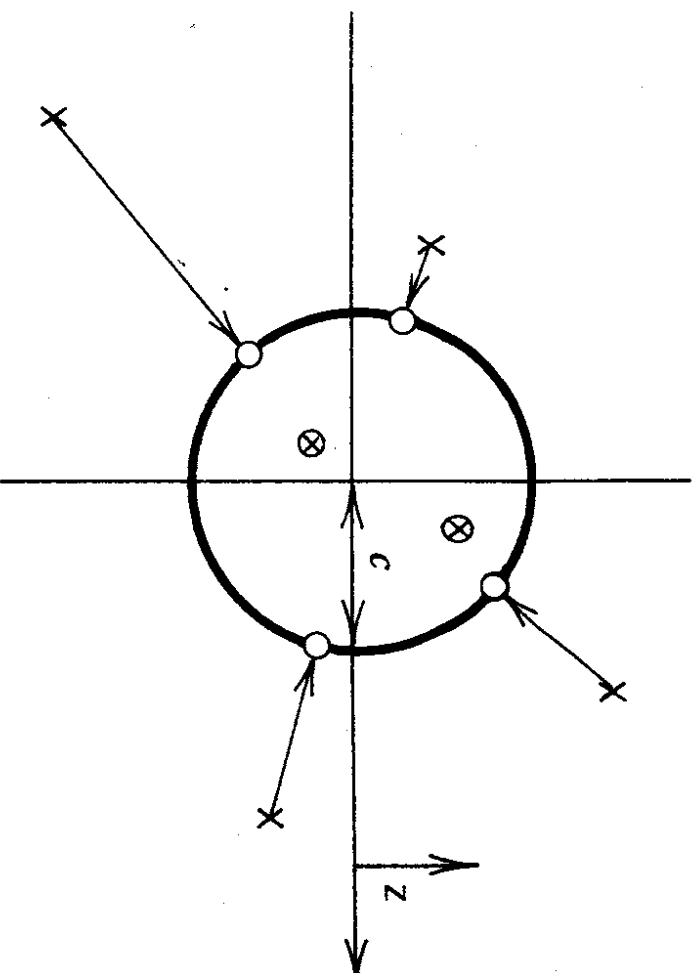


Figure 1. Sketch of the Huber function $z \mapsto h_c(z)$.

j Influence Function of M-estimators

Calculation: $\underline{T}_\varepsilon := \underline{T} \langle (1 - \varepsilon)F + \varepsilon \Delta_{\underline{y}} \rangle$ fulfills

$$(1 - \varepsilon) \int \underline{\psi} \langle \underline{x}, \underline{T}_\varepsilon \rangle dF \langle \underline{x} \rangle + \varepsilon \underline{\psi} \langle \underline{y}, \underline{T}_\varepsilon \rangle = 0 \quad (*)$$

Approximate: Linearization around $\underline{T}_0 = \underline{T} \langle F \rangle$:

$$\underline{\psi} \langle \underline{x}, \underline{T}_\varepsilon \rangle \approx \underline{\psi} \langle \underline{x}, \underline{T} \langle F \rangle \rangle + \frac{\partial}{\partial \underline{\theta}} \underline{\psi} \langle \underline{x}, \underline{T} \langle F \rangle \rangle (\underline{T}_\varepsilon - \underline{T} \langle F \rangle)$$

We have $\int \underline{\psi} \langle \underline{x}, \underline{T}_0 \rangle dF \langle \underline{x} \rangle = 0$

Plug approximation into (*)! – We need $\widehat{M} = \int \frac{\partial}{\partial \underline{\theta}} \underline{\psi} \langle \underline{x}, \underline{T}_0 \rangle dF \langle \underline{x} \rangle$ and get

$$(1 - \varepsilon) \widehat{M}(\underline{T}_\varepsilon - \underline{T}_0) + \varepsilon \underline{\psi} \langle \underline{y}, \underline{T}_0 \rangle + \varepsilon (\underline{T}_\varepsilon - \underline{T}_0) \frac{\partial}{\partial \underline{\theta}} \underline{\psi} \approx 0$$

$$\underline{T}_\varepsilon - \underline{T}_0 \approx -\frac{\varepsilon}{1 - \varepsilon} \widehat{M}^{-1} \underline{\psi} \langle \underline{y}, \underline{T}_0 \rangle$$

since $\varepsilon(\underline{T}_\varepsilon - \underline{T}_0) \dots$ is “small²”.

Thus,

$$\text{IF} \langle \underline{y}; \underline{I}, F \rangle = -\widetilde{M} \langle \underline{\psi}, \underline{I}_0 \rangle^{-1} \underline{\psi} \langle \underline{y}, \underline{I}_0 \rangle$$

An alternative formula for parametric distr. $F = F_{\underline{\theta}}$ and Fisher consistent M-estimators – meaning

$$\begin{aligned} \int \underline{\psi} \langle \underline{x}, \underline{\theta} \rangle f \langle \underline{x}; \underline{\theta} \rangle d\underline{x} &= 0 \\ \int \frac{\partial}{\partial \underline{\theta}} \psi \langle \underline{x}, \underline{\theta} \rangle dF_{\underline{\theta}} \langle \underline{x} \rangle + \int \underline{\psi} \langle \underline{x}, \underline{\theta} \rangle \frac{\partial}{\partial \underline{\theta}} f \langle \underline{x}; \underline{\theta} \rangle^T d\underline{x} &= 0 \end{aligned}$$

Note that $s \langle \underline{x}, \underline{\theta} \rangle = \frac{\partial}{\partial \underline{\theta}} f \langle \underline{x}; \underline{\theta} \rangle f \langle \underline{x}; \underline{\theta} \rangle^{-1}$

$$\begin{aligned} \widetilde{M} &= -\int \underline{\psi} \langle \underline{x}, \underline{\theta} \rangle s \langle \underline{x}, \underline{\theta} \rangle^T f \langle \underline{x}; \underline{\theta} \rangle d\underline{x} = -M \langle \underline{\psi}, \underline{\theta} \rangle \\ M \langle \underline{\psi}, F_{\underline{\theta}} \rangle &:= \int \underline{\psi} \langle \underline{y}, \underline{\theta} \rangle \underline{s} \langle \underline{y}, \underline{\theta} \rangle^T dF_{\underline{\theta}} \langle \underline{y} \rangle \end{aligned}$$

Finally: **IF** for M-estimators:

$$\begin{aligned}\text{IF} \langle \underline{y}; \underline{\psi}, F_{\underline{\theta}} \rangle &= M \langle \underline{\psi}, \underline{\theta} \rangle^{-1} \underline{\psi} \langle \underline{y}, \underline{\theta} \rangle \\ M \langle \underline{\psi}, \underline{\theta} \rangle &:= \int \underline{\psi} \langle \underline{y}, \underline{\theta} \rangle \underline{s} \langle \underline{y}, \underline{\theta} \rangle^{\text{T}} dF_{\underline{\theta}} \langle \underline{y} \rangle\end{aligned}$$

^k Asymptotic Covariance Matrix

$$\begin{aligned}V \langle \underline{\psi}; F_{\underline{\theta}} \rangle &= \int \text{IF} \langle \underline{y}; \underline{T}, F_{\underline{\theta}} \rangle \text{IF} \langle \underline{y}; \underline{T}, F_{\underline{\theta}} \rangle^{\text{T}} dF_{\underline{\theta}} \langle \underline{y} \rangle \\ &= M \langle \underline{\psi}, \underline{\theta} \rangle^{-1} \mathbf{Q} \langle \underline{\psi}, F_{\underline{\theta}} \rangle M \langle \underline{\psi}, \underline{\theta} \rangle^{-\text{T}} \\ Q \langle \underline{\psi}, F \rangle &:= \int \underline{\psi} \langle \underline{y}, \underline{T} \langle F \rangle \rangle \underline{\psi} \langle \underline{y}, \underline{T} \langle F \rangle \rangle^{\text{T}} dF \langle \underline{y} \rangle\end{aligned}$$

Maximum Likelihood Estimator:

$$\begin{aligned} M\langle \underline{s}, \underline{\theta} \rangle &= Q\langle \underline{s}, F_{\underline{\theta}} \rangle \\ &= J\langle \underline{\theta} \rangle := \int \underline{s} \langle \underline{y}, \underline{\theta} \rangle \underline{s} \langle \underline{y}, \underline{\theta} \rangle^T dF_{\underline{\theta}}\langle \underline{y} \rangle \\ V\langle \underline{s}, F_{\underline{\theta}} \rangle &= J\langle \underline{\theta} \rangle^{-1} \end{aligned}$$

$J\langle \underline{\theta} \rangle$: Fisher Information Matrix

I Optimality of the MLE

In 1 dim., MLE is optimal: it has minimal variance,

$$V\langle T, F_{\underline{\theta}} \rangle \geq V\langle T_{MLE}, F_{\underline{\theta}} \rangle$$

for all Fisher-consistent T .

What does this mean for multidim. estimators $\underline{T}$ of $\underline{\theta}$?

Let $\underline{T}^*$ be any Fisher consistent est. of $\underline{\theta}$, $V \langle \underline{T}^*, F_{\underline{\theta}} \rangle =: V^*$

Take any linear combination $\phi = \underline{b}^T \underline{\theta}$.

MLE for ϕ ? $\hat{\phi}_{MLE} = \underline{b}^T \underline{T}_{MLE}$; $V \langle \hat{\phi}_{MLE}, F_{\underline{\theta}} \rangle = \underline{b}^T V_{MLE} \underline{b}$.

For every $\tilde{T}$ that estimates ϕ the as. variance is at least as big.

... also for $\underline{b}^T \underline{T}^*$, whose as.var is $\underline{b}^T V^* \underline{b}$.

$$\implies \underline{b}^T V^* \underline{b} \geq \underline{b}^T V \langle \underline{T}_{MLE}, F_{\underline{\theta}} \rangle \underline{b}$$

$$\longrightarrow V^* - V \langle \underline{T}_{MLE}, F_{\underline{\theta}} \rangle \geq 0$$

in the sense: The difference is “positive semidefinite”.

... a rather **strong optimality** statement for the MLE!

m

Gross Error Sensitivity

Straightforward:

$$\gamma^* \langle \underline{T}, F \rangle = \sup_{\underline{y}} \langle \|\text{IF} \langle \underline{y}; \underline{T}, F \rangle\| \rangle$$

Afterthought: Bias should be small in directions that can be estimated precisely (small as.var.) and larger bias can be tolerated in directions with larger as.var.

→ Change norm. “Measure in the metric of as.var.”

→ Change norm. “Measure in the metric of as.var.”

- ... of the MLE → Fisher Information $\mathbf{J}^{-1}$

Information standardized GES:

$$\gamma_i^* \langle \underline{T}, F \rangle^2 = \sup_x \langle \text{IF} \langle \underline{y}; \underline{T}, F \rangle^T \mathbf{J} \langle \underline{T} \langle F \rangle \rangle \text{IF} \langle \underline{y}; \underline{T}, F \rangle \rangle$$

- ... of the estimator itself:

Self standardized GES:

$$\gamma_s^* \langle \underline{T}, F \rangle^2 = \sup_x \langle \text{IF} \langle \underline{y}; \underline{T}, F \rangle^T \mathbf{V} \langle \underline{T}, F \rangle^{-1} \text{IF} \langle \underline{y}; \underline{T}, F \rangle \rangle$$

n

Multivariate Location

Things are simple!

$$\begin{aligned} M \langle \underline{h}_b, \mathcal{N}_m \langle \underline{\mu}, \mathbf{I} \rangle \rangle &= \int \underline{h}_b \langle \underline{x} \rangle \underline{x}^T d\mathcal{N}_m \langle \underline{\mu}, \mathbf{I} \rangle = \gamma_1 \mathbf{I} \\ \gamma_1 &= \int u \min \langle b/u, 1 \rangle d\chi_m^2 \langle u \rangle \\ Q \langle \underline{h}_b, \mathcal{N}_m \langle \underline{\mu}, \mathbf{I} \rangle \rangle &= \int \underline{h}_b \langle \underline{x} \rangle \underline{h}_b \langle \underline{x} \rangle^T d\mathcal{N}_m \langle \underline{\mu}, \mathbf{I} \rangle = \gamma_2 \mathbf{I} \\ \gamma_2 &= \int u \min \langle (b/u)^2, 1 \rangle d\chi_m^2 \langle u \rangle \\ V \langle \underline{h}_b, \mathcal{N}_m \langle \underline{\mu}, \mathbf{I} \rangle \rangle &= \frac{\gamma_2}{\gamma_1^2} \mathbf{I} \\ \gamma_i^* \langle \widehat{\underline{\mu}}, \mathcal{N}_m \langle \underline{\mu}, \mathbf{I} \rangle \rangle &= \gamma_i^* \langle \widehat{\underline{\mu}}, \mathcal{N}_m \langle \underline{\mu}, \mathbf{I} \rangle \rangle = \frac{b}{\gamma_1} \\ \gamma_i^* \langle \widehat{\underline{\mu}}, \mathcal{N}_m \langle \underline{\mu}, \mathbf{I} \rangle \rangle &= \frac{b}{\sqrt{\gamma_2}} \end{aligned}$$

- **Hampel's Optimal Robust Estimators**

“Bound the scores”: What does this mean in > 1 dimensions?

Guess:

- Use the scores where they are small enough.
 - Change them as little as possible otherwise.
- Multivariate Huber Function:

$$\underline{h}_b \langle \underline{x} \rangle = \underline{x} \min \langle b / \|x\|, 1 \rangle$$

Derivation of the optimal estimator:

Another look at the solution for a single parameter:

Write $h_b \langle x \rangle = \max \langle \min \langle x, b \rangle, -b \rangle = x \min \langle 1, b/|x| \rangle$

Note that $h_{b\beta} \langle x \rangle = \beta h_b \langle x/\beta \rangle$.

IF for the optimal estimator was

$$\text{IF} \langle x; T_{opt}, F_\theta \rangle = \frac{1}{d} h_b \langle s \langle x, \theta \rangle - a \rangle = h_{b/d} \langle d(s \langle x, \theta \rangle - a) \rangle$$

where a such that $\int h_b \langle s \langle x, \theta \rangle - a \rangle dF_\theta \langle x \rangle = 0$ and

$$d = \int h_b \langle s \langle x, \theta \rangle - a \rangle s \langle x, \theta \rangle dF_\theta \langle x \rangle.$$

The achieved bound γ on the IF (the GES) was $\gamma = b/d$.

Multivariate version:

$$\text{IF} \langle \underline{x}; \underline{\psi}_{opt}, F_{\underline{\theta}} \rangle = h_{\gamma} \langle M^{-1}(\underline{s} \langle \underline{x}, \underline{\theta} \rangle - \underline{a}) \rangle$$

where $\underline{a}$ and M are such that

$$\begin{aligned} h_{\gamma} \langle M^{-1}(\underline{s} \langle \underline{x}, \underline{\theta} \rangle - \underline{a}) \rangle dF_{\underline{\theta}} \langle \underline{x} \rangle &= \underline{0} \\ \int h_{\gamma} \langle M^{-1}(\underline{s} \langle \underline{x}, \underline{\theta} \rangle - \underline{a}) \rangle s \langle \underline{x}, \underline{\theta} \rangle^T dF_{\underline{\theta}} \langle \underline{x} \rangle &= \underline{I} \end{aligned}$$

Lemma 5, multivariate:

- Existence of $\underline{a}$ and M ...
- Optimality:

$$\begin{aligned} V &= \int \mathbf{IF} \mathbf{IF}^T dF \\ &= \int (\mathbf{IF} \langle \underline{x}; \underline{T}, F_{\underline{\theta}} \rangle - M(\underline{s} \langle \underline{x}, \underline{\theta} \rangle - \underline{a}))^{*T} dF_{\underline{\theta}} \langle \underline{x} \rangle \\ &\quad + \text{constant terms} \end{aligned}$$

where $\underline{a}^{*T} := \underline{a} \underline{a}^T$.

Pointwise minimum outside $\|\mathbf{IF}\| = \gamma$?

No simple answer!

Choose $\text{tr} \langle V \rangle$ as a criterion to minimize. $\longrightarrow$

$$\begin{aligned} \text{tr} \langle V \rangle &= \int \text{IF}^T \text{IF} \, dF \\ &= \int \| \text{IF} \langle \underline{x}; \underline{T}, F_{\underline{\theta}} \rangle - M(\underline{s} \langle \underline{x}, \underline{\theta} \rangle - \underline{a}) \|^2 dF_{\underline{\theta}} \langle \underline{x} \rangle \\ &\quad + \text{constant terms} \end{aligned}$$

Pointwise minimum is the multidimensional Huber function.

Other choices: E.g., focus on V_{11} .

$\longrightarrow$ Keep first component of $M(\underline{s} \langle \underline{x}, \underline{\theta} \rangle - \underline{a})$ wherever possible!
(This changes the determination of M !)

$\longrightarrow$ In the multiparameter case, there is no *strongly* optimal estimator in the Hampel sense.

Variations

- **Optimal estimator for Information Standardized Sensitivity:**

Change parameter by linear transformation (or base change)

$$\tilde{\underline{\theta}} = \underline{B} \underline{\theta} \text{ such that } \underline{J} \left\langle \tilde{\underline{\theta}} \right\rangle = \underline{I},$$

then find optimal estimator as above, finally transform back.

- **Optimal estimator for the Self Standardized Sensitivity:**

For $\underline{M}$, require

$$\int h_{\gamma} \left\langle \underline{M}^{-1} (\underline{s} \langle \underline{x}, \underline{\theta} \rangle - \underline{a}) \right\rangle^{*T} dF_{\underline{\theta}} \langle \underline{x} \rangle = \underline{I}$$

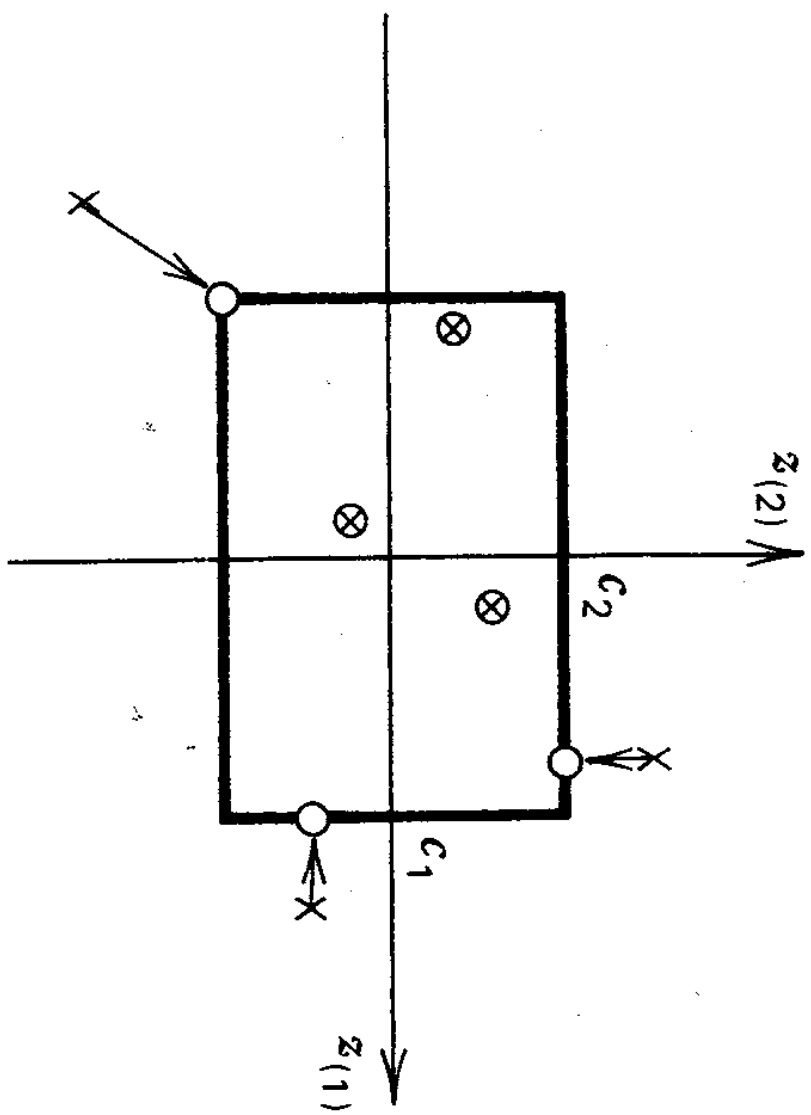
- q • **Partitioned Parameter**

Often, the parameter vector consists of

- a “primary part” $\underline{\theta}_{[1]}$, which is of main interest
- a “nuisance part” $\underline{\theta}_{[2]}$

Then, it seems appropriate to measure sensitivity separately.

→ Use correspondingly modified Huber function (see Figure)



10 Multivariate Scatter Matrices

10.1 Models

a The multivariate Normal Distribution

Density

$$f(\underline{y}; \underline{\mu}, \underline{\Phi}) = c \cdot \exp \left\langle -\frac{1}{2} (\underline{Y} - \underline{\mu})^T \underline{\Phi}^{-1} (\underline{Y} - \underline{\mu}) \right\rangle$$

$$1/c = (2\pi)^{m/2} \det \langle \underline{\Phi} \rangle^{1/2}$$

More insight is provided by the following derivation.

b **Normal distribution as a transformation family**

$\underline{Z} \sim \mathcal{N}_m(\underline{0}, \underline{I})$: centered, spherically distributed observations
= m random variables $Z^{(j)}$, standard normal, independent.

Consider, for any vector $\underline{\mu}$ and regular matrix B ,

$$\underline{X} = \underline{\mu} + B \underline{Z}$$

= linear (or affine) transformation of $\underline{Z}$.

Call the distribution of $\underline{X}$ $F(\underline{\mu}, B)$.

Transformations generate a family of distributions.

- c General: Use any group of transformations on any “starting” distribution.
- d Expectation and cov. matrix for linear transformations, if $\mathcal{E}\langle \underline{Z} \rangle = \underline{0}$ and $\text{var}\langle \underline{Z} \rangle = \mathbf{I}$,

$$\mathcal{E}\langle \underline{X} \rangle = \underline{\mu}, \quad \text{var}\langle \underline{X} \rangle = \mathbf{B}\mathbf{B}^{\text{T}}$$

e Identifiability

Problem: **Orthogonal** transformations

$$\underline{X} = \underline{\mu} + \underline{Q}\underline{Z}, \quad \underline{Q}\underline{Q}^T = \underline{I}$$

leave the distribution of $\underline{Z}$ unchanged

(= definition of “spherically symmetric” distr.)

→ Parameter $\underline{B}$ is **not identifiable!**

Covariance matrix: $\text{var}(\underline{X}) = \underline{B}\underline{B}^T =: \underline{\Phi}$ is
a suitable parameter instead of $\underline{B}$.

(different $\underline{\Phi}$ → different distributions)

f Elliptical distributions

Let F_0 be any spherically symmetric distribution. $\underline{Z} \sim F_0$.

$$F\langle \underline{\mu}, \Sigma \rangle = \text{distribution of } \underline{X} = \underline{\mu} + B\underline{Z}, \\ B B^T = \Sigma, \quad \Sigma \text{ regular.}$$

Family is called the family of **elliptical distr. generated by F_0** .

Σ scatter matrix, not covariance in general.

- Covariance may not exist,
- $\text{var}\langle Z \sim F_0 \rangle$ may be $\sigma_0^2 \mathbf{I}$ with $\sigma_0 \neq 1$.

g Densities

Density of F_0 must be a function of the (squared) “radius”,

$$f_0 \langle \underline{z} \rangle = \widetilde{f_0} \langle \|\underline{z}\|^2 \rangle.$$

$$\text{For normal: } \widetilde{f_0} \langle \|\underline{z}\|^2 \rangle = c \cdot \exp \langle -\|\underline{z}\|^2/2 \rangle$$

General $\underline{\mu}, \mathbf{\Phi}$: Let $\mathbf{B}$ such that $\mathbf{B}\mathbf{B}^T = \mathbf{\Phi}$.

$$\underline{z} = \mathbf{B}^{-1}(\underline{x} - \underline{\mu}) \longrightarrow \text{Jacobian } \det \langle \mathbf{B} \rangle = \det \langle \mathbf{\Phi} \rangle^{1/2}$$

$$\begin{aligned} f \langle \underline{x}; \underline{\mu}, \mathbf{\Phi} \rangle &= \det \langle \mathbf{\Phi} \rangle^{-1/2} f_0 \langle \underline{z} \rangle = \det \langle \mathbf{\Phi} \rangle^{-1/2} \widetilde{f_0} \langle \|\underline{z}\|^2 \rangle \\ &= \det \langle \mathbf{\Phi} \rangle^{-1/2} \widetilde{f_0} \langle (\underline{x} - \underline{\mu})^T \mathbf{\Phi}^{-1} (\underline{x} - \underline{\mu}) \rangle \end{aligned}$$

- h Examples of elliptical distributions: Multivariate t distr.
= distr. of $\hat{\underline{B}}^{-1} \hat{\underline{\mu}}$ for $\underline{X}_i \sim \mathcal{N}_m(\underline{0}, \underline{I})$

Not “natural” for observations!

If $\underline{Z} \sim F_0$, the components are not independent.

Independent $X^{(j)}$ cannot be modelled by any elliptical family
except for the Normal!

i **Size and Shape**

(Log-) Size of the covariance matrix:

$$\tau := \log \langle \det \langle \mathbb{P} \rangle \rangle = \frac{1}{m} \sum_k \log \langle \lambda_k \rangle \quad \lambda_k = \text{eigenvalues of } \mathbb{P}.$$

Shape: “standard deviation” of the log eigenvalues:

$$\eta^2 = \sum_k (\log \langle \lambda_k \rangle - \tau)^2.$$

Both size & shape are invariant under *orthogonal* transf.

10.2 Equivariant Estimators

- a Estimators $\hat{\underline{\mu}}$ and $\hat{\Phi}$ should “respect” the transformation nature of the family:

If $\underline{Y}_i = \underline{a} + C\underline{X}_i$, then

$$\hat{\underline{\mu}} \langle \underline{Y}_1, \dots, \underline{Y}_n \rangle = \underline{a} + C \hat{\underline{\mu}} \langle \underline{X}_1, \dots, \underline{X}_n \rangle$$

$$\hat{\Phi} \langle \underline{Y}_1, \dots, \underline{Y}_n \rangle = C \hat{\Phi} \langle \underline{X}_1, \dots, \underline{X}_n \rangle C^T$$

as these relations hold if $\hat{\underline{\mu}}$ and $\hat{\Phi}$ are replaced by the parameters $\underline{\mu}$ and Φ .

Estimators for which these relations hold are called equivariant (with respect to affine transf.s, the generating family)

b Other way 'round: Given $\hat{\underline{\mu}} \langle \underline{X}_1, \dots, \underline{X}_n \rangle$ and $\hat{\underline{\Psi}} \langle \underline{X}_1, \dots, \underline{X}_n \rangle$

Let $\hat{B}$ be a "square root" of $\hat{\underline{\Psi}}$, $\hat{B} \hat{B}^T = \hat{\underline{\Psi}}$. Then,

$$\begin{aligned} \hat{\underline{\mu}} \langle \dots, \hat{B}^{-1}(\underline{X}_i - \hat{\underline{\mu}}), \dots \rangle &= \underline{0} \\ \hat{\underline{\Psi}} \langle \dots, \hat{B}^{-1}(\underline{X}_i - \hat{\underline{\mu}}), \dots \rangle &= \underline{I} \end{aligned}$$

→ equation determining $\hat{\underline{\mu}}$ and $\hat{B}$.

$\hat{\underline{\mu}}$ and $\hat{B} \hat{B}^T = \hat{\underline{\Psi}}$ usually uniquely determined by these eqns.

Thus, it is sufficient to define when $\hat{\underline{\mu}} = \underline{0}$ and $\hat{\underline{\Psi}} = \underline{I}$.

Every such definition yields potentially suitable estimating functionals $\mu \langle F \rangle$ and $\Psi \langle F \rangle$.

10.3 M-estimators

^a Maximum likelihood for elliptical distributions

Write $\tilde{f}_0 \langle \|z\|^2 \rangle =: \exp \langle -\rho \langle \|z\|^2 \rangle \rangle$.

(Notice the ² in the argument!)

$$\begin{aligned}
 \ell \langle \underline{x}; \underline{\mu}, \mathbf{\Phi} \rangle &= \log \langle f \langle \underline{x}; \underline{\mu}, \mathbf{\Phi} \rangle \rangle \\
 &= -\rho \langle (\underline{x} - \underline{\mu})^T \mathbf{\Phi}^{-1} (\underline{x} - \underline{\mu}) \rangle - \frac{1}{2} \log \det \langle \mathbf{\Phi} \rangle \\
 \frac{\partial \ell}{\partial \underline{\mu}} &= \rho' \langle \|z\|^2 \rangle 2 \mathbf{\Phi}^{-1} (\underline{x} - \underline{\mu}) = \rho' \langle \|z\|^2 \rangle 2 \mathbf{B}^{-1} \underline{z} \\
 \frac{\partial \ell}{\partial \mathbf{\Phi}} &= \rho' \langle \|z\|^2 \rangle \mathbf{\Phi}^{-1} (\underline{x} - \underline{\mu}) (\underline{x} - \underline{\mu})^T \mathbf{\Phi}^{-1} - \mathbf{\Phi}^{-1} \\
 &= \mathbf{B}^{-T} \left(\rho' \langle \|z\|^2 \rangle \underline{z} \underline{z}^T - \mathbf{I} \right) \mathbf{B}^{-1}
 \end{aligned}$$

For the second equation, we have used the formulas

$$\begin{aligned} \frac{\partial(\log \det \langle A \rangle)}{\partial A} &= A^{-T} \\ u = \underline{x}^T V^{-1} \underline{x} &\Rightarrow \frac{\partial u}{\partial V} = -V^{-1} \underline{x} \underline{x}^T V^{-1} \end{aligned}$$

Thus, maximum likelihood solves

$$\begin{aligned} \text{ave} \langle \rho' \langle \|\underline{z}_i\|^2 \rangle \underline{z}_i \rangle &= \underline{0} \\ \text{ave} \langle \rho' \langle \|\underline{z}_i\|^2 \rangle \underline{z}_i \underline{z}_i^T \rangle &= \underline{I} \end{aligned}$$

Here, ρ' is a weighting function, not a ψ function (because of the 2 in the argument of ρ).

b M-estimators

$$\begin{aligned} \text{ave}_i \left\langle w^{(\mu)} \left\langle \|\underline{z}_i\|^2 \right\rangle \underline{z}_i \right\rangle &= \underline{0} \\ \text{ave}_i \left\langle w^{(\eta)} \left\langle \|\underline{z}_i\|^2 \right\rangle \underline{z}_i \underline{z}_i^T \right\rangle &= \text{ave}_i \left\langle w^{(\delta)} \left\langle \|\underline{z}_i\|^2 \right\rangle \right\rangle \mathbf{I} \end{aligned}$$

Note that for MLEs, $w^{(\mu)} = w^{(\eta)}$ and $w^{(\delta)} \langle u \rangle = 1$.

As in the location-scale case, the M-estimators are

more general than MLE for long-tailed distributions.

$w^{(\delta)}$ is not introduced in the usual literature (Maronna+...).

Alternative form: Weighted classical estimators with implicitly defined weights

$$\begin{aligned}\hat{\underline{\mu}} &= \frac{\text{ave}_i \langle w^{(\mu)} \langle D_i^2 \rangle \underline{X}_i \rangle}{\text{ave}_i \langle w^{(\mu)} \langle D_i^2 \rangle \rangle} \\ \hat{\underline{\Phi}} &= \frac{\text{ave}_i \langle w^{(\eta)} \langle D_i^2 \rangle (\underline{X}_i - \hat{\underline{\mu}})(\underline{X}_i - \hat{\underline{\mu}})^T \rangle}{\text{ave}_i \langle w^{(\delta)} \langle D_i^2 \rangle \rangle} \\ D_i^2 &= (\underline{X}_i - \hat{\underline{\mu}})^T \hat{\underline{\Phi}}^{-1} (\underline{X}_i - \hat{\underline{\mu}})\end{aligned}$$

Weighted classical estimators can only be obtained when allowing

for a $w^{(\delta)}$, and choosing $w^{(\delta)} = w^{(\eta)}$ ($= w^{(\mu)}$).

Computation: Iteratively reweighted classical estimates.

c Influence Function

Abbreviate $\text{IF}(\underline{z}) = \text{IF}\left(\underline{z}; [\underline{\hat{\mu}}, \hat{\Phi}], \mathcal{N}_m(\underline{0}, \mathbf{I})\right)$

$$\text{IF}^{(\mu)}(\underline{z}) = a^{(\mu)} w^{(\mu)} \langle \|\underline{z}\|^2 \rangle \underline{z}$$

$$\text{IF}^{(\Phi)}(\underline{z}) = a^{(\eta)} w^{(\eta)} \langle \|\underline{z}\|^2 \rangle \underline{z} \underline{z}^T - \tilde{w}^{(\delta)} \langle \|\underline{z}\|^2 \rangle \mathbf{I}$$

where $\tilde{w}^{(\delta)}$ depends on $w^{(\delta)}$ and $w^{(\eta)}$.

General case:

$$\text{IF}^{(\mu)}\left(\underline{x}; [\underline{\hat{\mu}}, \hat{\Phi}], \mathcal{N}_m(\underline{\mu}, \Phi)\right) = \mathbf{B} \text{IF}^{(\mu)}(\underline{z})$$

$$\text{IF}^{(\Phi)}\left(\underline{x}; [\underline{\hat{\mu}}, \hat{\Phi}], \mathcal{N}_m(\underline{\mu}, \Phi)\right) = \mathbf{B} \text{IF}^{(\Phi)}(\underline{z}) \mathbf{B}^T$$

For all equivariant estimators, the IF must have this form.
(not only M-estimators)

Calculation of $a^{(\mu)}, a^{(\eta)}, a^{(\tau)}$: We need integrals of the form

$$\boldsymbol{D} = \int \underline{a} \langle u \rangle \underline{b} \langle u \rangle^{\text{T}} d\tilde{F}_0 \langle u \rangle$$

where $\underline{a} \langle ||\underline{z}|| \rangle = \underline{\psi} \langle \underline{z}; \underline{0}, \boldsymbol{I} \rangle$, $\underline{b} \langle ||\underline{z}|| \rangle = \underline{s} \langle \underline{z}, \underline{0}, \boldsymbol{I} \rangle$

For the Fisher Information and the as.var. of an equivariant est.

the same type of integral is needed

with $a = b = s$ and $a = b = \psi$, respectively.

These integrals are characterized by $d^{(\mu)}, d^{(\eta)}, d^{(\tau)}$, where

$$d^{(\mu)} = \int \frac{u}{m} \omega_a^{(\mu)} \langle u \rangle \omega_b^{(\mu)} \langle u \rangle d\tilde{F}_0 \langle u \rangle$$

$$d^{(\eta)} = (1 + 2/m)^{-1} \int \left(\frac{u}{m} \right)^2 \omega_a^{(\mu)} \langle u \rangle \omega_b^{(\mu)} \langle u \rangle d\tilde{F}_0 \langle u \rangle$$

$$d^{(\tau)} = \frac{1}{2m} \int \psi_a^{(\tau)} \langle u \rangle \psi_b^{(\tau)} \langle u \rangle d\tilde{F}_0 \langle u \rangle$$

Proof: See Hampel et al.

D = “d-type matrix” = special structure!

This shows that $\psi^{(\tau)} \langle u \rangle = u \cdot \omega^{(\eta)} \langle u \rangle - m \cdot \omega^{(\delta)} \langle u \rangle$

is the function that characterizes the M-estimator of Ψ
together with $\omega^{(\eta)}$ — not $\omega^{(\delta)}$!

d **Sensitivities**

Maximum of the norm of the IF.

It is natural to focus on the standardized sensitivities.

Standardized sensitivities $\longrightarrow$ invariant w.r.t. affine transf.

$\equiv$ independent of $\underline{\mu}$, $\underline{\Sigma}$.

$\longrightarrow$ only need to consider IF at F_0 .

The norm splits into 2 parts, $\|IF\|^2 = \|IF^{(\mu)}\|^2 + \|IF^{(\Sigma)}\|^2$

For the $\mathbb{V}$ term, we get a further split:

Let $\psi^{(\tau)}\langle u \rangle := u \cdot w^{(n)}\langle u \rangle - m \cdot w^{(\delta)}\langle u \rangle$

$$\begin{aligned} \|\text{IF}^{(\mathbb{V})}\langle \underline{z} \rangle\|^2 &:= \sum_{j,k} \left(\text{IF}^{(\mathbb{V})}\langle \underline{z} \rangle \right)_{j,k}^2 \\ &= \left(a^{(n)} w^{(n)} \langle \|\underline{z}\|^2 \rangle \|\underline{z}\|^2 \right)^2 + \left(a^{(\tau)} \psi^{(\tau)} \langle \|\underline{z}\|^2 \rangle \right)^2 \\ &=: \|\text{IF}^{(n)}\langle \underline{z} \rangle\|^2 + |\text{IF}^{(\tau)}\langle \underline{z} \rangle|^2 \end{aligned}$$

$\psi^{(\tau)}$ & $w^{(n)}$ connected to size and shape parameters, see 10.1.i.

Natural choice $w^{(\delta)} = w^{(n)} \longrightarrow$

$$\|\text{IF}^{(\mathbb{V})}\langle \underline{z} \rangle\|^2 = w^{(n)} \langle \|\underline{z}\|^2 \rangle \left((a^{(n)} + a^{(\tau)}) \|\underline{z}\|^2 - a^{(\mu)} m \right)$$

allows for easy definition of **redescending estimators**

$w\langle u \rangle \rightarrow 0$ for $u \rightarrow \infty$ (cf. “weighted scale” 7.1.e)

e

Optimal robust estimation

Bounded influence function, but as efficient for central model (normal distribution) as possible!

Bound 3 “parts” separately,

$$\| \text{IF}^{(\mu)} \| \leq \gamma_\mu, \quad \| \text{IF}^{(\eta)} \| \leq \gamma_\eta, \quad | \text{IF}^{(\tau)} | \leq \gamma_\tau.$$

$$\begin{aligned} \text{Result: } w^{(\mu)} \langle u \rangle &= \min \left\langle 1, \sqrt{b_\mu / u} \right\rangle, & w^{(\eta)} \langle u \rangle &= \min \langle 1, b_\eta / u \rangle, \\ \psi^{(\tau)} \langle u \rangle &= h_{b_\tau} \langle u - m\beta \rangle \quad (\text{Huber function}) \text{ or} \\ w^{(\delta)} \langle u \rangle &= u w^{(\eta)} \langle u \rangle - (u - m\beta) \min \langle 1, b_\tau / (u - m\beta) \rangle, \end{aligned}$$

β a suitable constant for Fisher consistency, depending on b_η and b_τ .

10.4 Breakdown

^a Breakdown of covariance matrix estimators

No breakdown of $\hat{\Sigma}$ means

eigenvalues $< \infty$ and $> \gamma \langle \Sigma \rangle > 0$ (bounded away from 0).

If $\Sigma \rightarrow \text{singular}$, $\hat{\Sigma} \rightarrow$ should also be singular (equal subspace).

Goal of multivariate statistics:

Study **relations between variables**

For all affinely equivariant estimators, boundedness of $\hat{\Sigma}$ for $\Sigma = I$ is sufficient.

b Breakdown of M-estimators

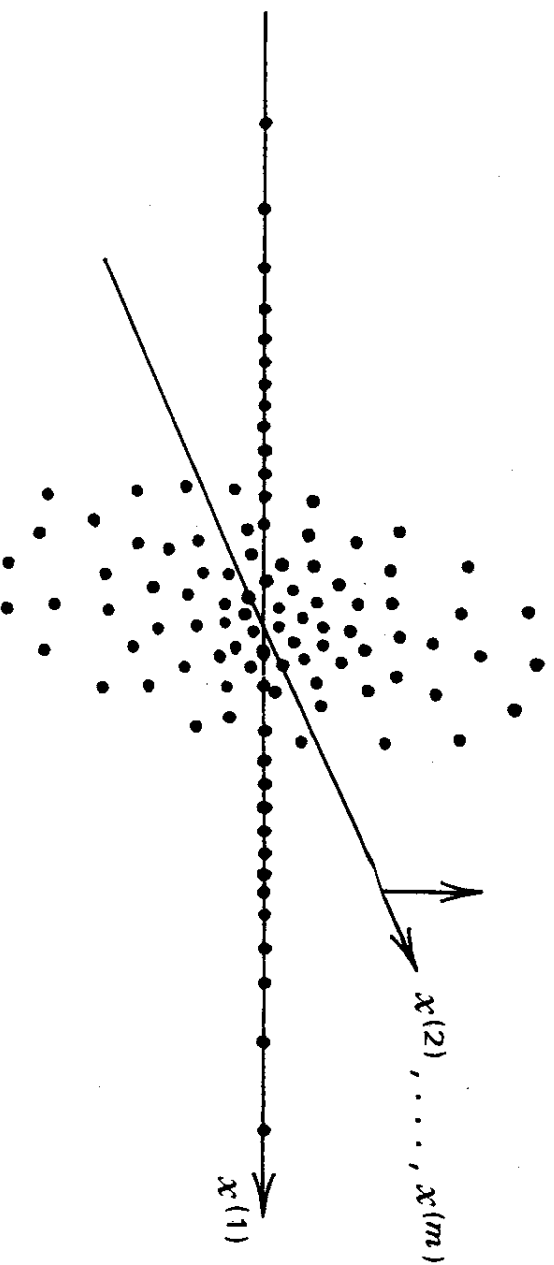
Result: **breakdown point** $< 1/m$.

Proof: Consider the “**barrow wheel**” distribution:

Let $F_\tau = \mathcal{N}(\underline{0}, \text{diag}(\tau, 1, \dots, 1))$,

H distribution of $[U, 0, \dots, 0]$ with $U \sim \chi_{p-1}^2$ Chisqu. distr.,

$G_{\varepsilon, \tau} = (1 - \varepsilon)F_\tau + \varepsilon H$ looks like the wheel of a barrow.



Let $\varepsilon = 1/m$ and $\tau \rightarrow 0$. Then, $T \langle G_{\varepsilon,0} \rangle = cI$ is a solution of the M-estimating equation for $\mathbb{P}$.

→ Breakdown!

The barrow wheel situation is easily detected in practise – unless rotated such that the first axis is mapped to the space diagonal.

Then, the scatterplot of every pair of variables looks spherical if the dimension is not small.

c What does this mean?

Consider 100 obs. of 7 variables = 700 data.

Assume 2% gross errors

- about 13-14 (multivar.) obs. contain ≥ 1 error(s)
- **breakdown at 2% of the data, not $1/m = 1/7$**
- another “curse of dimension”

d **Non affinely equivariant estimators?**

Idea: **Estimate each element of the covariance matrix robustly!**

E.g., Gnanadesikan and Kettenring (1972): Note that

$$\text{cov} \left\langle X^{(j)}, X^{(k)} \right\rangle = \frac{1}{4} \left(\text{var} \left\langle X^{(j)} + X^{(k)} \right\rangle - \text{var} \left\langle X^{(j)} - X^{(k)} \right\rangle \right)$$

Robust version:

$$\begin{aligned} \hat{\rho} \left\langle X^{(j)}, X^{(k)} \right\rangle &= \frac{1}{4} \left(\hat{\sigma}^2 \left\langle \frac{X^{(j)}}{\hat{\sigma} \left\langle X^{(j)} \right\rangle} + \frac{X^{(k)}}{\hat{\sigma} \left\langle X^{(k)} \right\rangle} \right\rangle - \right. \\ &\quad \left. \hat{\sigma}^2 \left\langle \frac{X^{(j)}}{\hat{\sigma} \left\langle X^{(j)} \right\rangle} - \frac{X^{(k)}}{\hat{\sigma} \left\langle X^{(k)} \right\rangle} \right\rangle \right) \\ \widehat{\text{cov}} \left\langle X^{(j)}, X^{(k)} \right\rangle &= \hat{\rho} \left\langle X^{(j)}, X^{(k)} \right\rangle \cdot \hat{\sigma} \left\langle X^{(j)} \right\rangle \hat{\sigma} \left\langle X^{(k)} \right\rangle \end{aligned}$$

e **Problem:** The resulting matrix will in general

not be positive (semi-) definite.

Trick to make them positive definite:

Apply Eigenvalue analysis

$$\tilde{\Sigma} = Q D Q^T$$

with diagonal D and orthogonal Q .

If $\tilde{\Sigma}$ was a covariance matrix, then D would contain the variances D_{ii} of the uncorrelated variables $\underline{Z} = Q^{-1} \underline{X}$.

→ Estimate robust variances $\hat{v}_k$ of the $Z^{(k)}$

Then use back-transformed matrix as estimate,

$$\hat{\Sigma} = Q^T \text{diag} \langle \hat{v}_1, \dots, \hat{v}_m \rangle Q$$

f On the other hand:

- These estimators are **not equivariant**.

But they *want* to react to “outliers in single variables”, which is not an affinely invariant model.

- **Singular** covariance matrices will most probably **not be detected**.

(Think of the barrow wheel!)

10.5 Estimation with high breakdown point

a Projection estimator

Outlyingness of an observation:

One-dimensional: $|x_i - \hat{\mu}| / \hat{\sigma}$

$\hat{\mu}, \hat{\sigma}$: robust estimators of location and scatter.

Multi-dimensional: Examine projections $\underline{b}^T \underline{X}_i$, $|\underline{b}| = 1$.

Outlyingness: $|x_i - \hat{\mu}| \langle \dots, \underline{b}^T \underline{X}_h, \dots \rangle / \hat{\sigma} \langle \dots, \underline{b}^T \underline{X}_h, \dots \rangle$

Multivariate outlyingness:

Maximize over projection directions $\underline{b}$

Use outlyingness to define weights for weighted estimator:

$$\begin{aligned}
 D_i &= \max_{\underline{b}: \|\underline{b}\|=1} \left\langle |x_i - \hat{\mu} \langle \dots, \underline{b}^T \underline{X}_h, \dots \rangle| / \hat{\sigma} \langle \dots, \underline{b}^T \underline{X}_h, \dots \rangle \right\rangle \\
 \underline{\hat{\mu}} &= \frac{\text{ave}_i \langle w \langle D_i^2 \rangle \underline{X}_i \rangle}{\text{ave}_i \langle w \langle D_i^2 \rangle \rangle} \\
 \hat{\Sigma} &= \frac{\text{ave}_i \langle w \langle D_i^2 \rangle (\underline{X}_i - \underline{\hat{\mu}})(\underline{X}_i - \underline{\hat{\mu}})^T \rangle}{\text{ave}_i \langle w \langle D_i^2 \rangle \rangle}
 \end{aligned}$$

Disadvantage: Computationally “impossible”!

Way out: see below.

b **S-estimators**

“Backward definition”: Assume that $\hat{\mathbf{\Sigma}}$ is the estimate.

Shape matrix $\mathbf{S} = \hat{\mathbf{\Sigma}} / (\det \hat{\mathbf{\Sigma}})^{1/m}$ determines

outlyingness $d_i^2 = (\underline{x}_i - \underline{\hat{\mu}})^T \mathbf{S}^{-1} (\underline{x}_i - \underline{\hat{\mu}})$.

Scale estimate $\hat{\sigma} \langle d_1^2, \dots, d_n^2 \rangle$.

Minimize this scale! ... and use it as the estimate of size.

$$[\underline{\hat{\mu}}, \hat{\mathbf{S}}] = \arg \min \langle \hat{\sigma}^2 \langle \dots, (\underline{x}_i - \underline{\hat{\mu}})^T \mathbf{S}^{-1} (\underline{x}_i - \underline{\hat{\mu}}), \dots \rangle \mid$$

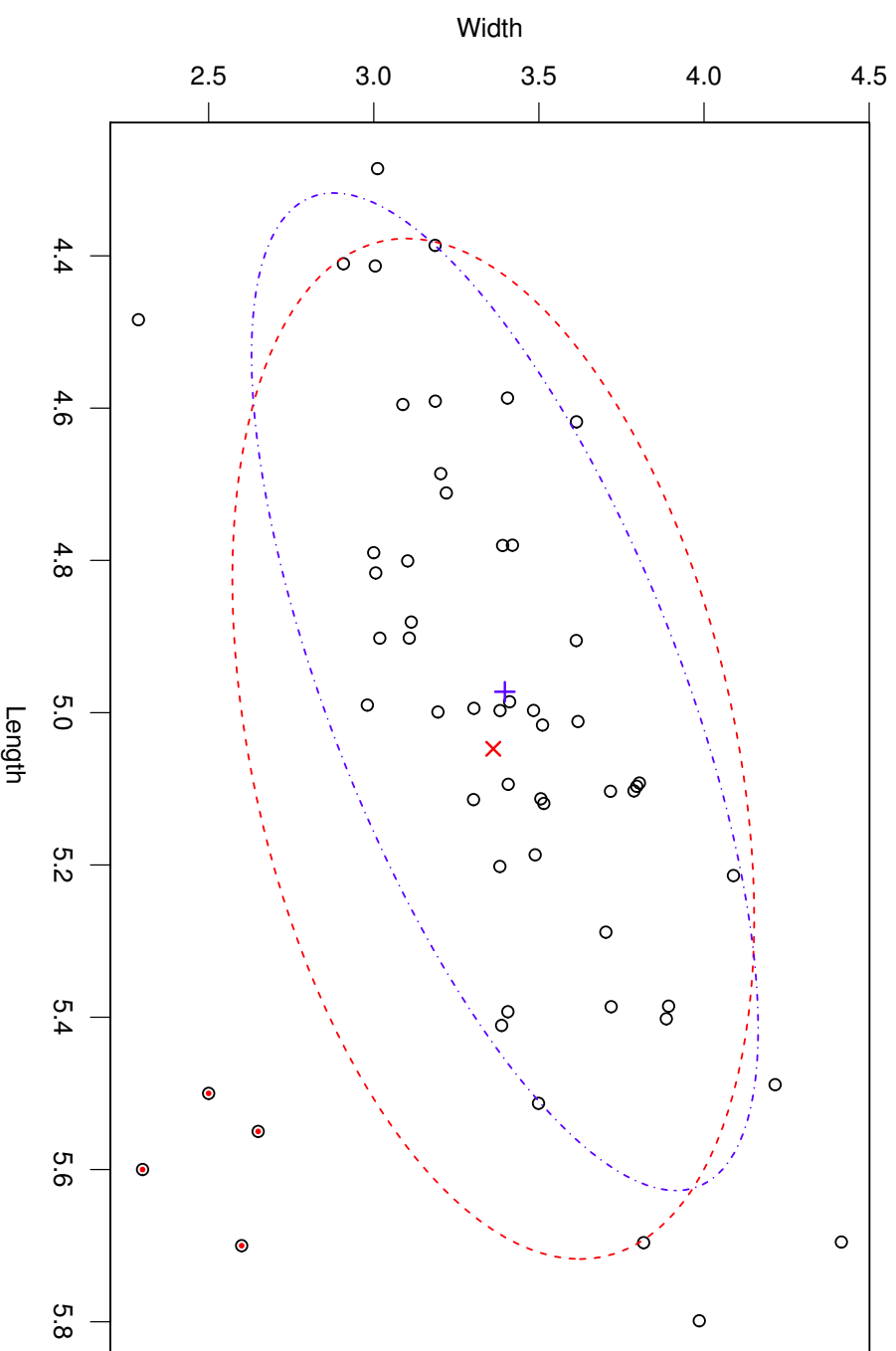
$$\hat{\mathbf{\Sigma}} = \hat{\mathbf{S}} \hat{\sigma}^2 \left\langle \dots, (\underline{x}_i - \underline{\hat{\mu}})^T \mathbf{S}^{-1} (\underline{x}_i - \underline{\hat{\mu}}), \dots \right\rangle^{\mu, \mathbf{S} : \det \langle \mathbf{S} \rangle = 1}$$

$\hat{\sigma}$ can be an M-estimator of scale or another robust estimator.

Early example: **Minimal Covariance Determinant (MCD)**:
uses trimmed scale.

Computation: Similar complexity as Projection estimator.

Finally, a data example: Iris data, first species = 50 obs., first 2 vars. – contaminated by 4 gross errors (colored). Ellipses are sized such that they should contain 80% of the data. Classical estimates (larger, red ...) and MCD (smaller, blue, - - -).



c **Computation of high breakdown point estimators**

Minima needed – over all directions in $\mathbb{R}^m$ or all shape matrices cannot be computed (except for very small dim.).

Need an “intelligent” selection, over which the minimum is “hopefully” close to the global minimum.

At least: Breakdown should be avoided.

Basic idea:

Choose “candidate” for minimum from a random minimal subset of obs. (called “**elemental subset**”).

Hope for candidates that are free of contamination.

This should give a reliable starting point for some kind of local minimization, which will define the estimate.

Projections: m points determine a hyperplane.

Project on the direction perpendicular to it.

Shape matrices: need $m + 1$ points.

Probability to get a “clean” elemental subset:

$$P = 1 - (1 - (1 - \varepsilon)^m)^r \text{ for } r \text{ random choices.}$$

Exercise: Calculate the number of replicates needed

to obtain $P = 0.9$, say, for several combinations of ε and m !

d Messages

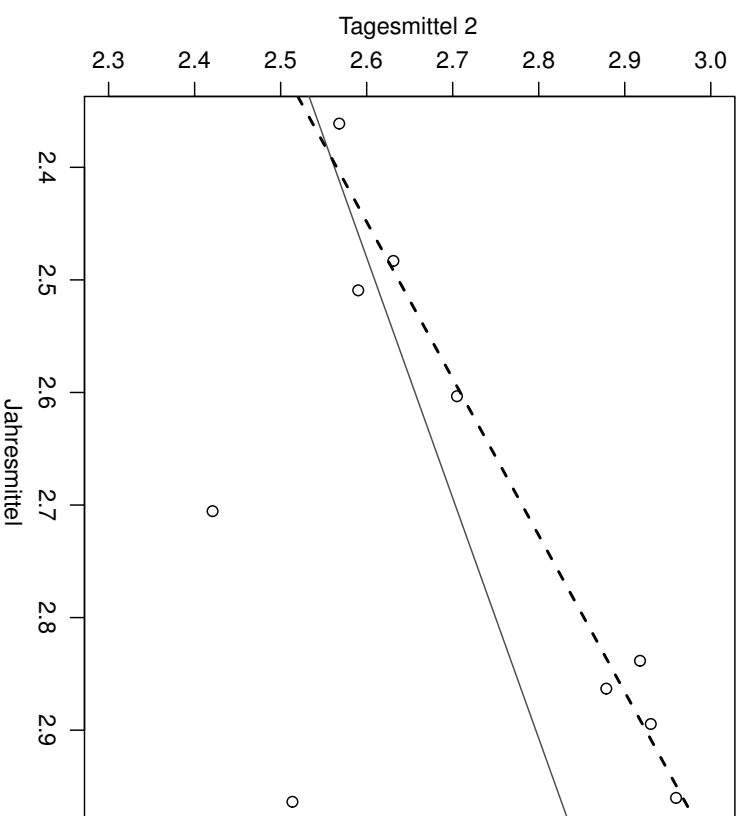
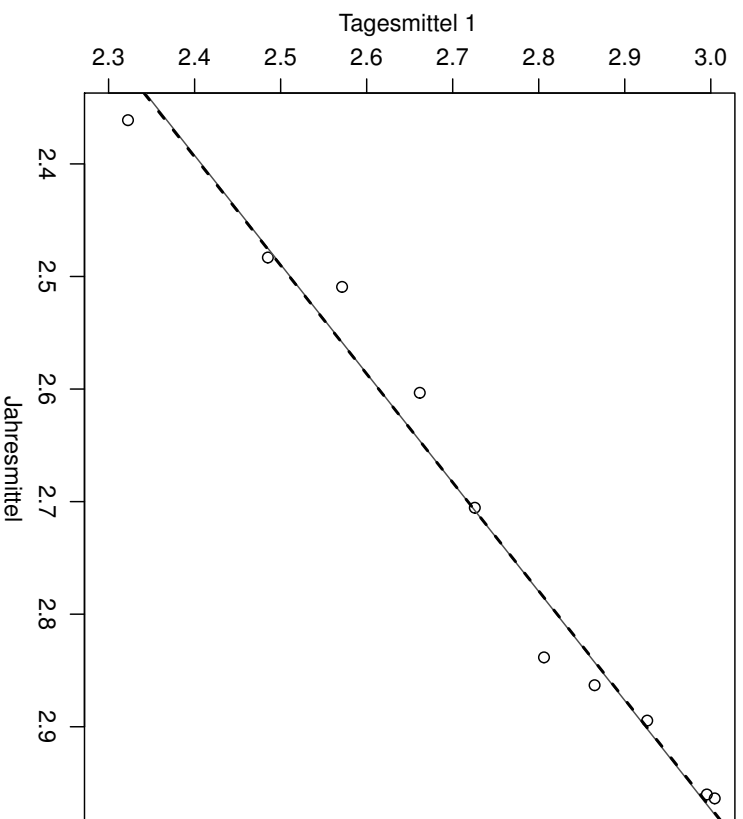
- **Normal distribution:** The “only realistic” model!
 $f(\underline{x}; \underline{\mu}, \underline{\Sigma}) = c \cdot \exp -\frac{1}{2} d^2$, where $d^2 = (\underline{x} - \underline{\mu})^T \underline{\Sigma}^{-1} (\underline{x} - \underline{\mu})^T$
- Affine **transformations**, applied to $\mathcal{N}_m(\underline{0}, \underline{I})$ define the family of normal distributions, $\mathcal{N}_m(\underline{\mu}, \underline{\Sigma})$
 → **transformation family**, → Elliptical distributions,
 → asks for **equivariant estimators**.
- **M-estimators** given by only 3 scalar functions $w^{(\mu)}$, $w^{(\eta)}$, $\psi^{(\tau)}$.

- Breakdown of M-estimators is poor:
 $\leq 1/m$ of the (vector) obs. $\longrightarrow \leq \approx 1/m^2$ of the data!
- Estimators with high breakdown point:
Projection- and **S-estimators**
Computation is “impossible”.
The best we can do is “**elemental subsets**”
which avoids breakdown with a probability that depends on
the dimension m and the number of replicates r .

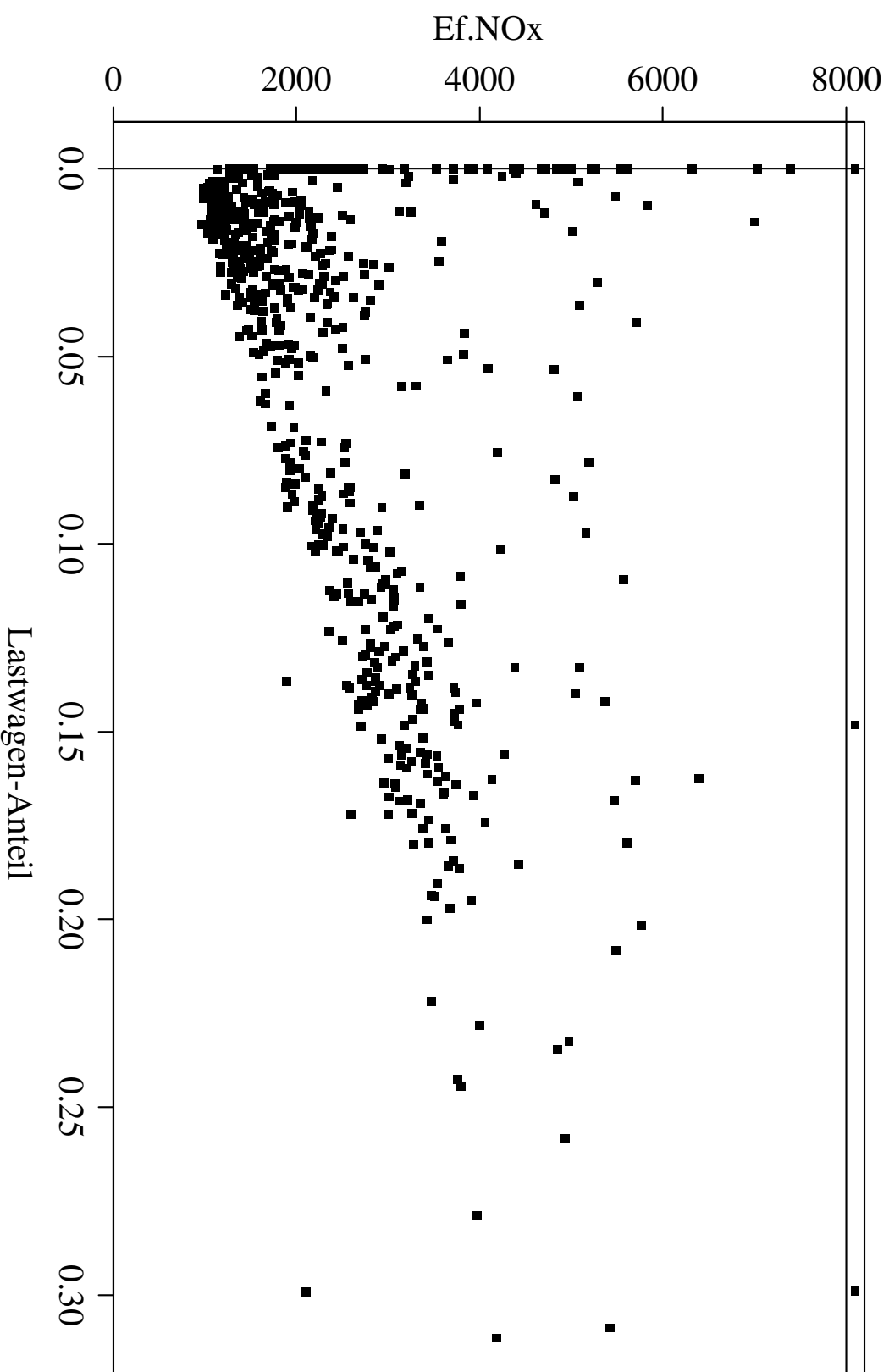
11 Linear Regression

11.1 Examples

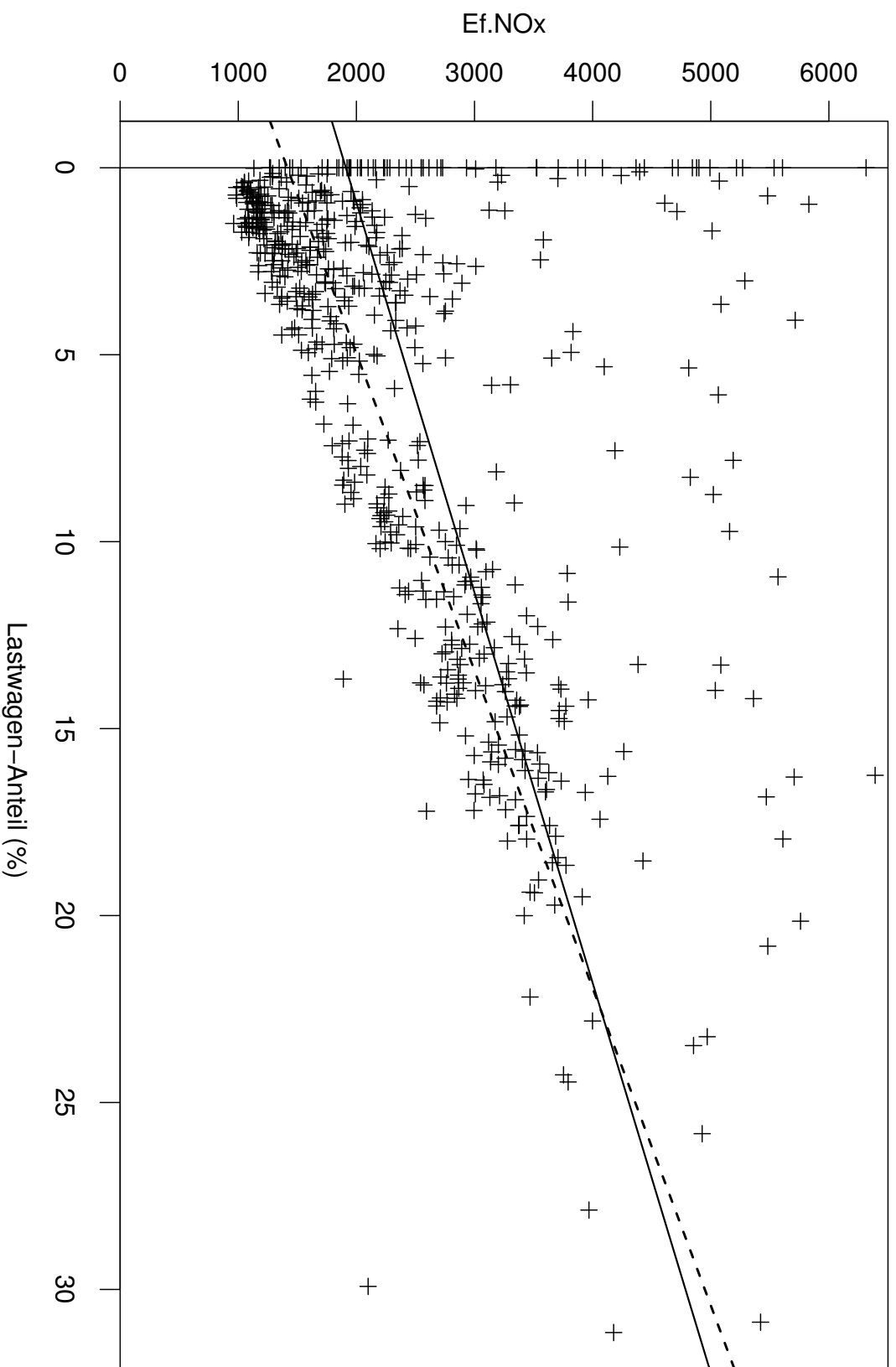
a Example NO₂-means



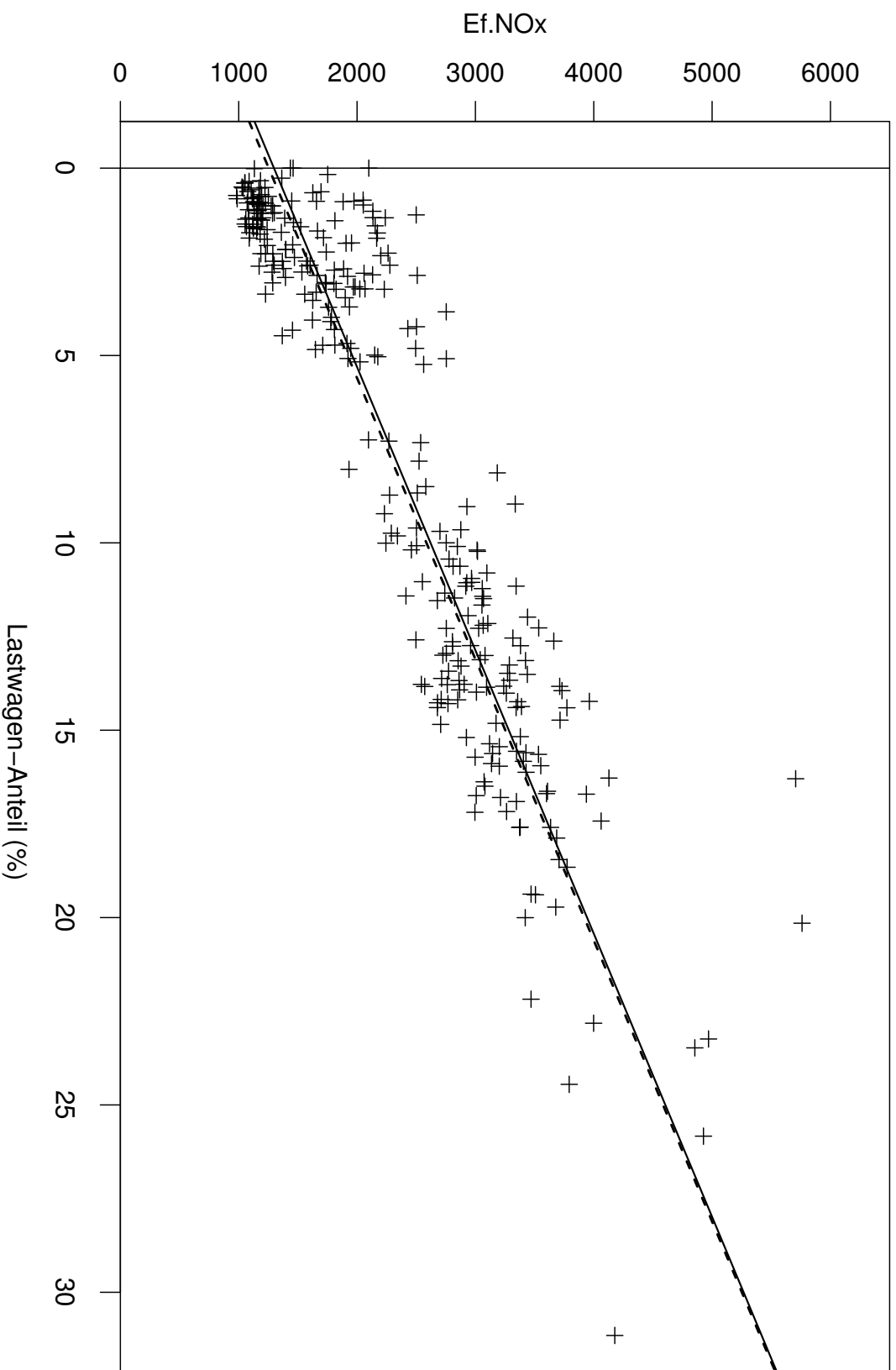
Example Pollutants in a tunnel



Estimated regression lines



Only data with sufficient air velocity



11.2 The Model

a The model of linear regression is

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \dots + \beta_m x_i^{(m)} + E_i$$

Y_i : **target** or **response** variable, i th observation

$x_i^{(j)}$: j th **explanatory** variable (input variable)

β_j : **coefficients** (parameters)

E_i : Random deviation, “**error term**”

Vector notation: $Y_i = \underline{x}_i^T \underline{\beta} + E_i$

Matrix notation: $\underline{Y} = \underline{X} \underline{\beta} + \underline{E}$

$\underline{X}$: **Design matrix**, each row belongs to 1 observation.

Assumptions

- $E_i \sim \mathcal{N}\langle 0, \sigma \rangle$, indep. $\underline{E} \sim \mathcal{N}_n\langle \underline{0}, \sigma^2 \mathbf{I} \rangle$
 - Design $\mathbf{X}$: Different approaches
 - not random, but **fixed values** (designed?)
 - If also random (observational studies)
 - **conditional analysis**
 - also **random** variables, $\underline{x}_i \sim K$ indep.
 - makes asymptotics straightforward
- Distinguish from “observed with error”
- “errors-in-variables” model

b Least Squares

Max.li. for normal assumption = **Least Squares**. Minimize

$$Q(\underline{\beta}) = \sum_i r_i^2, \quad r_i = Y_i - \underline{X}_i^T \underline{\beta}: \text{residual}$$

Scores

$$\underline{s}^{(\beta)} \langle y, \underline{x}; \underline{\beta}, \sigma^2 \rangle = r \cdot \underline{x} / (2\sigma^2) = (y - \underline{x}^T \underline{\beta}) \underline{x} / (2\sigma^2)$$

c Estimation of the scale σ

$\hat{\sigma}^2 = \frac{1}{n-p} \sum_i r_i^2$. Standard deviation of the residuals, up to a finite sample correction $n/(n-p)$ (or $(n-1)/(n-p)$).

Scores (w.r.t σ^2): $s^{(\sigma^2)} \langle y, \underline{x}; \underline{\beta}, \sigma^2 \rangle = (r^2 / \sigma^2 - 1) / (2\sigma^2)$

d Equivariance

- Location-scale family.

- **Regression equivariance:** Linear transf. of the design,

$$\underline{x}_i \mapsto \tilde{\underline{x}}_i = T \underline{x}_i, \text{ leads to } \hat{\underline{\beta}} = T^{-T} \underline{\beta}$$

$$\text{In matrix terms: } \tilde{X} = X T^T \longrightarrow \tilde{\underline{\beta}} = T^{-T} \underline{\beta}$$

If the design matrix X contains a column of 1s, then all affine transf.s $\tilde{\underline{x}}_i = \underline{a} + B \underline{x}_i$ have the form $\tilde{X} = X T^T$.

An estimator $\hat{\underline{\beta}}$ should react in the same way: $\hat{\tilde{\underline{\beta}}} = T^{-T} \hat{\underline{\beta}}$

Consequence: **It is sufficient to consider the case of orthogonal regressors,** $\frac{1}{n} \tilde{X}^T \tilde{X} = I$.

(Let T be such that $T^{-1} T^{-T} = \frac{1}{n} X^T X$!)

11.3 M-estimators for regression

- a Minimize $\sum_i \rho \left((Y_i - \underline{x}_i^T \underline{\beta}) / \sigma \right)$ or solve

$$\sum_i \psi \left(\frac{Y_i - \underline{x}_i^T \hat{\underline{\beta}}}{\sigma} \right) \quad \underline{x}_i = \underline{0}, \quad \psi \langle r \rangle = \rho' \langle r \rangle$$

Scale: See later.

- b **Influence Function**

Random $\underline{x}_i$:

$$\text{IF} \left\langle \underline{x}, y; \hat{\underline{\beta}}, F \right\rangle = \psi \left\langle \frac{Y - \underline{x}^T \hat{\underline{\beta}}}{\sigma} \right\rangle (\tilde{\kappa} \mathbf{C})^{-1} \underline{x}$$

$$\begin{aligned} \tilde{\kappa} \mathbf{C} &= \int \psi \langle u \rangle \underline{x} \underline{x}^T u \mathcal{F}_0 \langle u, \underline{x} \rangle \\ &= \int \psi \langle z \rangle z d\Phi \langle z \rangle \cdot \text{var} \langle \underline{X} \rangle \end{aligned}$$

$\mathbf{C}$ estimated by $\frac{1}{n} \mathbf{X}^T \mathbf{X}$.

c For fixed $\underline{x}_i$, it is not clear what “asymptotic” means.

- Repetition of the same design matrix X , with new $\underline{Y}$?
Very restrictive.

- Sequence X_n with $\frac{1}{n}X_n^T X_n \rightarrow C \rightarrow$ better in principle.
Result as for random $\underline{x}_i$ (?)

- Let dimension increase with n . Realistic, very demanding.

→ We stick with random $\underline{x}_i$!

d **Gross Error Sensitivity.**

infinite! ... unless $\underline{x}_i$ s are bounded (and ψ is bounded)

e **Generalized M-estimation**

$$\sum_i \eta \left\langle (Y_i - \underline{x}_i^T \hat{\underline{\beta}}) / \sigma, \underline{x}_i \right\rangle \underline{x}_i = \underline{0}$$

“ ψ function” depends on $\underline{x}$ –

best through “leverage” = Mahalanobis Distance $h_i = \underline{x}_i^T \mathbf{C}^{-1} \underline{x}_i$

- Bound on $\psi \langle r \rangle$ depends on h_i , $\eta \langle r_i, \underline{x}_i \rangle = \psi_{c(h_i)} \langle r_i \rangle$ “Schweppe type”
- Weights according to h_i , $\eta \langle r_i, \underline{x}_i \rangle = w \langle h_i \rangle \psi_c \langle r_i \rangle$ “Mallows type”

f **Estimation of the scale parameter**

$$\begin{aligned} \text{Scores (w.r.t } \sigma^2): \quad s^{(\sigma^2)} \langle y, \underline{x}; \underline{\beta}, \sigma^2 \rangle &= (r^2 / \sigma^2 - 1) / (2\sigma^2) \\ \longrightarrow \sum_i (r_i^2 / \hat{\sigma}^2 - 1) &= 0 \end{aligned}$$

Robust (see Chapter on scale, 7.1.c, now squared argument)

$$\sum_i \chi \langle r_i^2 / \hat{\sigma}^2 \rangle = 0 \quad \int \chi \langle r^2 / \hat{\sigma} \langle F \rangle^2 \rangle dF \langle r \rangle = 0$$

- Huber, “Proposal 2”: $\chi \langle z^2 \rangle = h_c \langle z^2 \rangle - \kappa$
- Weighted Scale $\chi \langle z^2 \rangle = \omega^2 \langle z^2 \rangle (z^2 - \kappa)$,
 κ such that $\int \omega^2 \langle z^2 \rangle (z^2 - \kappa) dF_1 \langle z \rangle = 0$

g

Design Adaptive Scale (DAS)

Problem: Residuals r_i are not the errors $\mathcal{E}_i!$ – For LSq-estimation,

$$\mathbb{R}_i \sim \mathcal{N}\langle 0, \sigma^2 / (1 - h_{ii}) \rangle$$

where h_{ii} are the diagonal elements of the “hat matrix”

$$H = X(X^T X)^{-1} X^T.$$

Robust: $\text{var} \langle r_i \rangle \approx \tau_i \sigma^2$. τ_i see below.

Solve

$$\begin{aligned} \sum_i \chi \left\langle \frac{r_i^2}{\tau_i^2 \hat{\sigma}^2} \right\rangle &= 0 \\ \sum_i \tau_i^2 \cdot \omega^2 \left\langle \frac{r_i^2}{\tau_i^2 \hat{\sigma}^2} \right\rangle \left(\frac{r_i^2}{\tau_i^2 \hat{\sigma}^2} - \kappa \right) &= 0 \end{aligned}$$

The first τ_i^2 is introduced to get usual $\hat{\sigma}$ back for $\omega = 1$.

Calculation of τ_i . We use $C = \frac{1}{n} X^T X$,

$$\gamma_1 = \mathcal{E} \langle \psi' \rangle, \quad \gamma_2 = \mathcal{E} \langle \psi^2 \rangle$$

$$\begin{aligned}
 r_i = Y_i - \underline{x}_i^T \hat{\underline{\beta}} &\approx Y_i - \underline{x}_i^T \left(\underline{\beta} + \frac{1}{n} \sum_h \text{IF} \left\langle y_h, \underline{x}_h; \hat{\underline{\beta}}, \sigma \right\rangle \right) \\
 &= E_i - \underline{x}_i^T \frac{1}{n} \text{IF} \left\langle y_i, \underline{x}_i; \hat{\underline{\beta}}, \sigma \right\rangle + U_{-i} \\
 U_{-i} &= \underline{x}_i^T \frac{1}{n} \sum_{h \neq i} \text{IF} \left\langle y_h, \underline{x}_h; \hat{\underline{\beta}}, \sigma \right\rangle \\
 &= \frac{1}{n} \sum_{h \neq i} \sigma \gamma_1^{-1} C^{-1} \underline{x}_h \psi \langle E_h / \sigma \rangle \\
 \text{var} \langle U_{-i} \rangle &= \frac{1}{n} \sigma^2 \gamma_2 / \gamma_1^2 \sum_{h \neq i} \left(\underline{x}_i^T C^{-1} \underline{x}_h \right)^2 \\
 &= \frac{1}{n} \sigma^2 \gamma_2 / \gamma_1^2 (h_i - h_i^2)
 \end{aligned}$$

because

$$\sum_h \left(\underline{x}_i^T C^{-1} \underline{x}_h \right)^2 = \underline{x}_i^T C^{-1} \left(\sum_h \underline{x}_h \underline{x}_h^T \right) C^{-1} \underline{x}_i = h_i$$

$$\begin{aligned}
\text{var} \left\langle E_i - \underline{x}_i^T \frac{1}{n} \text{IF} \left\langle y_i, \underline{x}_i; \hat{\underline{\beta}}, \sigma \right\rangle \right\rangle &= \text{var} \left\langle E_i - h_i \sigma \gamma_1^{-1} \psi \left\langle E_i / \sigma \right\rangle \right\rangle \\
&= \sigma^2 g \left\langle h_i, n, \psi \right\rangle \\
\text{var} \left\langle r_i \right\rangle / \sigma^2 &= \tau_i^2 = g \left\langle h_i, n, \psi \right\rangle + \gamma_2 / \gamma_1^2 (h_i - h_i^2)
\end{aligned}$$

Koller & Stahel (2011)

The scale is also used for the standardization of ψ in the estimation of $\underline{\beta}$. $\longrightarrow$ **Re-estimate $\underline{\beta}$!**
 (but do not use τ_i s to standardize ψ ,
 because derivation would fail.)

h

Distribution of estimators

$\hat{\underline{\beta}}$ is asymptotically normal, with

$$\text{var} \left(\hat{\underline{\beta}} \right) = \frac{1}{n} \sigma^2 \cdot \gamma \mathbf{C}^{-1}, \quad \mathbf{C} = \frac{1}{n} \sum_i \underline{x}_i \underline{x}_i^T$$

$$\gamma = \frac{\int \psi^2 \langle u \rangle f_1 \langle u \rangle du}{\left(\int \psi' \langle u \rangle f_1 \langle u \rangle du \right)^2}$$

Often used: Factor $1/(n - p)$ and

“finite sample correction” introduced by Huber (1972).

Replace by DAS!

Versions

- γ estimated from the data: $\hat{\gamma} = \text{ave}_i \langle \psi_i^2 \rangle / \text{ave}_i \langle \psi_i' \rangle^2$
- C “robustified” by using a weighted version $\sum_i w_i \underline{x}_i \underline{x}_i^T / \sum_i w_i$
weights = robustness weights $\psi \langle r_i / \hat{\sigma} \rangle / (r_i / \hat{\sigma})$

i Tests, confidence regions

Wald type: using the asymptotic distribution just described.

Other: see next Chapter

11.4 Breakdown point

a Breakdown of M-estimators

- Without leverage points: depends on design see Maronna, Martin, Yohai.

– Random x s: Breakdown point 0 for non-generalized M-est.

For GM: $< 1/p$ (in order to bound h_i , rob.cov. is needed!)

Again: It is worse! See multivariate model

→ proportion $\approx 1/p^2$ contaminated data values are enough!

b M-estimators with redescending ψ .

→ often more than 1 solution of the equation $\sum_i \psi \langle r_i \rangle \underline{x}_i = \underline{0}$.

Version of minimizing $\sum_i \rho \langle r_i \rangle$ → higher breakdown point.

Computation? → need initial value with high breakdown pt.!

11.5 Estimators with high breakdown point

^a S-estimators

Minimize robust scale estimate!

- Least Median of Squares: Minimize

$$Q\langle \underline{\beta} \rangle = \text{median} \langle (y_i - \underline{x}_i^T \underline{\beta})^2 \rangle$$

- Least trimmed Squares: Minimize

$$Q\langle \underline{\beta} \rangle = \sum_{k=1}^{hn} ((y_i - \underline{x}_i^T \underline{\beta})_{[k]})^2, \quad 0 < h < 1$$

$[k]$: sorted. (Sum over hn smallest squared residuals.)

- More general: M scale est.: Solve

$$\frac{1}{n-p} \sum_i p \left\langle \frac{r_i}{s} \right\rangle = \kappa$$

for $s \neq 0$ $\kappa =$ function of ρ , such that $\hat{\sigma} = 1$ for $E \sim \Phi$.

$\longrightarrow s = s \left\langle \underline{\beta} \right\rangle.$ $\longrightarrow$ Minimize $s \left\langle \underline{\beta} \right\rangle$ over $\underline{\beta}$.

$=$ **S-estimator**

b **Computation?**

$s \langle \underline{\beta} \rangle$ has bad properties (not smooth, local minima)

→ **Elemental subsets:**

p observations determine coefficients $\underline{\tilde{\beta}}$ (without error)

by linear equation system (linear algebra)

(unless they are linearly dependent)

Take a many random “subsamples” r of size p → $s_r, \underline{\tilde{\beta}}_r$.

Choose the $\underline{\tilde{\beta}}_r$ with the smallest s_r as initial value
for a local minimization of s .

c

MM-estimators

S-estimators are inefficient.

→ combination of S- and M-estimators

Use S-estimator as initial value for a redescending M-estimator!

Recommendation: Use “SMDM”-estimator:

1. S-estimate → $\hat{\underline{\beta}}_S, \hat{\sigma}_S$
2. M-estimate with redescending ψ → $\tilde{\underline{\beta}}$
(use $\hat{\underline{\beta}}_S$ as initial value, $\hat{\sigma}_S$ to standardize resid.)
3. DAS scale estimate → $\hat{\sigma}$
4. M-estimate with redescending ψ → $\hat{\underline{\beta}}$
(use $\tilde{\underline{\beta}}$ as initial value, $\hat{\sigma}$ to standardize resid.)
5. Estimate var $\langle \hat{\underline{\beta}} \rangle$ using robustness weights
for estimating $C = \frac{1}{n} \sum_i \underline{x}_i \underline{x}_i^T$ (see 11.3.h).
This allows for “Wald type” inference

Implemented in R as `lmrob(..., method='MS2011', ...)`

11.6 Random Effects Models

a Repeated Measures

$$Y_{ghj} = \underbrace{\beta_0}_{\text{interc.}} + \underbrace{\beta_g}_{\text{betw.eff}} + \underbrace{\tilde{\beta}_j}_{\text{within eff}} + \underbrace{B_{gh}}_{\text{random eff}} + \underbrace{\varepsilon_{ghj}}_{\text{error}}$$

subject h in **treatment group** g , measured under **condition** j .

Regression form:

$$\begin{bmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{112} \\ Y_{131} \\ Y_{132} \\ Y_{211} \\ Y_{212} \\ \dots \\ Y_{232} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ \dots & \dots & \dots \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_g \\ \tilde{\beta}_j \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} B_{11} \\ B_{12} \\ B_{13} \\ B_{21} \\ B_{22} \\ B_{23} \end{bmatrix} + \underline{\varepsilon}$$

b General model

$$\underline{Y} = \underbrace{\underline{X}\underline{\beta}}_{\text{fixed effects}} + \underbrace{\underline{Z}\underline{B}}_{\text{random effects}} + \underbrace{\underline{\varepsilon}}_{\text{error (nugget)}}$$

$$\underline{B} \sim \mathcal{N}_q(\underline{0}, \underline{V}_b \langle \underline{\theta} \rangle), \quad \underline{\varepsilon} \sim \mathcal{N}_n(\underline{0}, \sigma^2 \underline{I}), \quad \underline{B} \perp \underline{\varepsilon}$$

Model includes

- Variance Components,
- Repeated Measures, Split Plot,
- Random Slopes, crossed Random Effects,

- **Geostatistics:** $\underline{Z} = \underline{I}, \quad \underline{V}_b \langle \underline{\theta} \rangle = \eta \underline{V}_\alpha$

where $(V_\alpha)_{ij} = \text{fctn}(|\underline{s}_i - \underline{s}_j|/\alpha)$. $\underline{\varepsilon}$: nugget effect.

c MLE

More precisely: REML (restricted / reduced ML)

$$\sigma^{-2} \underline{\mathbf{Z}}^{\mathrm{T}} \underline{\mathbf{e}} = \mathbf{V}_b^{-1/2} \underline{\mathbf{b}}^*$$

$$\underline{\mathbf{X}}^{\mathrm{T}} \underline{\mathbf{e}} = \underline{\mathbf{0}}$$

$$\underline{\mathbf{e}}^{\mathrm{T}} \underline{\mathbf{e}} = \text{expectation}$$

$$\underline{\mathbf{b}}^{*\mathrm{T}} \mathbf{Q}_{\ell} \underline{\mathbf{b}}^* = \text{expectation,}$$

$$\mathbf{Q}_{\ell} := \mathbf{V}_b^{-1/2} \frac{\partial \mathbf{V}_b \langle \underline{\boldsymbol{\theta}} \rangle}{\partial \underline{\boldsymbol{\theta}}} \mathbf{V}_b^{-1/2}$$

$\underline{\mathbf{b}}^*$: “Estimated” standardized random effect.

Robust estimation:

Replace $\underline{\mathbf{e}}/\sigma$ by $\psi \langle \underline{\mathbf{e}}/\sigma \rangle$, $\underline{\mathbf{b}}^*$, by $\psi \langle \underline{\mathbf{b}}^* \rangle$.

11.7 Other Regression Models

a Quantile Regression

Consider the location problem. ψ -function

$$\psi\langle z \rangle = \begin{cases} b_u & \text{for } z \geq 0 \\ -b_l & \text{for } z < 0 \end{cases}$$

for $b_u = b_l$ ($= 1$). $\longrightarrow$ M-estimator = Median.

What happens if $b_u \neq b_l$?

$\longrightarrow$ **bias** for symmetric distributions! What bias?

$$b_l \int_{z < \hat{\mu}} d\Phi\langle z \rangle = b_u \int_{z > \hat{\mu}} d\Phi\langle z \rangle \longrightarrow \text{quantile } q_p, \quad p = b_l / (b_l + b_u).$$

Regression:

Use this ψ -function for M-estimator in linear regression!

→ Estimates the dependence of the quantile of the distribution of Y (instead of the center of symmetry) on $\underline{x}$!

Applications: medical, risk, ...

b Logistic Regression

Response variable is **binary** (or binomially distributed)

→ logistic regression.

$$P \langle Y_i = 1 \rangle = g \langle \underline{x}_i^T \underline{\beta} \rangle, \quad g \langle \eta \rangle = \exp \eta / (1 + \exp \langle \eta \rangle)$$

Outliers in Y are difficult to conceptualize:

A success for conditions where its probability is minimal.

If downweighted, this must lead to serious bias!

→ **correction!**

→ weights depending on $\underline{x}_i$ (or h_i) are natural.

c Poisson Regression

Response variable is

- a **frequency** of an event $\longrightarrow$ Poisson distribution, Poisson regression

$$Y_i \sim \mathcal{P} \langle \lambda_i \rangle, \quad \log \langle \lambda_i \rangle = \underline{x}_i^T \underline{\beta}$$

For $\lambda_i \leq 3$, no outlier on the left, since $P \langle Y_i = 0 \rangle \geq 0.05$.

- a **count** (# individual plants on plot, animals in flock,...)
 $\longrightarrow$ “overdispersed Poisson”, “quasi-Poisson”,
Negative Binomial

d Generalized Linear Model

... encompasses other useful models besides
Normal, Logistic, Poisson

Messages

Robust Regression

- **Model** for linear regression: $\underline{Y} = \underline{X}\underline{\beta} + \underline{E}$, $\underline{E} \sim \mathcal{N}_n(\underline{0}, \sigma^2 \underline{I})$.
Want regression **equivariance**!
- **M-estimation** is straightforward:
Replace residuals r_i/σ by $\psi(r_i/\sigma)$.
- Scale est.: Weighted scale $\longrightarrow$ **Design Adaptive Scale**.
Useful for efficient **inference** (tests and confidence regions)

- **Breakdown** is a problem! Need high breakdown estimators
→ S -est.
- **Computation** of S -estimators is “impossible”.
→ Probabilistic algorithm, “**elemental subsets**”
- For good properties, combine S - and M -estimators → **MM-, SMDM-est.**
- Robust methods also exist for **Generalized Linear Models**.
- **Try them, use them!**

12 Tests

12.1 Basic Concepts

a How to construct a test

- **Model**, here parametric, $F_{\underline{\theta}}$.
- **Null hypothesis** H_0 and alternatives need **exact model under** H_0 , but only

general idea about alternatives

Common H_0 : **Fix parameter value** $\theta = \theta_0$ (for scalar θ)
 multidim. parameter: split into two parts,

$$\underline{\theta} = [\underline{\theta}_{[1]}^T, \underline{\theta}_{[2]}^T]^T, \quad \underline{\theta} \in \mathbb{R}^p, \quad \underline{\theta}_{[1]} \in \mathbb{R}^q, \quad \underline{\theta}_{[2]} \in \mathbb{R}^{p-q}.$$

$$H_0: \underline{\theta}_{[2]} = \underline{\theta}_{[2],0}$$

Alternative: $\underline{\theta}_{[2]} \neq \underline{\theta}_{[2],0}$

- Invent a “raw” test statistic $\underline{U}$ that has more extreme values under the alternative than under the H_0
Common choice: $\underline{U} = \hat{\underline{\theta}}_{[2]}$

- Get the distribution of $\underline{U}$ under the H_0
This will often depend on “nuisance parameters”
(formally contained in $\underline{\theta}_{[1]}$)

In the light of this distribution, find

- a suitable norm of $\underline{U}$ if it is multivariate
(Testing a multidim. test statistic directly is rare.)
- a standardization that makes the distribution independent of the nuisance parameter(s) in $\underline{\theta}_{[1]}$

This involves the estimator $\hat{\underline{\theta}}_{[1]}$

If a scale param. is involved, this is called “Studentization”

Result: Standardized test statistic $T \left(\hat{\underline{\theta}} \right)$

- Get the **distribution of T under the H_0**
 - hope for a good **approximation by a well-known distr.** like normal, t-, F-distr. Often need “correction factor”.
 - **brute force: simulation** of data under the null hypothesis with estimated nuisance parameters, $\underline{\theta}_{[1]} = \hat{\underline{\theta}}_{[1]}$. “parametric bootstrap”
 - if feasible, use **randomization test**
 - nonparametric bootstrap ??? has problems with testing unless (approx.) “translation model”
 - bootstrap confidence interval
 - **P value $\hat{p}$**

b Properties of tests

Level: $P_0 \langle \hat{p} < \alpha \rangle = \alpha$ by construction.

Functional $\hat{p} \langle F \rangle$. – Level $\hat{p} \langle F_{\underline{\theta}_0} \rangle$: Robust?

Can get an Influence Function, “Level Influence Function”
(Note: This is different from studying the correctness of approximations to the distribution of T under H_0 .)

Power? Need to specify alternative precisely: $F_{\underline{\theta}}$

$\rightarrow$ Power $P_{\underline{\theta}} \langle \hat{p} \langle \hat{F}_n \rangle < \alpha \rangle$.

For normal models, distr. of T under the alternative is often known. Otherwise: Simulation!

Robust? Study $\hat{p} \langle F_{\underline{\theta}} \rangle$. – Asymptotics?

$\hat{p} \langle F_{\underline{\theta}} \rangle \rightarrow 0$ for $n \rightarrow \infty$. Not useful!

c Asymptotics for tests

Alternative must $\rightarrow H_0$ for $n \rightarrow \infty$,
 with the appropriate “rate”, $\underline{\theta} = \underline{\theta}_0 + \frac{1}{\sqrt{n}} \Delta \underline{\theta}$.
 “Contiguous alternatives”.

→ Can get an Influence Function, “Power Influence Function”.

Message: Use robust estimator $\hat{\underline{\theta}}$

d **Recipe for robust tests**

- Use robust estimator $\hat{\underline{\theta}}$ to define $\underline{U}$.
- Summarize & standardize analogously to the classical case.
- Use correction factors to get approximately the correct level when taking critical values from the “classical distribution”
If feasible, adjust degrees of freedom if
a t or F distribution is involved.

e **Huber’s Robust Likelihood Ratio Test**

See History

12.2 Tests for Regression

^a Null Hypothesis

Parameters to be tested: $\underline{\beta}_{[2]} = [\beta_{p-q+1}, \dots, \beta_p]^T$.

$$H_0: \underline{\beta}_{[2]} = \underline{\beta}_{[2],0}.$$

Nuisance parameter: σ (and $\underline{\beta}_{[1]} = [\beta_1, \dots, \beta_{p-q}]^T$)

More general: Give any matrix of “contrasts” A .

Null hypothesis $A\underline{\beta} = \gamma_0$.

By transforming X linearly $\longrightarrow$ back to simpler case.

If regr. equivar. methods are studied $\longrightarrow$ no loss of generality.

- b **“Full model”**: Parameter space for $\underline{\beta}$ unrestricted
all x variables are “in the model”

“Reduced model”: Param. space for $\underline{\beta}$ restricted to $\beta_{[2]} = \beta_{[2],0}$
If $\beta_{[2],0} = \underline{0}$, the last x variables are
“dropped from the model”

If $\beta_{[2],0} \neq \underline{0}$: replace $\underline{Y}$ by $\underline{Y} - \underline{X}^{([2])} \beta_{[2],0}$
→ back to $\beta_{[2],0} = \underline{0}$

c **Wald type test**

$$\underline{U} = \hat{\underline{\beta}}_{[2]} \approx \mathcal{N}_{p-q} \left(\underline{\beta}_{[2],0}, \underline{V}/n \right)$$

Natural standardization: $\underline{T} = \underline{U}^T \underline{V}^{-1} \underline{U}$

Need $\underline{V} = \gamma \sigma^2 (\underline{X}^T \underline{X} / n)^{-1}$

→ Studentization by replacing σ^2 by $\hat{\sigma}^2$

Distribution under H_0 : F distribution. – Exact for classical case.

Robust case: Need

- correction factors
- robust estimator of σ^2 : design adaptive estimator!
- robust “estimator” of $\underline{X}^T \underline{X} / n = \text{ave}_i \langle \underline{x}_i \underline{x}_i^T \rangle$

Use weighted average $\sum_i w_i \underline{x}_i \underline{x}_i^T / \sum_i w_i$, where w_i are the robustness weights $\psi \langle r_i / \hat{\sigma} \rangle / (r_i / \hat{\sigma})$

d Likelihood ratio type test

One way to motivate robust estimators is by

MLE with respect to a long-tailed distribution

Likelihood for fixed σ : $Q\left\langle \underline{\hat{\beta}} \right\rangle = \sum_i \rho\left\langle r_i/\sigma \right\rangle$, $r_i = y_i - \underline{x}_i^T \underline{\hat{\beta}}$

LR test statistic: $U = Q\left\langle \underline{\hat{\beta}}_0 \right\rangle - Q\left\langle \underline{\hat{\beta}} \right\rangle$ where $\underline{\hat{\beta}}_0$ is the estimator of $\underline{\beta}$ under the restriction $\underline{\hat{\beta}}_{[2]} = \underline{\beta}_{[2],0}$.

For classical setup ($\rho\left\langle r \right\rangle = r^2/2$), this leads to $T \sim \chi_{p-q}^2$.

When σ is estimated from the full model $\longrightarrow T \sim F_{p-q,n-p}$.

Robust ρ : Need correction factors.

Note: We want the distribution of U under H_0
for *normal* data, not long-tailed! (?)

e

Score type test

Estimate only the reduced model!

$$\underline{U} = \sum_i \psi \left(\underline{x}_i, y_i; \hat{\underline{\beta}}_0, \hat{\sigma}_0 \right)$$

Standardization: $T = \underline{U}^T \mathbf{J}^{-1} \underline{U} = \underline{U}^T (\mathbf{X}^T \mathbf{X})^{-1} \underline{U} / \hat{\sigma}^2$

$\hat{\sigma}$? This test is currently under study.

13 Time Series, Dependent Observations

13.1 Models

a Classical concepts for time series

$X_1, X_2, \dots, X_t, \dots$ observations

Full description: need joint distribution of all X_t

Expectation $\mathcal{E} \langle X_t \rangle$, variance $\text{var} \langle X_t \rangle$,

Autocovariance: $\text{cov} \langle X_t, X_{t+h} \rangle \longrightarrow \text{autocorrelation}$.

Stationary time series: joint distributions do not depend on t :

$$F \langle X_t, X_{t+1}, \dots, X_{t+h} \rangle = F \langle X_1, X_{1+1}, \dots, X_{1+h} \rangle$$

Weakly stationary time series: **Moments are independent of t :**

$$\mathcal{E} \langle X_t \rangle = \mu, \quad \text{var} \langle X_t \rangle = \sigma^2, \quad \text{cov} \langle X_t, X_{i+h} \rangle = \dots$$

$$\text{autocorrelation function } \rho \langle h \rangle = \text{cov} \langle X_t, X_{t+h} \rangle / \sigma^2.$$

b **Gaussian time series**

Joint distribution is multivariate normal.

→ First & second moments describe distribution completely.

→ only need μ (often assumed $= 0$); σ^2 ; $\rho \langle h \rangle$.

c Autoregression

$$X_t = \alpha X_{t-1} + E_t,$$

$E_t \sim \mathcal{N}\langle 0, \sigma_e^2 \rangle$ independent, indep. of $X_1, \dots, X_t$

More general: AR $\langle p \rangle$: $X_t = \sum_{h=1}^p \alpha_h X_{t-h} + E_t$

“Prototype” of stationary time series model

Stationarity needs assumptions about $\alpha_1, \dots$

$$\mu = 0, \quad \sigma^2 = \sigma_e^2 / (1 - \alpha^2). \quad \text{For } p = 1, \quad \rho \langle h \rangle = \alpha^h$$

- d Basic concepts become difficult when obs. are dependent:
mainly **contamination, Influence Function**

13.2 Contamination

^a Observation Outliers

Replace X_t by gross error Y_t = **replacement outlier**, or

by $\tilde{Y}_t = X_t + Y_t$ = **additive outlier** with $Y_t \sim H$

Model: Binary variable Z_t ,

$$X_t = \tilde{X}_t + Z_t \cdot Y_t$$

$\tilde{X}_t$: un-contaminated time series.

Z_t can be independent $\longrightarrow$ **“isolated outliers”** or
strongly dependent $\longrightarrow$ **“patchy outliers”**

b **Innovation Outliers**

... for AR models: E_t is replaced by gross error
→ the following observations will also be affected.
Effect decays exponentially for isolated outliers.

c **Level shift**

A rather different deviation from assumptions:

$\mathcal{E}(X_t)$ is only piecewise constant.

d **Message**

The concept of contamination / gross error is ambiguous!

13.3 Influence Function, Breakdown

a ???

An IF and a breakdown point for each type of contamination?

→ Many approaches to “robustification”!

See Maronna et al., Ch. 8.

13.4 Other types of dependent observations

a **Spatially correlated data**

X_i is measured at location $\underline{s}_i$.

Correlation of X_i and X_h depends on $\underline{s}_i - \underline{s}_h$,

or, more simply, on $d_{ih} = \|\underline{s}_i - \underline{s}_h\|$

→ **spatial autocorrelation** $\rho\langle d \rangle$.

→ “spatial data”, “geostatistics”.

b **Mixed Effects Models**

see Chapter 11.6.

14 Outlook

14.1 Exam Topics

- Basic notions:
 - Gross error model, long-tailed distributions
 - Estimators as functionals
 - Maximum likelihood
 - M-estimators
 - Huber's Minimax Theorem, without proof
 - Influence function, relation to asymptotic variance
 - Optimality problem of Hampel
 - Optimality result: basic idea, proof
 - Breakdown: finite sample and asymptotic

- Location and Scale:
 - Model
 - Scale: Max. Lik. $\longrightarrow$ Proposal 2 $\longrightarrow$ Weighted Scale
- Multiparameter Problem:
 - Generalization of concepts (M-estimators, IF)
 - Generalization of Hampel's optimality problem
- Multivariate Statistics:
 - Normal and elliptical distributions
 - Equivariance
 - M-estimators

- Regression:
 - Model, Maximum Likelihood
 - M-estimators
 - S-estimators (why?), MM-estimators
 - Tests: Wald type and Likelihood Ratio
- Time Series: What is the difficulty?