Lecture 10
An Introduction to Robustness

1. Basic idea of robustness

All along this course, we have studied theories about how to construct optimal procedures (be they Likelihood-based or Bayesian) when certain parametric model $F(X, \theta)$ is given. These theories say nothing about the behaviour of the optimal procedures when the models are only **approximately** valid. By **approximately** here we mean that we believe that the sample has been generated by a model that is **not exactly** one from the set $\{F(X, \theta), \theta \in \Theta\}$ but nevertheless, the "distance" between the actual model and some representative of the family $\{F(X, \theta), \theta \in \Theta\}$ is **relatively small.** (For the purpose of this course, we avoid the precise definition of distance between distributions here and leave its interpretation to your intuition). Such situations occur often in practice. For example, we may have a sample that comes from a normal distribution but there might be a (small number of) outliers that do not follow this model. If this is the case, we could model our data as a mixture of normal with another (unknown) "contaminating" distribution (see Section 4 below for a more precise description).

The question is how do we do inference in cases like these where our parametric model is believed to be only **approximately** valid? Of course, one could suggest in cases like these, to decide to simply *abandon* the parametric model since it is not exact and does not include all distributions that may have generated the data. One could argue that a non-parametric approach should be adopted instead in order to be "closer to reality" with our set of models. But, again relying on our intuition, we understand that adopting the purely non-parametric approach would *not* properly address the described situation since it would mean to allow for *too many* modelling distributions $G(X)$ to be considered, some of them being "too far away" from any of the distributions in the set $\{F(X, \theta), \theta \in \Theta\}$. The inference methods applied for such a large class of distributions would be with low efficiency when applied to a model that is "not too far away" from the distributions in the set $\{F(X, \theta), \theta \in \Theta\}$, and the assumption about *relatively small* deviation form a baseline parametric model would be lost in a completely nonparametric treatment. The proper approach would be the robustness approach where we still keep the idea about the *ideal parametric model* but allow for small deviations from it. Speaking loosely, nonparametric statistics allows "almost all" possible probability distributions as models. Classical parametric statistics allows only a very "thin" finite-dimensional subset of probability distributions, i.e the ideal parametric model of interest for which usually optimal inferences are available. Robust statistics allows a full-dimensional neighbourhood of a parametric model, thus being more realistic and yet, at a price of a relatively small loss of efficiency at the ideal model, tries to provide almost the same advantages as a strict parametric model in a "broader" neighbourhood of the ideal parametric model.

The problem in robustness is to construct estimators that are **close to efficient** if the parametric model holds but are at the same time **less sensitive** to small deviations from the ideal model.

2. Simple example

One of the simple examples to start with, is estimating the location parameter of a continuous symmetric distribution. Assume a sample $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ is available from a

location parameter family

$$F(x,\theta) = F(x-\theta), \theta \in R^1.$$

Here θ is interepreted as the location parameter.

Denote the density by $f(x,\theta) = f(x-\theta)$. If F is a normal distribution for example then θ coincides with its mean, median and mode. As we know, in this case the estimator $\bar{\mathbf{x}}$ is efficient for θ for any fixed sample size.

But assume now that F is Cauchy with a density

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}.$$

$\varphi_{x\,normal}(t) = e^{i\mu - \frac{1}{2}\sigma^2 t^2}$

$\varphi_{\bar{x}}(t) = E\left[e^{i\frac{t}{n}\sum_{i=1}^{n} X_i}\right] = E\left[e^{i\frac{t}{n}X_1} \cdot \ldots \cdot e^{i\frac{t}{n}X_n}\right]$

$\left[E\left(e^{i\frac{t}{n}X_n}\right)\right]^n = \left[e^{i\theta\frac{t}{n} - \frac{|t|}{n}}\right]^n$

$= e^{i\theta t - |t|}$

$= \varphi_{cauchy}(t)$

$\bar{X} \to$ Cauchy

Then $f(x,\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$. The parameter θ in this model does **not** coincide with the mean of the distribution (in fact, the Cauchy distribution does not have a finite mean) but coincides with its median and mode. It can be shown that $\bar{\mathbf{x}}$ is even **not** consistent for θ in the Cauchy model! If we consider the characteristic function of a single variable X_1 from the Cauchy distribution $\varphi_{X_1}(t) = E(e^{itX_1})$ and the characteristic function of the average of n i.i.d. Cauchy variables then it holds:

$$\varphi_{\bar{X}}(t) = \varphi_{X_1}(t) = e^{i\theta t - |t|}$$

and this means that the average of n i.i.d. Cauchy variables has **the same** distribution as any one of them; hence $\bar{X}$ is not consistent for the location parameter θ of the Cauchy distribution.

The reason for the good behaviour of $\bar{\mathbf{x}}$ as an estimator of location parameter θ in the normal family and for its "bad" behaviour in the Cauchy family are the **heavy tails** of the Cauchy distribution, i.e. it allows with a large probability for very large (in absolute value) realisations to occur. Because of this, we would decide to ignore the observations with a large absolute value and use the **empirical median instead of the sample mean** when estimating the location parameter of the Cauchy distribution. The empirical median $\tilde{\theta}_n$ is **not sensitive** to large realisations in the tail of the distribution, hence it is more robust as a location parameter estimator.

Assume for simplicity that sample size n is odd. From theoretical derivations it can be shown that for a symmetric F (i.e. $F(0) = 1/2$) with a density $f(0) > 0$,

$* f(x-\theta)$

$$\sqrt{n}(\tilde{\theta}_n - \theta) \to N(0, \frac{1}{4f^2(0)}).$$

Normal: $f(0) = \frac{1}{\sqrt{2\pi}}$

Median: $\sqrt{n}(X_{med} - \theta) \xrightarrow{d} N(0, \pi)$

Mean: $\sqrt{n}(\bar{X} - \theta) \xrightarrow{d} N(0,1)$

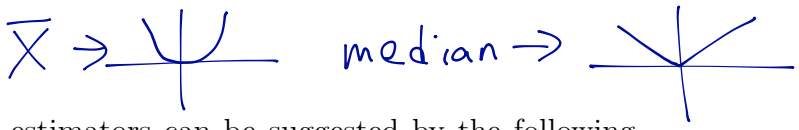
$\bar{X} \sim N(\theta, \frac{1}{n})$

LOSSING efficiency!

$\frac{\pi}{2} > 1$!

Hence, if F is a standard normal, the asymptotic variance of the median will be $\frac{\pi}{2}$ whereas the variance of $\bar{x}$ would be one in this case. This means that $\tilde{\theta}_n$ is **not** asymptotically efficient when the family $f(x,\theta)$ is the normal family **but** it is consistent and even asymptotically normal as a location parameter estimator **simultaneously** for both the normal **and** the Cauchy family. Note that, in contrast, the arithmetic mean **is** efficient for estimating the location parameter if the family was normal **but**, as pointed out, is **very bad** if the family was Cauchy. Hence, the median is more robust than the mean when estimating the location parameter- at the price of some efficiency loss at the normal family, the median still behaves reasonably well even if the family was not normal (i.e. it does a **reasonable compromise.)**

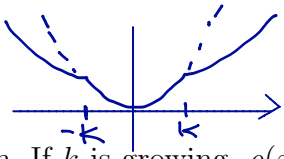
Often, in practise it would be not sure if the family is normal, of Cauchy, or another location family with tails that are heavier than the ones of the normal but less heavy than the ones of the Cauchy family. So how to estimate the location parameter then? Could we also consider another robust estimators besides the median? Another compromise is the α- **trimmed mean** $\bar{x}_\alpha = \frac{1}{n-2k}[x_{(k+1)} + x_{(k+2)} + \dots x_{(n-k)}]$ where $k/n = \alpha < 1/2$ (i.e. we trim symmetrically $2\alpha 100\%$ of the observations and average the rest). It can be shown that this estimator also has an asymptotically normal distribution and when α is small, it has a high efficiency at the normal family (higher than the median). In this example, we trim some of the extreme observations, i.e. we attach **zero weight** to them thus making the location estimator more robust.



3. M-estimators of location.

Alternative way of generating robust estimators can be suggested by the following observation. It is well known that $\bar{x}$ minimises the sum $\sum_{i=1}^n (x_i - a)^2$ for all possible values of a whereas $\tilde{\theta}_n$ minimises the sum $\sum_{i=1}^n |x_i - a|$. When the information is incomplete, it would be a good idea to choose to minimise a function in the form $\sum_{i=1}^n \rho(x_i - a)$ where ρ is symmetric nonnegative and $\rho(0) = 0$. An example of such a function is given by:

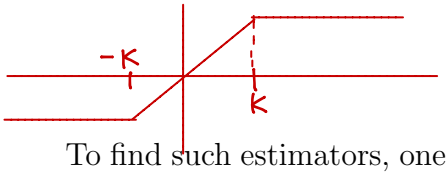
$$\rho(x) = \begin{cases} \frac{1}{2}x^2 \text{ if } |x| \le k \\ k|x| - \frac{1}{2}k^2 \text{ if } |x| \ge k, \end{cases}$$



k being a positive constant. This is Huber's famous ρ-function. If k is growing, $\rho(x)$ will coincide with $\frac{1}{2}x^2$ on almost the whole interval and the estimator will approach $\bar{x}$. In the opposite case, when k is getting smaller, one gets values closer to $\tilde{\theta}_n$. As a compromise values, in practice, values of $k = 1.5$ or $k = 2$ are suggested. The estimators we get through the minimisation:

$$\min_a \sum_{i=1}^n \rho(x_i - a)$$

have the common name **M-estimators**. When ρ is differentiable (such is the case with the Huber function) they can also be considered as a solution of the equation



$$\sum_{i=1}^n \psi(x_i - a) = 0. \qquad (1)$$

To find such estimators, one needs to apply iterative procedures. Under certain regularity conditions, these estimators are asymptotically normal with asymptotic variance equal to

$$\sigma^2(F, \psi) = \frac{\int \psi^2(x) f(x) dx}{(\int \psi'(x) f(x) dx)^2}.$$

Obviously, when $\rho(x) = -lnf(x)$ (in the ideal case when f was known), one gets the MLE estimator for the location parameter of the family $f(x, \theta)$.

4. Optimality in robustness theory for the location model.

After we suggested so many different robust estimators, which one should we take as our "favourite" to use? Obviously, if no information whatsoever is available about how much "contaminated" our ideal parameter family is, it is difficult (and hopeless) to suggest the right choice. But if information about the amount of "contamination" of the

ideal parametric family is available, we can decide, for example, which would be the most suitable M-estimator to use.

Below, we discuss how we could proceed. Instead of assuming that F is exactly in the parametric family, we assume now that F is in a ϵ -neighbourhood of certain distribution G from the ideal parameter family, i.e. $F(x) = (1-\epsilon)G(x) + \epsilon H(x) = F_H(x)$ where $0 < \epsilon < 1/2$ and G are given. We can also say that F is a *mixture* of the ideal model with some other (unknown) distribution H that might be outside the parametric family.

The interpretation of ϵ is that it reflects the "amount of contamination" of the "ideal" G (if $\epsilon = 0$, there is no contamination). We are interested in finding our favourite M-estimator (defined via its $\psi-$ function in (1)) that delivers

$$\inf_{\psi} \sup_{F_H} \sigma^2(F_H, \psi). \tag{2}$$

Here $\sup_{F_H}$ is taken over all symmetric continuous distributions H. The optimality of the estimator is interpreted in a minimax sense (compare (2): we are looking for the estimator that performs best under the worst possible conditions). It can be shown that if G is the standard normal , the minimax M-estimator is a special ρ- estimator of Huber where the constant k can be determined as a function of the contamination ϵ and is the root of the equation

$$\frac{1}{1-\epsilon} = \int_{-k}^{k} \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2})dx + \frac{2}{k}\frac{1}{\sqrt{2\pi}} exp(-k^2/2) \tag{3}$$

Hence, assuming that information about the amount of contamination ϵ is available, we can obtain the **optimal** M-estimator with its k-value defined in Equation (3). (Some computer packages end up with default values in case the customer does not have a good idea about ϵ.)

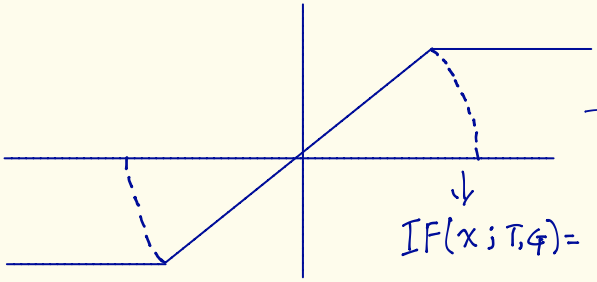
We mention that a more modern and general approach to optimality in robustness has been suggested recently. It is based on investigating the properties of the so-called **influence function**. This function can be defined for any estimator and is used to express its robustness properties. In a nutshell, this approach amounts to investigate the robustness behaviour of the M-estimators when varying the choice of the $\psi-$ function in (1) and to choosing a $\psi-$ function as to meet some (formulated in advance) robustness properties of the estimator. In the Wiley monograph of Hampel, Ronchetti, Rousseeuw and Stahel: "Robust Statistics" this approach has been thoroughly discussed.

5. Robustness isssues in linear regression.

Finally, we should mention that robust estimation is a very important and practically relevant topic in linear regression theory. We only discussed robustness issues in relation to location parameter estimation but it can be seen that a regression model can also be interpreted as location model. Let, for example,

$$Y = X\beta + e, Y = (Y_1, Y_2, \ldots, Y_n) \in R^n, e = (e_1, e_2, \ldots, e_n) \in R^n, \beta \in R^p,$$

X a $(n \times p)$ matrix describes a linear regression model. The errors $e_i, i = 1, 2, \ldots, n$ are independent identically distributed zero-mean random variables. We denote the $i-$th



$$T(F) = \int x \, dF$$

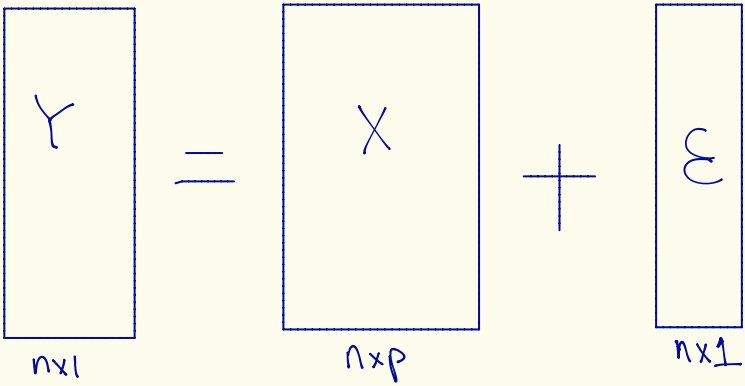
$$T(\hat{F}) = \int x \, d\hat{F}$$

$$IF(x; T, G) = \lim_{t \to 0+} \frac{T((1-t)G + t\Delta x) - T(G))}{t}$$

M - estimators $IF(2) = \Psi(x)$ (bounded)

Breakdown point approach

very robust



$$Y_i = X_i \beta + \varepsilon_i$$

$$y_i = \theta_i + \varepsilon$$

$$Y_i \sim N(X_i \beta, \sigma^2)$$

input row vector $x_i = (x_{ij}), j = 1, 2, \ldots, p$ of the design matrix X. Then if $\theta_i = x_i\beta$ and the error distribution has a density $f(e_i)$ then we can write

$$f_\beta(y_i) = f(y_i - \theta_i) \tag{4}$$

where $\theta_i = x_i\beta$. Equation (4) indicates that treatment of robust estimation of the regression parameter vector β can be approached similarly to the treatment of the simple location model. The need for robust approach in regression arises when there is a belief or evidence that the error density is not exactly normal but a contaminated normal. In that case, robustness in regression can be discussed similarly to robustness in the location model. We shall skip the theoretical details completely but some basic capabilities of S-Plus to perform robust linear regression will be explained and demonstrated during the laboratory exercise.