INSTITUTE OF ECONOMIC STUDIES, FACULTY OF SOCIAL SCIENCES
CHARLES UNIVERSITY IN PRAGUE (*established 1348*)

ROBUST STATISTICS AND ECONOMETRICS

INSTITUTE OF ECONOMIC STUDIES
FACULTY OF SOCIAL SCIENCES
CHARLES UNIVERSITY IN PRAGUE

JAN ÁMOS VÍŠEK

Week 5

Content of lecture

1. Repetiton is mother of wisdom(Jan Amos Komensky)
2. Estimators alternative to the classical ones
   - Locatin parameter
   - Scale parameter
   - General parameter

The most popular families of robust estimators

The first four lectures established all the prerequisites
for starting to study basic families of (optimal) robust estimators.

Prior to it let, repeat some basic findings from previous lectures.

We'll start with influence function
and the four robustness characteristics of estimators.

Recalling influence function

Returning to IF - the second return

The mathematical part of definition of the influence function is :

$$IF(x, T, F) = \lim_{\delta \to 0} \frac{T\left((1-\delta)F(.) + \delta \cdot \Delta_x\right) - T\left(F(.)\right)}{\delta}.$$

Influence function $IF(x, F, T)$ (IF) predetermines or predestinates (many) properties of estimator.

$$T(F) + \frac{1}{n}\sum_{i=1} IF(x_i, F, T) \quad \text{to} \quad T(F) + \frac{1}{n+1}\sum_{i=1} IF(x_i, F, T).$$

So, $\frac{1}{n+1} IF(x_{n+1}, F, T)$ approximately represents

a contribution of the observation x_{n+1} to the functional $T(F_n)$.

That was the reason

why we have defined a couple of new requirements by it.

The “robustness” characteristics for basic estimators

At the end of the fourth lecture we have established
the influence functions and the robustness characteristics
for

1. the location

and

2. scale parameter.

The IF and “robustness” characteristics for the location parameter

1. Fix $T(F) = \mathbb{E}_F(Z) = \int Z \mathrm{d}F = \int z \cdot f(z)\mathrm{d}z$.
2. $T\left(F(.)\right) = \int z \cdot f(z)\mathrm{d}z = \mu$.
3. $T\left((1-\delta)F(.) + \delta \cdot \Delta_x\right)$
   $= \int z \left\{(1-\delta)f(x) + \delta \cdot \Delta_x\right\} \mathrm{d}z = (1-\delta) \cdot \mu + \delta \cdot x$.
4. Finally, $IF(x, T, F) = \lim_{\delta \to 0} \frac{\delta \cdot (-\mu + x)}{\delta} = -\mu + x$.

The IF and "robustness" characteristics for the location parameter

So, from previous slide $IF(x, T, F) = -\mu + x$.

As the IF isn't bounded,
the "robustness" characteristics of $T(F) = \mathbb{E}_F(X)$ are:

1. The gross error sensitivity $\gamma^* = \sup_{x \in R} |IF(x, T, F)| = \infty$.
2. The local-shift sensitivity $\lambda^* = \sup_{x,y \in R} \left| \frac{IF(x,T,F) - IF(y,T,F)}{x-y} \right| = 1$.
3. The rejection point
$$\rho^* = \inf \{ r \in R^+ \; : \; IF(x, T, F) = 0, \; |x| > r \} = \infty.$$
4. The breakdown point $\varepsilon^* = 0$

(the last characteristic is "derived heuristically"
from the finite version of breakdown point).

The IF and “robustness” characteristics for the scale parameter

1. Fix $T(F) = \boldsymbol{E}_F(Z - \boldsymbol{E}Z)^2 = \int (Z - \boldsymbol{E}Z)^2 \, \mathrm{d}F = \int (z - \boldsymbol{E}Z)^2 \cdot f(z)\mathrm{d}z.$
2. $T\left(F(.)\right) = \int (z - \boldsymbol{E}Z)^2 \cdot f(z)\mathrm{d}z = \sigma^2.$
3. $T\left((1-\delta)F(.) + \delta \cdot \Delta_x\right)$
   $= \int (z - \boldsymbol{E}Z)^2 \left\{(1-\delta)f(z) + \delta \cdot \Delta_x\right\} \mathrm{d}z = (1-\delta) \cdot \sigma^2 + \delta \cdot (x - \boldsymbol{E}Z)^2.$
4. Finally, $IF(x, T, F) = \lim_{\delta \to 0} \frac{\delta \cdot \left(-\sigma^2 + (x - \boldsymbol{E}Z)^2\right)}{\delta} = -\sigma^2 + (x - \boldsymbol{E}Z)^2.$

The IF and "robustness" characteristics for the scale parameter

So, from previous slide $IF(x,T,F) = -\sigma^2 + (x - \mathbb{E}Z)^2$.

As the IF isn't bounded,
the "robustness" characteristics of $T(F) = \mathbb{E}_F(Z - \mathbb{E}Z)^2$ are:

1. The gross error sensitivity $\gamma^* = \sup_{x \in R} |IF(x,T,F)| = \infty$.
2. The local-shift sensitivity $\lambda^* = \sup_{x,y \in R} \left| \frac{IF(x,T,F) - IF(y,T,F)}{x-y} \right| = \infty$.
3. The rejection point
$$\rho^* = \inf \{ r \in R^+ \; : \; IF(x,T,F) = 0, \;\; |x| > r \} = \infty.$$
4. The breakdown point $\varepsilon^* = 0$

(the last characteristic is again "derived heuristically"
from the finite version of breakdown point).

We have discussed the general reasons causing instability of estimator.

Maximum likelihood - solving an extremal problem

$$\hat{\theta}^{(ML,n)} = \underset{\theta \in \Theta}{\arg\max} \prod_{i=1}^{n} f(x_i, \theta)$$

$$\hat{\theta}^{(ML,n)} = \underset{\theta \in \Theta}{\arg\max} \sum_{i=1}^{n} log(f(x_i, \theta))$$

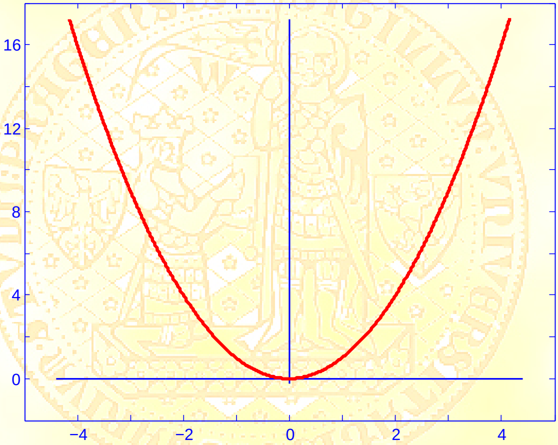
Let $f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{\frac{(x-\mu)^2}{2\sigma^2}\right\}$ and consider only μ

$$\Rightarrow \quad \hat{\mu}^{(ML,n)} = \underset{\mu \in R}{\arg\min} \left\{\sum_{i=1}^{n} (x_i - \mu)^2\right\}$$

The observations with large $(x_i - \mu)^2$ have a large influence on solution.

Evidently, low robustness is consequence of quadratic objective function

We have such objective function.



We should depress influence of large residuals.

Let's study general reasons causing it - an alternative way.

Maximum likelihood - solving the normal equations

$$\hat{\theta}^{(ML,n)} = \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} f(x_i,\theta) = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} log\left(f(x_i,\theta)\right)$$

$$\hat{\theta}^{(ML,n)} = \arg_{\theta\in\Theta} \left\{ \sum_{i=1}^{n} \frac{1}{f(x_i,\theta)} \cdot \frac{\partial f(x_i,\theta)}{\partial\theta} = 0 \right\}$$

Let again $f(x,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{\frac{(x-\mu)^2}{2\sigma^2}\right\}$, i. e. $\frac{\partial f(x_i,\theta)}{\partial\mu} = f(x_i,\mu,\sigma^2) \cdot \frac{(x_i-\mu)}{\sigma^2}$

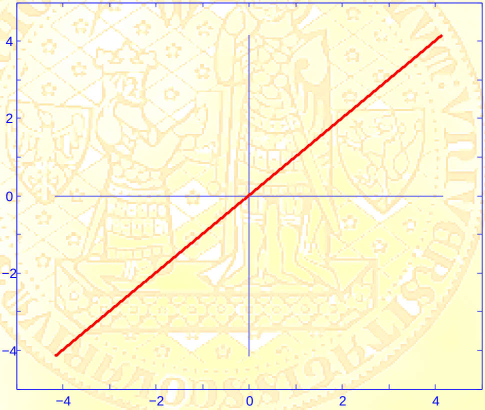
and consider only μ $\Rightarrow$ $\hat{\mu}^{(ML,n)} = \arg_{\mu\in R} \left\{ \sum_{i=1}^{n} (x_i-\mu) = 0 \right\}$

The same conclusion:

The observations with large $|x_i - \mu|$
have a large influence on solution.

Equivalently, low robustness is consequence of identity in normal equations

We have such influence function.



We should depress influence of large residuals

We have recalled everything what will be helpful

and bringing an inspiration for today discussion. So, let’s start.

Reviewing the basic families of robust estimators - location.

Up to now we spoke several times about estimating the location parameter. Let's give its definition:

Let $F(x)$ be a (parent) d. f. . Then $\{F(x-\mu)\}_{\mu \in R}$ is called the family with location parameter.

Let's start with estimating the location parameter:

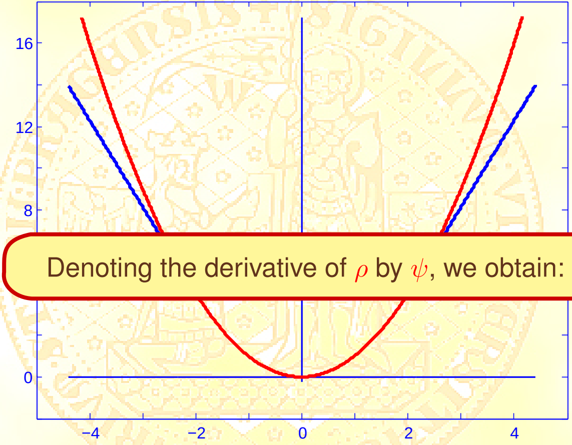
The solution of the extremal problem

$$\hat{\mu}^{(M,n)} = \arg\min_{\mu \in R} \sum_{i=1}^{n} \rho\left(x_i - \mu\right)$$

is called *Maximum likelihood-like estimators of location* or *M-estimators of location*, for short.

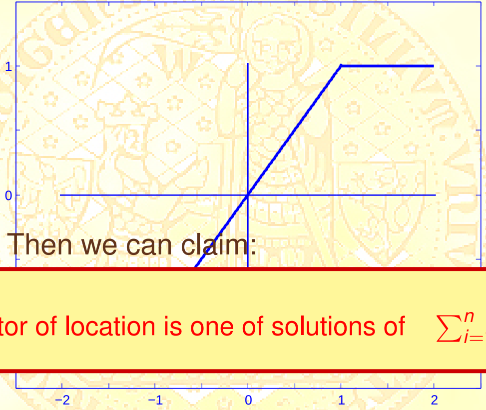
(Example of ρ is on the next slide.)

Reviewing the basic families of robust estimators - location.



Denoting the derivative of ρ by ψ, we obtain:

Reviewing the basic families of robust estimators - location.

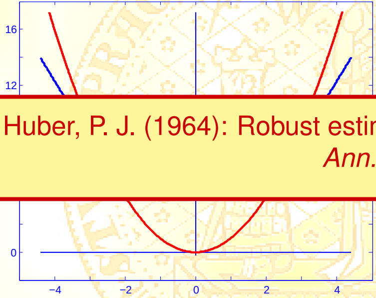


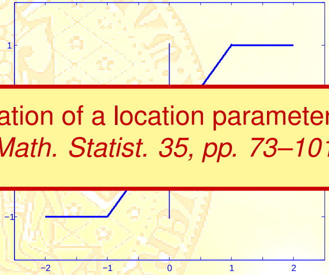
Then we can claim:

M-estimator of location is one of solutions of $\sum_{i=1}^{n} \psi\left(x_i - \mu\right) = 0$.

Reviewing the basic families of robust estimators - location and scale.

The functions ρ and ψ were firstly proposed in:





Huber, P. J. (1964): Robust estimation of a location parameter. *Ann. Math. Statist. 35, pp. 73–101.*

Hence they are usually referred to as Huber's ρ and Huber's ψ.

Reviewing the basic families of robust estimators - scale.

Up to now we spoke also several times about estimating the scale parameter. Let's give its definition:

Let $F(x)$ be a (parent) d. f. . Then $\{F(x/\sigma)\}_{\mu \in R}$ is called the family with scale parameter.

Let's continue with estimating the scale parameter:

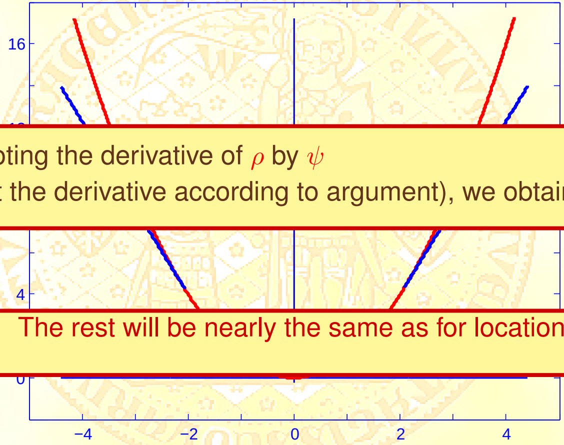
The solution of the extremal problem

$$\hat{\sigma}^{(M,n)} = \arg\min_{\sigma \in R^+} \sum_{i=1}^{n} \rho\left(x_i/\sigma\right)$$

is called *Maximum likelihood-like estimators of scale* , or *M-estimators of scale* for short.

(An example of ρ is the same - see the next slide.)

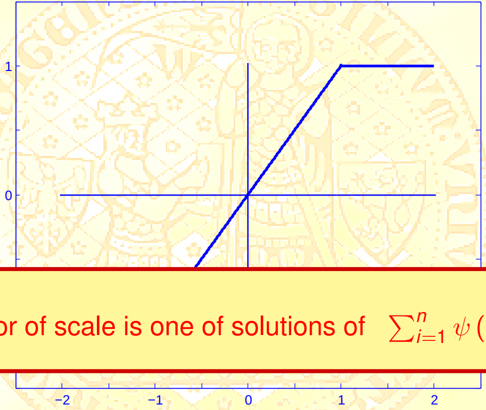
Reviewing the basic families of robust estimators - scale.



Denoting the derivative of ρ by ψ
(just the derivative according to argument), we obtain again:

The rest will be nearly the same as for location:

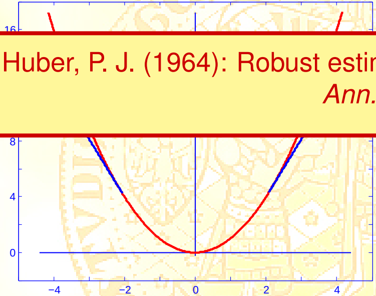
Reviewing the basic families of robust estimators - scale.



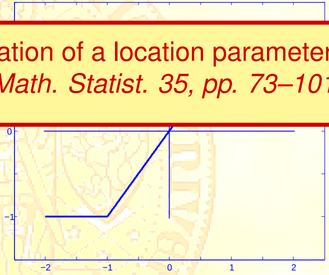
M-estimator of scale is one of solutions of $\sum_{i=1}^{n} \psi\left(x_i/\sigma\right) = 0$.

Reviewing the basic families of robust estimators - scale.

Recalling once again that ρ and ψ were proposed in pioneering paper:



Huber, P. J. (1964): Robust estimation of a location parameter. *Ann. Math. Statist. 35, pp. 73–101.*



We can start to consider a general parameter.

Reviewing the basic families of robust estimators - general parameter.

Now, let's consider a general parameter family:

In what follows, let $\{F(x,\theta)\}_{\theta\in\Theta}$ and $\{f(x,\theta)\}_{\theta\in\Theta}$ be families of d. f.'s and densities, respectively.

Then:

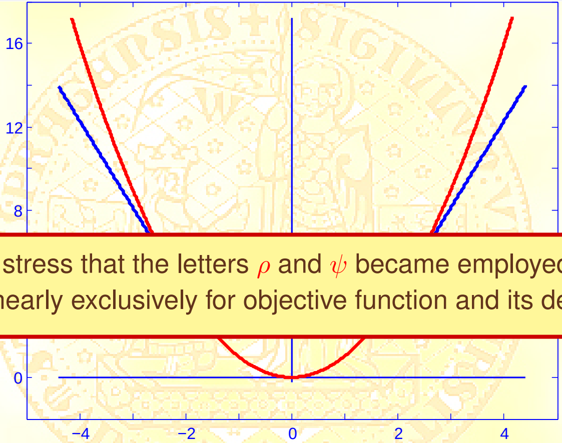
The solution of the extremal problem

$$\hat{\theta}^{(M,n)} = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} \rho\,(x_i,\theta)$$

is called *Maximum likelihood-like estimators of the parameter* θ or *M-estimators of* θ, for short.

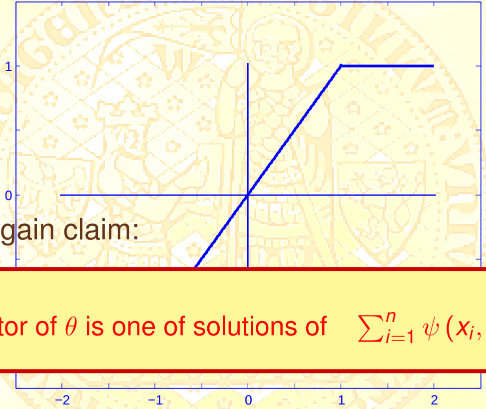
(We can use the same ρ as for location and scale.)

Reviewing the basic families of robust estimators - general parameter.



Let's stress that the letters ρ and ψ became employed nearly exclusively for objective function and its derivative.

Reviewing the basic families of robust estimators - general parameter.



Then we can again claim:

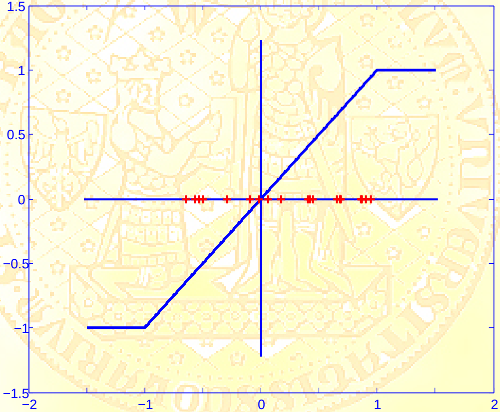
M-estimator of θ is one of solutions of $\sum_{i=1}^{n} \psi(x_i, \theta) = 0$.

M-estimators - general parameter.

What is a drawback of *M*-estimators ?

M-estimators - general parameter.

To learn it, let's consider the following data:



M-estimators - general parameter.

How to cope with the problem ?

M-estimators - general parameter.

Let $\hat{\sigma}$ be a (highly) robust estimator of the standard deviation of data x_i's and solve:

$$\sum_{i=1}^{n} \psi\left(x_i/\hat{\sigma}, \theta\right) = 0.$$

The solution $\hat{\theta}^{(M,n)}$ is then scale-equivariant.

An example of such estimator is

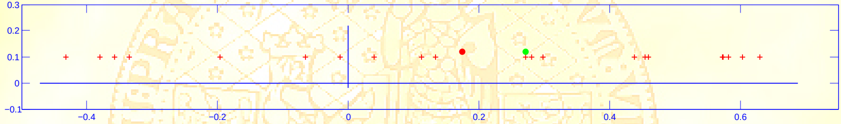
$$\hat{\sigma}_{MAD} = 1.483 \operatorname{med}_i \left\{ \left| x_i - \operatorname{med}_j (x_j) \right| \right\}.$$

(A comparison of $1.483 * MAD$ and s_n is on the next slide.)

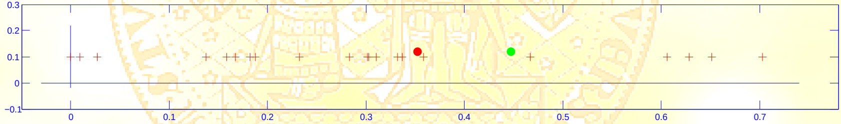
Demonstrating abilities of MAD

Observe the mean • and the median •
and standard deviation s_n • and $\hat{\sigma}_{MAD}$ •.

Non-contaminated data - normal d. f. $\mu = 0$ and $\sigma^2 = \frac{1}{9}$

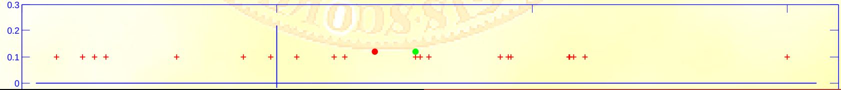


Absolute values of data



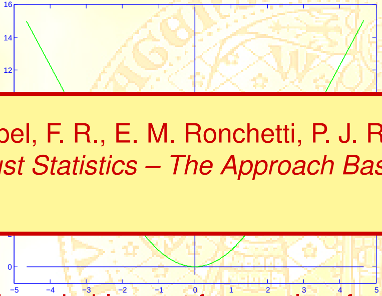
Conta

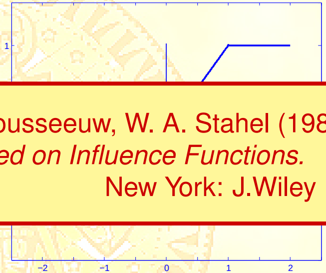
at point 1



Reviewing the basic families of robust estimators - general parameter.

For the nearly exhaustive explanation see:





Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, W. A. Stahel (1986): *Robust Statistics – The Approach Based on Influence Functions.* New York: J.Wiley *&* Son.

This is probably most frequently referred book having an extremely interesting first chapter - which is without mathematics and can be read as a detective story.

The rest of book is mostly beyond the scope of this basic lecture but we shall (without the proofs) quote some results from it.
(Let's give only one example.)

Example of searching for an optimal *M*-estimator of location.

Assume the underlying parent d. f. $F(x)$

with differentiable density $f(x)$ which is symmetric

and ask for the *M*-estimator solving the location problem

and having following properties:

1. The efficiency as high as possible,
2. a priori given gross-error sensitivity.

The solution is given by

$$\psi(x) = \max\{-b, \min\{b, -f'(x)/f(x)\}\}.$$

An example of the likelihood function $f'(x)/f(x)$

Let's consider the standard normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{x^2}{2}\right\},$$

i. e.

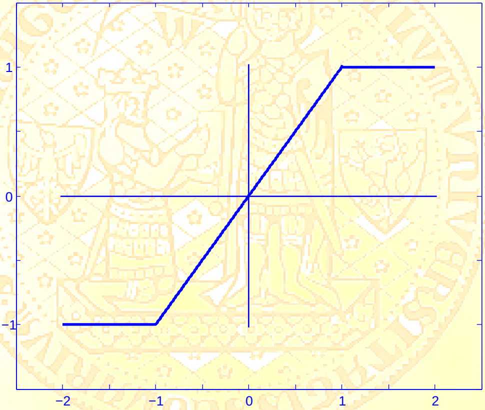
$$f'(x) = \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{x^2}{2}\right\} \cdot \{-x\}, = f(x) \cdot \{-x\},$$

hence

$$-\frac{f'(x)}{f(x)} = x.$$

Example of searching for an optimal M-estimator of location.

Specifying $F(x) = \Phi(x)$, we obtain



Example of searching for an optimal *M*-estimator of location.

Assume the underlying d. f. $F(x)$ with differentiable density $f(x)$ which is symmetric and ask for the *M*-estimator having:

1. The efficiency as high as possible,
2. a priori given gross-error sensitivity,
3. a priori given rejection point c.

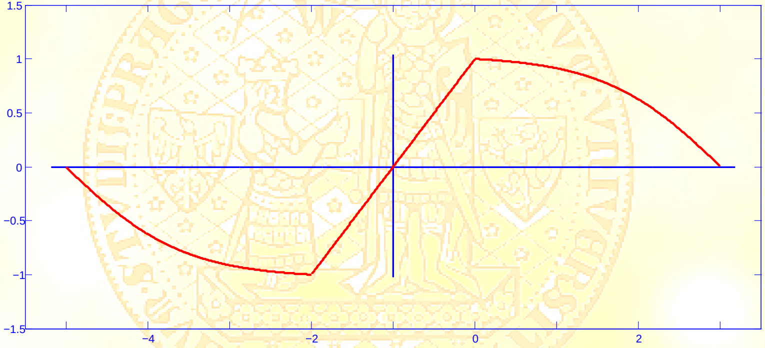
The solution is given by

$$\psi(x) = \max\{-h(x), \min\{h(x), f'(x)/f(x)\}\}$$

where the shape of the function $h(x)$ is given by employment of *tangh*(x) - see next slide.

Example of searching for an optimal *M*-estimator of location.

Specifying $F(x) = \Phi(x)$, we obtain



Other types of estimators

Estimators based on linear (hence the name) combination of order statistics - *L*-estimators

Estimating the location

Observations $z_1, z_2, ..., z_n \quad \Rightarrow \quad \underbrace{z_{(1)} \leq z_{(2)} \leq ... \leq z_{(n)}}_{\textit{These statistics are called order statistics}}$

$\hat{\mu}^{(L,n)} = \sum_{i=1}^{n} a_i \cdot z_{(i)}$

where a_i's are a priori selected weights.

Estimating the scale

Put $r_i = \left| z_i - \hat{\mu}^{(L,n)} \right| \quad \Rightarrow \quad r_{(1)} \leq r_{(2)} \leq ... \leq r_{(n)}$

$\hat{\sigma}^{(L,n)} = \sum_{i=1}^{n} b_i \cdot r_{(i)}$

where b_i's are again a priori selected weights.

Other types of estimators

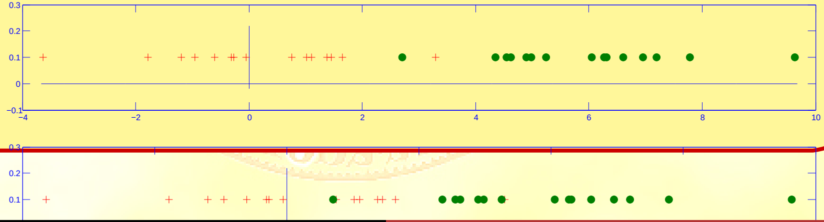
Estimators based on rank statistics (hence the name)
R-estimators

Estimating the location

Let $x_1, x_2, ..., x_n$ be observations, $\Delta \in R$ and consider data

$$x_1, x_2, ..., x_n, 2\Delta - x_1, 2\Delta - x_2, ..., 2\Delta - x_n.$$

The situation can looks like this for
$\Delta = 3 \Delta = 2 \Delta = 1 \Delta = 0 \Delta = -0.5 \Delta = -0.25 \Delta = -0.125$



Other types of estimators

Estimators based on rank statistics (hence the name)
R-estimators

Estimating the location

Let $x_1, x_2, ..., x_n$ be observations and $\Delta \in R$.
Let R_i be the rank of the i-th observations in the pooled sample

$$x_1, x_2, ..., x_n, 2\Delta - x_1, 2\Delta - x_2, ..., 2\Delta - x_n$$

and put
$$S_n(\Delta) = \frac{1}{n} \sum_{i=1}^{n} a_n(R_i)$$

where $a_n(R) = n \int_{\frac{R-1}{n}}^{\frac{R}{n}} \Psi(u)\mathrm{d}u$ with $\Psi(u) = \Psi(1-u)$ $(\rightarrow \int_0^1 \Psi(u)\mathrm{d}u = 0)$.
Then put

$$\hat{\mu}^{(R,n)} = \underset{\Delta \in R}{\arg\min}\ S_n(\Delta).$$

Other types of estimators

Minimal distance estimators
Estimating a general parameter

Let $\{F_\theta(x)\}_{\theta\in\Theta}$ $\quad x_1, x_2, ..., x_n \quad \rightarrow \quad F^{(n)}(x)$ empirical d. f.

$\mathcal{D}(F, G)$ a distance on the space of all d. f.'s,
e. g. Prokhorov metric π or some I-divergence

$$\hat{\theta}^{(MD,n)} = \arg\min_{\theta\in\Theta} \mathcal{D}(F_\theta, F^{(n)})$$

Kullbac-Leibler divergence

Let F and G are absolutely continuous d. f.
and f and g the corresponding densities, respectively.

Then

$$KL(F, G) = \int log \left(\frac{g(x)}{f(x)} \right) \cdot g(x) \mathrm{d}x$$

is called *Kullbac-Leibler divergence*.

By Jensen's inequality we easy prove that

$$KL(F, G) \geq 0.$$

The problem with orthogonality - Igor Vajda.

Jensen's inequality

Let $h(x)$ be convex and X a random variable having the mean value $\mathbb{E}X$. Then

$$\mathbb{E}\{h(X)\} \geq h(\mathbb{E}X).$$

Proof: As (see the next slide)

$$h(x) \geq h(\mathbb{E}X) + b \cdot (X - \mathbb{E}X),$$

we have

$$\mathbb{E}\{h(X)\} \geq h(\mathbb{E}X) + b \cdot \mathbb{E}(X - \mathbb{E}X) = h(\mathbb{E}X).$$

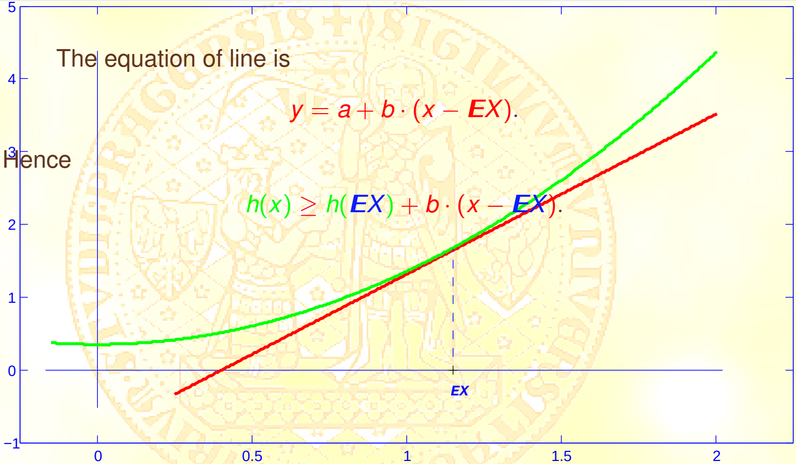
Jensen's inequality

The equation of line is

$$y = a + b \cdot (x - \boldsymbol{EX}).$$

Hence

$$h(x) \geq h(\boldsymbol{EX}) + b \cdot (x - \boldsymbol{EX}).$$



Kullbac-Leibler divergence

By Jensen's inequality we easy prove that

$$KL(F,G) = \int log\left(\frac{g(x)}{f(x)}\right) \cdot g(x)\mathrm{d}x = \mathbb{E}_G log\left(\frac{g(x)}{f(x)}\right) = -\mathbb{E}_G log\left(\frac{f(x)}{g(x)}\right)$$

As $-log(z)$ is convex function, we have

$$KL(F,G) = -\mathbb{E}_G log\left(\frac{f(x)}{g(x)}\right) \geq log\left(\int \frac{f(x)}{g(x)} g(x)\mathrm{d}x\right) = 0.$$

I-divergence

Let F and G are absolutely continuous d. f.,
f and g the corresponding densities, respectively,
and $h(z)$ a convex function.

Then

$$I(F,G) = \int h\left(\frac{g(x)}{f(x)}\right) \cdot g(x)\mathrm{d}x$$

is called *I-divergence*.

By Jensen's inequality we again easy prove that

$$I(F,G) \geq 0.$$

Frequently used divergences

A great contribution to study of I-divergences:

Csiszár, I. (1975): I-divergence geometry of probability distributions and minimization problems. *Ann. Probab. 3, 146-158*.

One of the most frequently employed function $h(z)$

$$h(z) = \frac{z^\alpha - 1}{\alpha}, \qquad \alpha \in (0, 1].$$

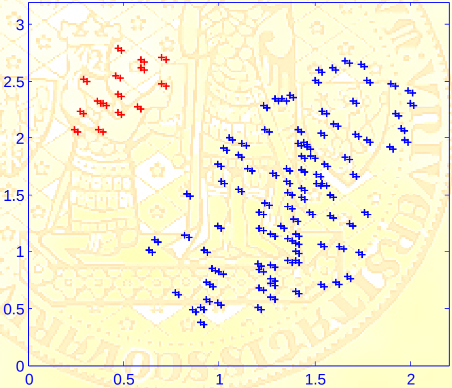
Box, G. E. P., D. R. Cox (1964): An analysis of transformations. *Journal of the Royal Statistical Society, Series B, 26, 211 - 243.*

The *I-divergence* is then called α-divergence.

Other types of estimators

Minimal volume estimator

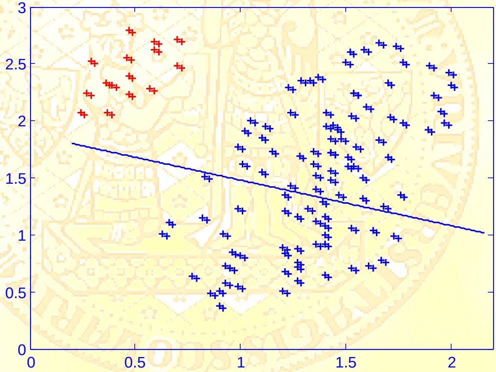
Estimating a regression model



Other types of estimators

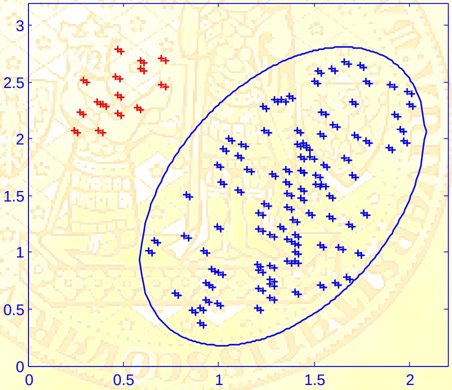
By the way, the Ordinary Least Squares gives

Estimating a regression model



Other types of estimators

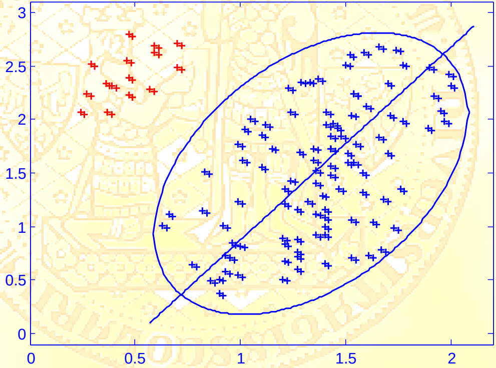
Minimal volume estimator
Estimating a general parameter



Other types of estimators

So, it seems we have nearly unmistakeable tool

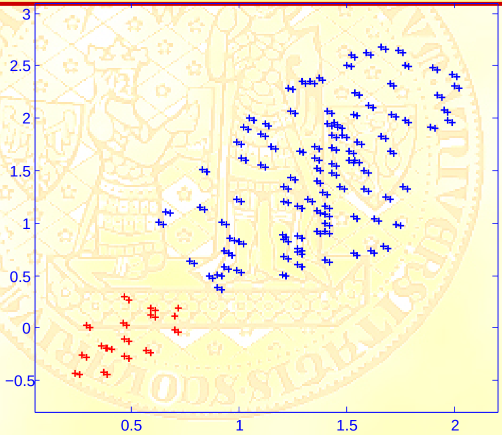
Estimating a regression model



Other types of estimators

But what about such a situation

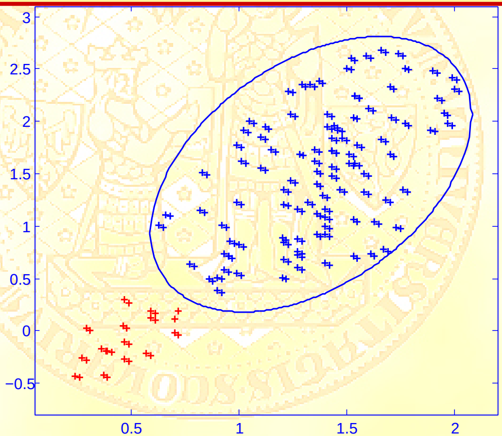
Estimating a regression model



Other types of estimators

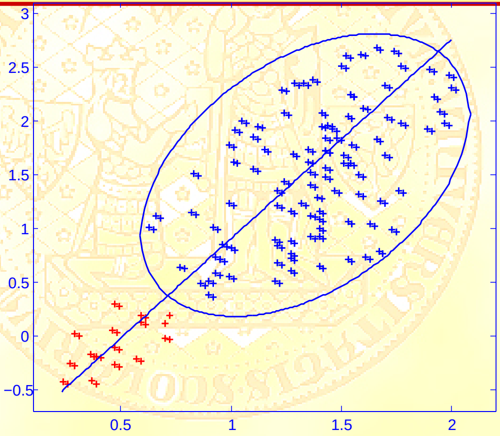
We can proceed as in previous case

Estimating a regression model



Other types of estimators

And the model is reasonable
but we lose idly some information



THANKS FOR ATTENTION