9 Lecture 9: Robustness. Estimating statistical functionals.

9.1 Motivation. Basic idea of robustness

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. Going over in such cases directly to purely Nonparametric approach would also not address properly the situation since the idea about (relatively small) deviation form a baseline parametric model would be lost. The proper approach would be the robustness approach where we still keep the idea about the ideal parametric model but allow for deviations from it. Speaking loosely, nonparametric statistics allows "all" possible probability distributions and reduces the ignorance about them only by one or a few dimensions. 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, provides 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.

9.1.1 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 $F(x, \theta) = F(x - \theta), \theta \in R^1$. Denote the density by $f(x, \theta) = f(x - \theta)$. If F is a normal distribution 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 $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$. 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 (see lecture) that $\bar{\mathbf{x}}$ has **the same distribution** as the distribution of a single observation from the Cauchy model! Therefore $\bar{\mathbf{x}}$ is even **not** consistent for θ in the Cauchy model! 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) realizations to occur. Because of this observation, we would decide to ignore the observations with a large absolute value and use the **empirical median** instead when estimating the location parameter of the Cauchy distribution. The empirical median $\tilde{\theta}_n$ is **not sensitive** to large realizations 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 to be demonstrated later (in the Section about influence functions in this lecture) we know that for a symmetric F (i.e. $F(0) = 1/2$) with a density $f(0) > 0$,

$$\sqrt{n}(\tilde{\theta}_n - \theta) \to^d N(0, \frac{1}{4f^2(0)}). \tag{29}$$

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 $\tilde{\theta}_n$ 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.

Often, in practice 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? A compromise between the mean and the median is the α**- trimmed mean** $\bar{x}_\alpha = \frac{1}{n-2k}[x_{(k+1)} + x_{(k+2)} + \ldots x_{(n-k)}]$ 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.

9.1.2 Extension of the discussion. The notion of an M-estimator.

Another compromise is suggested by the following observation. It is well known that $\bar{x}$ minimizes the sum $\sum_{i=1}^n (x_i - a)^2$ for all possible values of a whereas $\tilde{\theta}_n$ minimizes the sum $\sum_{i=1}^n |x_i - a|$. When the information is incomplete, it would be a good idea to choose to minimize 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, $k = 1.5$ and $k = 2$ are suggested. The estimators one gets through the minimization:

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

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

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

To find such estimators, one needs to apply iterative procedures. Under certain regularity conditions, they are asymptotically normal with asymptotic variance equal to $\sigma^2(F, \psi) = \frac{\int \psi^2(x) f(x) dx}{(\int \psi'(x) f(x) dx)^2}$. The derivation of this result is in fact not difficult and follows the same steps as the proof of asymptotic normality of the maximum likelihood estimators. You

can find the details in Casella and Berger, p. 486. 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 $f(x, \theta)$.

The most common formulation of the robustness approach is the following. Instead of assuming that F is known precisely, we assume that F is in a ϵ -neighbourhood of certain distribution G, i.e. $F(x) = (1-\epsilon)G(x) + \epsilon H(x) = F_H(x)$ where $\epsilon < 1/2$ and G are given. They reflect the "amount of contamination" (ϵ) of the "ideal" G. We are interested in finding the M-estimator that makes $\inf \sup_{F_H} \sigma^2(F_H, \psi)$ (i.e., minimax approach). Here $\sup_{F_H}$ is taken over all symmetric continuous distributions H. Then 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

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

9.2 Robustness approach based on influence functions

Huber's approach to optimality in robustness as described above is very attractive from theoretical point of view. However it does have the drawback that it requires in the case of the contaminated normal the contamination itself to be symmetric. This is unrealistic. An alternative approach to robustness based on Influence functions was introduced by Hampel and nowadays it dominates the robustness theory. It does not need the assumption about symmetric contamination and has broader applicability, notably in regression.

Roughly speaking, the influence function (IF) describes the effect of an additional observation in any point x on a statistic $T(X)$, given a large sample with distribution F. To give a more precise idea, we consider paremeters of interest as functionals.

Note: Indeed, virtually every parameter of interest of a given distribution can be presented as a functional of the distribution: for example, the mean μ is $\mu = \int xdG(x)$, the variance σ^2 is $\sigma^2 = \int (x - \int xdG(x))^2 dG(x)$, etc. Then it becomes natural to define estimators of parameters as resulting from applying the coresponding defining functional on the empirical distribution function **(edf)** (i.e. $T_n(G_n) = T(G_n)$ where G_n is the empirical distribution function of the sample). Since it is known that the edf is a very good estimator of the true distribution it is then not unreasonable to expect that $T_n(G_n) = T(G_n)$ would be a good estimator of the parameter.

We say that T(G) is the **asymptotic value** of $T_n, n \geq 1$ at G. We shall also assume that the functional under study are **Fisher consistent**, i.e. $T(F_\theta) = \theta$ (which means that on the ideal family they estimate the right parameter asymptotically). Under regularity conditions, one can assume that the limit:

$$lim_{t\to 0}\frac{T((1-t)F + tG) - T(F)}{t} = \int a(x)dG(x) \tag{30}$$

holds which equivalently can be written as

$$\frac{\partial}{\partial t}[T((1-t)F + tG)]_{t=0} = \int a(x)dG(x) = L_F(G)$$

and we call it the *Gâteaux derivative* of the functional T at F in direction G. By putting G=F in the latter equality, we also see from (30) that $\int a(x)dF(x) = 0$ holds. If we put for G the empirical measure that puts the whole mass 1 at the point x we get the *influence function* of the functional T :

$$IF(x;T,F) = \lim_{t\to 0} \frac{T((1-t)F + t\Delta_x) - T(F)}{t} = a(x). \tag{31}$$

Heuristically, the IF describes the effect of infinitesimal contamination at the point x on the estimate, standardized by the mass of contamination. It measures the asymptotic bias caused by contamination.

(Note: The IF of a functional can naturally be estimated via the empirical influence function

$$IF(x;T,\hat{F}_n)$$

where $\hat{F}_n$ is the edf of the data.)

From the definition (30) we see that, since for large n, F_n and F are close, we can write then approximately

$$T(F_n) \approx T(F) + \int IF(x;T,F)dF_n(x) + remainder \approx T(F) + \frac{1}{n}\sum_{i=1}^{n} a(X_i) + remainder$$

If the remainder behaves well under further regularity conditions (the so-called Hadamard differentiability of the functional), from here we also get the asymptotic variance of the estimator. The role of the remainder would be negligible in the calculation of the asymptotic variance and we would end up the statement that $T_n = T(F_n)$ has the property that

$$\sqrt{n}[T_n - T(F)]$$

tends in distribution to $N(0, V(T,F))$ with $V(T,F) = \int IF(x;T,F)^2 dF(x)$. Of course, for large n, we can estimate $\hat{V}(T,F) = V(T,F_n) = \frac{1}{n}\sum_{i=1}^{n} a(X_i)^2$ and we have the approximate result

$$\sqrt{n}(T(F_n) - T(F)) \approx N(0, \hat{V}(T,F)). \tag{32}$$

Then a confidence interval for $T(F)$ can easily be constructed.

Note. The statement (32) is often referred to as the *nonparametric delta method.* The reason behing this name is that indeed in (32) we are "transferring" a result related to the asymptotic distribution of $\sqrt{n}(F_n(t) - F(t)), t \in (-\infty, \infty)$ as a stochastic process to a distribution of a functional T applied to its paths. It is a known fact from probability theory that the stochastic process $G_n(t) = \sqrt{n}(F_n(t) - F(t))$ converges to a zero mean Gaussian process with a covariance function $E[G(s)G(t)] = F(\min(s,t)) - F(s)F(t)$.

9.2.1 Simple Examples

1) Let $T(F) = \int xdF(x) = \mu(F)$ (the mean). An estimator of this functional is $T(F_n) = \int xdF_n(x) = \bar{X}$. We also see that

$$T((1-t)F + t\Delta_x) - T(F) = (1-t)T(F) + tx - T(F) = t(x - T(F))$$

and from here we get $a(x) = x - T(F) = x - \mu$. If we want to estimate it we get $\hat{a}(x) = x - \bar{X}$ and then $\hat{V}(T, F) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. An asymptotic confidence interval for $T(F)$ at level $(1-\alpha)$ would be, not surprisingly:

$$\bar{X} \pm z_{\alpha/2}\frac{1}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

2) Second example. Let $F(x)$ be strictly increasing with a positive density $f(x)$. The $p-th$ quantile ($p \in (0,1)$) of F is defined as the point $q_p : F(q_p) = p$ and clearly it is very important object in Statistical Inference. Symbollically, you can also denote $q_p(F) = T(F) = F^{-1}(p)$. The $p-th$ quantile is used to evaluate threshold constants in designing tests and it is also important in its own right. It became hugely prominet in financial applications (the so-called **VaR** (value at risk) is just a quantile). Estimating this quantile is an important problem in Statistical inference. One obvious candidate is just the empirical quantile $\hat{q}_p$ defined as a solution to the equation $\tilde{F}_n(\hat{q}_p) = p$ where $\tilde{F}_n$ is (linearised version of) the empirical distribution function. We want to derive the influence function of $q_p(F) = T(F) = F^{-1}(p)$. Defining the function

$$H(y, x) = H(y - x) = \left\{ \begin{array}{l} 1 \text{ if } y \geq x \\ 0 \text{ otherwise} \end{array} \right\}$$

we can represent a perturbation of $F(y)$ via point-mass in x as follows:

$$F_t(y) = (1-t)F(y) + tH(y - x), t \in [0, 1].$$

Now, per definition $p = F_t(q_p(F_t))$ holds and taking derivatives from both sides, and applying the chain rule, we get:

$$\frac{d}{dt}p_{|t=0} = 0 = \frac{d}{dt}F_t(q_p(F_t))_{|t=0} = f(q_p)\frac{d}{dt}q_p(F_t)_{|t=0} - p + H(q_p(F) - x)$$

This implies

$$\frac{d}{dt}q_p(F_t)_{|t=0} = IF(x; T, F) = \left\{ \begin{array}{l} \frac{p-1}{f(q_p)} \text{ if } x \leq q_p(F) \\ \frac{p}{f(q_p)} \text{ if } x > q_p(F) \end{array} \right\}$$

From here, we also get the asymptotic variance of $\sqrt{n}(T_n - T(F))$ for the empirical $p-th$ quantile as $V(T, F) = p\frac{(p-1)^2}{f(q_p(F))^2} + (1-p)\frac{p^2}{f(q_p(F))^2} = \frac{p(1-p)}{f(q_p(F))^2}$

Note Applying this result for the case of the median $p = 1/2$ and for F being the standard normal distribution, we discover (29). Note also that if we wanted to use the above result to approximate the asymptotic variance of the sample quantile estimator, we sould need to involve a non-parametric density estimator $\hat{f}$. More stable and simpler estimate of this asymptotic variance can be obtained using the bootstrap method.

9.3 Using the influence function in practice of robust inference.

A sample analogue of the influence function of an estimator $\hat{\theta}_n$ is its so-called **sensitivity curve (SC).** It is defined as

$$SC_{n-1}(x) = n[\hat{\theta}_n(x_1, x_2, \ldots, x_{n-1}, x) - \hat{\theta}_{n-1}(x_1, x_2, \ldots, x_{n-1})].$$

The SC is a function of x and it measures how much an estimator can change when an observation with value x is added to a data set consisting of $(n-1)$ observations $x_1, x_2, \ldots, x_{n-1}$. Robust estimators are the ones for which the sensitivity curve is bounded.

Note: An advantage of the SC in comparison to the IF is that the former can be calculated from the sample whereas the latter is just a theoretical concept. However the SC is sample dependent therefore it is not uniquely defined. One can in fact consider the SC as a (non-parametric) estimator of the IF.

9.3.1 Example

Find the SC for the sample mean and for the sample median. In particular show that the former is unbounded whereas the latter is bounded in x.

i) For the mean $\bar{X} = \hat{\theta}_n$:

$$SC_{(n-1)}(x) = n[\frac{\sum_{i=1}^{n-1} x_i + x}{n} - \frac{\sum_{i=1}^{n-1} x_i}{n-1}] = \frac{(n-1)x - \sum_{i=1}^{n-1} x_i}{n-1} = x - \bar{x}_{n-1}.$$

If we let $n \to \infty$ we get the influence function $x - \mu$ from Section 9.2.1 thus confirming again that the sensitivity curve is just a finite sample analogue of the influence function.

ii) For the median, we first consider the case of an odd sample size $n = 2k+1$. Then $\hat{\theta}_{n-1} = \frac{1}{2}[x_{(k)} + x_{(k+1)}]$ whereas $\hat{\theta}_n(x_1, x_2, \ldots, x_{n-1}, x) = \left\{ \begin{array}{l} x_{(k)} \text{ if } x < x_{(k)} \\ x \text{ if } x_{(k)} \leq x \leq x_{(k+1)} \\ x_{(k+1)} \text{ if } x > x_{(k+1)} \end{array} \right\}$ Hence for $n[\hat{\theta}_n - \hat{\theta}_{(n-1)}]$ we get:

$$SC_{(n-1)}(x) = \left\{ \begin{array}{l} -\frac{n}{2}[x_{(k+1)} - x_{(k)}] \text{ if } x < x_{(k)} \\ \frac{n}{2}[2x - x_{(k)} - x_{(k+1)}] \text{ if } x_{(k)} \leq x \leq x_{(k+1)} \\ \frac{n}{2}[x_{(k+1)} - x_{(k)}] \text{ if } x > x_{(k+1)} \end{array} \right\}$$

This implies that no matter what the value of x, we end up with

$$|SC_{(n-1)}(x)| \leq \frac{n}{2}[x_{(k+1)} - x_{(k)}].$$

The case of even sample $n = 2k$ is similar and is left for you to do it.

As we have seen, the IF can be used in assessing the asymptotic variance of an estimator. Having derived the influence function of an estimator, one can evaluate its robustness properties. The rough guide is that for an estimator to be robust, its IF must be bounded. However, finer robustness properties can also be investigated by studying the IF. For example:

- The **gross error sensitivity** $\gamma = \sup_x |IF(x : T, F)|$ measures the worst influence which a small amount of contamination can have on the value of the estimator. For robustness purposes, it has to be finite.
- Rejection of outliers in terms of IF means that IF should vanish outside certain area.

- **Breakdown point** is the maximal amount of model misspecification before an estimator breaks down (meaning that its bias becomes infinite). Formally the breakdown point is the value $\epsilon^* = \inf_{\epsilon \in (0,1/2]} |bias(\hat{\theta}, F_\theta, \epsilon) = \infty$. The breakdown point of the sample mean is 0 but for the median is the maximal possible ($= 1/2$.) WHY!