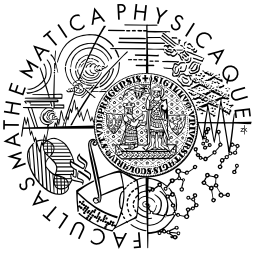

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

Robust Statistical Methods

NMST 444

Course notes

Stanislav Nagy

Last updated: April 25, 2024

Contents

The main references are [12] and [10]. Parts of what is presented here can also be found in [13]. An additional useful and accessible reference to the problem of robust estimation is [20, Sections 3 and 4].

1 Basic concepts of robustness

We work in a probability space $(\Omega, \mathcal{A}, \mathsf{P})$ where all random elements are defined. The set of all Borel probability on a topological space $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$, and $X \sim P \in \mathcal{P}(\mathcal{X})$ means that we have a random variable X with distribution P in $\mathcal{X}$. Typical examples of $\mathcal{X}$ are the real line $\mathbb{R}$ or the Euclidean spaces $\mathbb{R}^k$ with $k \geq 1$. We will often work with weak convergence of measures $\{P_n\}_{n=1}^{\infty} \subset \mathcal{P}(\mathcal{X})$ towards $P \in \mathcal{P}(\mathcal{X})$. We denote it by $P_n \xrightarrow[n\to\infty]{\mathrm{w}} P$. Weak convergence of measures is equivalent to the convergence of the corresponding random variables $X_n \sim P_n$ in distribution to $X \sim P$, we write $X_n \xrightarrow[n\to\infty]{\mathrm{d}} X$. Of course, we know that convergence in probability $X_n \xrightarrow[n\to\infty]{\mathsf{P}} X$ implies $X_n \xrightarrow[n\to\infty]{\mathrm{d}} X$.

1.1 Main ideas and motivation

In standard, parametric statistics we assume that we are given a statistical model $\{P_\theta \colon \theta \in \Theta\} \subset \mathcal{P}(\mathcal{X})$, where θ is our parameter of interest (possibly including nuisance parameters) that lives in the parameter space Θ. In most situations, $\Theta \subseteq \mathbb{R}^p$ with $p \geq 1$. We intend to estimate, or test about, the unknown parameter θ. To do that, we are often given a random sample $X_1, \ldots, X_n$ of independent variables, $X_i \sim P_{\theta_X}$, where $\theta_X \in \Theta$ is the true value of the parameter θ. We construct estimators $T_n = T_n(X_1, \ldots, X_n)$, find their (asymptotic) distributions when assuming $X_i \sim P_\theta$, and infer based on the assumption that each X_i had precisely distribution P_θ.

We already know many optimal estimators and testing procedures. For example, the Rao-Cramér bounds [19, Theorem 3 and 8] and the Lehmann-Scheffé theorems [19, Theorems 17 and 18] give several criteria for the optimality of unbiased estimators of θ. The Neyman-Pearson theorem [19, Theorem 26] establishes the optimal testing procedure for $H_0 \colon \theta = \theta_0$ against $H_1 \colon \theta = \theta_1$, and the theory of maximum likelihood estimation [19, Section 2] gives methods for asymptotically near-optimal inference.

All these tools are quite valuable, but they inherently rely on the assumption that the data $X_1, \ldots, X_n$ truly come from the ideal distribution P_θ. For all practical purposes, that is only a mathematical model. It is certainly necessary to assume something about the data-generating process to be able to say anything reasonable about θ. But, usually, it is also too idealistic to suppose that the model is exactly true in practice. Take, for example, the exact t-test.

There, we suppose that the data come from a normal distribution. In practice, we can never be sure about this:

- What if the data really comes from a Student's distribution?
- What if some rounding takes place that invalidates the assumption of normality?
- Can the true distribution fail to possess a second moment?
- What if some of the observations contain hidden measurement errors, and thus we deal with a mixture of two random samples instead of just the normal one?

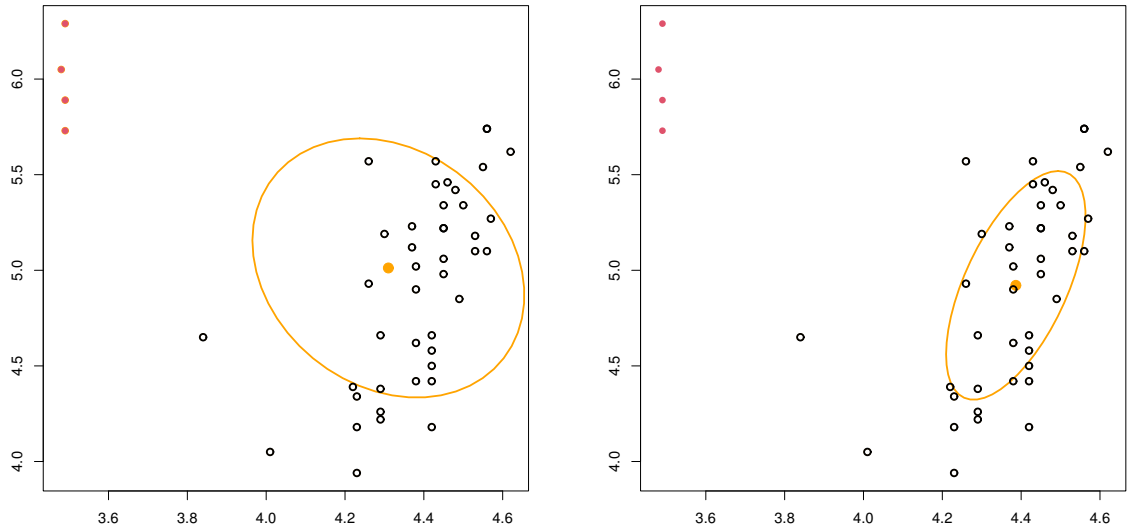

Figure 1: The bivariate stars dataset contains four suspicious observations (red points in the top left corner). With the data, we see the sample mean (orange point) and a Mahalanobis ellipsoid based on the sample covariance matrix (orange curve) both for the complete dataset (left panel) and the dataset without the suspicious observations (right panel). The red points severely affect both the center and the shape of the ellipsoid.

Example 1.1. To illustrate the problem with potential gross errors in data or suspicious (so-called outlying) observations, consider the real `starsCYG` dataset from R package robustbase. The bivariate data represent the Hertzsprung-Russell diagram of the star cluster CYG OB1, which contains 47 stars in the direction of the constellation Cygnus. The first variable is the logarithm of the effective temperature at the surface of the star and the second one is the logarithm of its light intensity. In the scatterplot of these data points in Figure 1, two groups of points are seen: the majority which tends to follow a bivariate normal distribution and

four stars in the upper corner (red points in Figure 1). In astronomy, the 43 stars plotted in black are said to lie on the main sequence and the four remaining stars are called "giants".

To visualise the geometry of the data, we use the sample Mahalanobis ellipse given as

$$\left\{x \in \mathbb{R}^2 \colon\ (x-\mu_n)^{\mathsf{T}}\, \Sigma_n^{-1}\, (x-\mu_n) \le c\right\},$$

where μ_n and Σ_n are the sample mean and the sample covariance matrix of the data, respectively, and $c > 0$ is an appropriate constant. This ellipse is plotted in Figure 1 in orange. As can be seen, the shape of the ellipse is affected profoundly by the four giant stars, and for full data neither the shape nor the center of the ellipse represents the geometry of the data well (left panel). When the giant stars are omitted, the same ellipse appears to capture the shape of the data much better (right panel). △

The problem, of course, is how to detect deviating data such as the giants in this dataset? Or even better, is it possible to devise statistical procedures that will be able to cope with the presence of potentially outlying observations or departures from the assumed model without a great loss in efficiency?

The main principle of robust statistics is to address these problems. It aims to study the effect of deviations from the hypothetical model P_θ on statistical procedures, understand and quantify it, and propose methods to deal with the potential problems. We do so not only by studying the behaviour of the procedure T_n at the model P_θ, but also by considering probability measures $P \in \mathcal{P}(\mathcal{X})$ is a small neighbourhood $\mathcal{P}_\varepsilon \subseteq \mathcal{P}(\mathcal{X})$ of P_θ, and examining $T_n(X_1, \ldots, X_n)$ in the case when $X_1, \ldots, X_n \sim P$ for $P \in \mathcal{P}_\varepsilon$. To put the idea simply, we will search for and design procedures that will be not only (close to) optimal at the hypothetical model P_θ, but also stable in the argument of P in neighbourhoods of P_θ. This approach always involves a trade-off; we usually need to sacrifice a portion of optimality to achieve robustness (that is, stability). An ideal output would be a procedure that is nearly optimal at P_θ, and at the same time not easily disturbed by small deviations from the idealistic model.

Example 1.2. As another simple example take the average $\bar{X}_n = \sum_{i=1}^n X_i/n$ of a random sample $X_1, \ldots, X_n$ from a distribution $P_\theta \in \mathcal{P}(\mathbb{R})$, whose expected value $\mathsf{E}\, X_1 = \theta$ is to be estimated. We know that the sample average is the optimal estimator of the mean $\theta \in \mathbb{R}$ in the location model

$$\mathcal{F} = \{X_1, \ldots, X_n \sim \mathsf{N}(\theta, 1) \text{ independent } : \theta \in \mathbb{R}\}.$$

We compare the sample mean with, e.g., the sample median $T_n = \mathsf{med}\,(X_1, \ldots, X_n)$. Both $\bar{X}_n$ and T_n estimate the same parameter θ in $\mathcal{F}$, and they are both asymptotically normal.

We have

$$\begin{aligned} \sqrt{n}\left(\bar{X}_n - \theta\right) &\xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}(0, \mathsf{var}\, X_1), \\ \sqrt{n}\left(T_n - \theta\right) &\xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}\left(0, \frac{1}{4(f_{X_1}(\theta))^2}\right), \end{aligned} \tag{1}$$

where f_{X_1} is the density of X_1. We can compare the performance of the two estimators by means of their asymptotic relative efficiency (ARE). This is defined as the ratio of the asymptotic variances

$$\mathsf{ARE} = \frac{\mathsf{var}\, X_1}{(f_{X_1}(\theta))^{-2}/4} = \frac{4}{2\,\pi} \approx 0.63. \tag{2}$$

We obtain that at the model $\mathcal{F}$, the sample average is almost twice more efficient than the sample median.

Suppose now, however, that the data $X_1, \ldots, X_n$ do not come exactly from the normal distribution in $\mathcal{F}$. Rather, assume that a small, ε-fraction of them comes from a different distribution $Q \in \mathcal{P}(\mathbb{R})$, for $\varepsilon > 0$ close to zero. This can be described by a contamination model, where the true distribution we sample from is a mixture of the ideal distribution P and the contaminating one Q. We take $\varepsilon \in (0, 1)$ small, and assign to the ideal distribution P a weight $(1 - \varepsilon)$, and the contaminating distribution Q weight ε. We obtain a measure

$$P_\varepsilon = (1 - \varepsilon)\, P + \varepsilon\, Q \in \mathcal{P}(\mathcal{X}) \quad \text{for } \varepsilon \in (0, 1), \tag{3}$$

with $\mathcal{X} = \mathbb{R}$. The measure P_ε is defined by

$$P_\varepsilon(B) = (1 - \varepsilon)\, P(B) + \varepsilon\, Q(B) \quad \text{for each } B \subseteq \mathcal{X} \text{ Borel and } \varepsilon \in (0, 1). \tag{4}$$

Consider an experiment where the true value of θ is 0, and where an ε-fraction of the observations does not come from the ideal distribution $P = \mathsf{N}(0, 1)$, but rather from some $Q \in \mathcal{P}(\mathbb{R})$ with density g. Writing $\varphi(x) = (2\,\pi)^{-1/2} \exp(-x^2/2)$ for the density of $\mathsf{N}(0, 1)$, the density of P_ε is

$$f_\varepsilon(x) = (1 - \varepsilon)\, \varphi(x) + \varepsilon\, g(x) \quad \text{for } x \in \mathbb{R}.$$

Observe that even if $Q \neq P$ is normal, P_ε is not normal. Instead, it is a mixture of two normals. That can be interpreted using the following sampling scheme:

- First, we take a Bernoulli random variable $Y \sim \mathsf{Bernoulli}(\varepsilon)$ such that $\mathsf{P}(Y = 1) = \varepsilon$; and then
- we take $X \sim P = \mathsf{N}(0, 1)$ if $Y = 0$, and $X \sim Q$ if $Y = 1$.

Thus, our random sample $X_1, \ldots, X_n$ from P_ε has approximately a fraction ε of observations sampled from Q. Those can be interpreted as data points containing measurement

errors or simply observations not following the assumed model P (such as the giant stars in Example 1.1).

If Q is a distribution without a finite second moment, with any $\varepsilon > 0$ we have

$$\mathsf{E}_{P_\varepsilon} X^2 = (1-\varepsilon)\,\mathsf{E}_P X^2 + \varepsilon\,\mathsf{E}_Q X^2 = \infty,$$

where $\mathsf{E}_Q X^2 = \int_{\mathbb{R}} x^2 \,\mathrm{d}\, Q(x)$ etc. We see that for arbitrarily small contamination, $\mathsf{var}_{P_\varepsilon} X = \infty$ and the central limit theorem for the sample mean $\bar{X}_n$ fails. Even worse, if Q does not have even the first moment, $\mathsf{E}_{P_\varepsilon} X$ is not defined, meaning that also $\mathsf{E}_{P_\varepsilon} \bar{X}_n$ does not exist. In words, arbitrarily small contamination can completely disrupt the behaviour of the sample mean $\bar{X}_n$ as an estimator of θ. The median of P_ε changes much less. Even if the whole ε-mass of Q is added strategically to one side of $\theta = 0$, the median of P_ε stays bounded as

$$\mathsf{med}_{P_\varepsilon} X \in \left[\Phi^{-1}\left(\frac{1}{2(1+\varepsilon)}\right), \Phi^{-1}\left(\frac{1}{2(1-\varepsilon)}\right)\right],$$

where we write $\mathsf{med}_Q X$ for the median of random variable $X \sim Q \in \mathcal{P}(\mathbb{R})$, and Φ is the distribution function of the standard normal distribution.

Let us, however, not consider only extreme contamination and take $Q = \mathsf{N}(0, 3^2)$, which is a fairly tame distribution. We compute the asymptotic relative efficiency (2) of the sample average and the sample mean in the model P_ε, $\varepsilon \in (0,1)$. The central limit theorem now still guarantees (1), only the asymptotic variances change appropriately. For the sample average we get for $X_1 \sim P_\varepsilon$ that

$$\mathsf{var}_{P_\varepsilon} X_1 = \mathsf{E}_{P_\varepsilon} X_1^2 - (\mathsf{E}_{P_\varepsilon} X_1)^2 = 1 + 8\,\varepsilon,$$

since

$$\mathsf{E}_{P_\varepsilon} X_1^s = (1-\varepsilon)\,\mathsf{E}\, Z^s + \varepsilon\,\mathsf{E}\,(3Z)^s = (1+8\,\varepsilon)\mathsf{E}\, Z^s, \quad \text{for } Z \sim \mathsf{N}(0,1) \text{ and } s \in \mathbb{R},$$

and for the sample median

$$\frac{1}{4(f_\varepsilon(0))^2} = \frac{1}{4((1-\varepsilon)\,\varphi(0) + \varepsilon\,\frac{1}{3}\varphi(0))^2} = \frac{9\,\pi}{2(3-2\varepsilon)^2}.$$

As a function of the contamination level $\varepsilon \in (0,1)$, we now obtain

$$\mathsf{ARE}(\varepsilon) = \frac{\mathsf{var}_{P_\varepsilon} X_1}{(4(f_\varepsilon(0))^2)^{-1}} = \frac{2(1+8\,\varepsilon)(3-2\varepsilon)^2}{9\,\pi}.$$

This function is displayed in the left-hand panel of Figure 2.

We already know that the sample average is the best possible estimator in a normal distribution, which corresponds to both $P_0 = \mathsf{N}(0,1)$ and $P_1 = \mathsf{N}(0,3^2)$. This is reflected in

$$\mathsf{ARE}(0) = \mathsf{ARE}(1) = \frac{4}{2\,\pi} \approx 0.63.$$

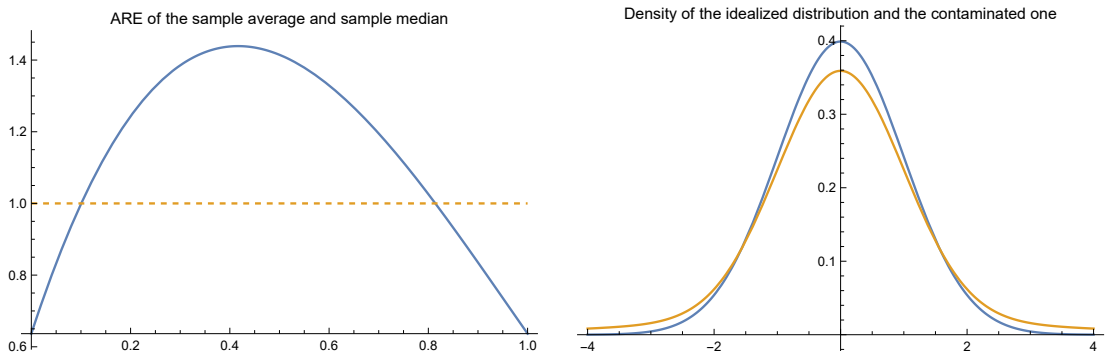

Figure 2: Example 1.2. Left panel: Asymptotic relative efficiency of the sample mean and the sample median in a contamination model. For about $\varepsilon = 0.15$, the sample median is more efficient than the sample mean, with $\mathsf{ARE}(0.15) \approx 1.13$. Right panel: The density of $P_0 = \mathsf{N}(0,1)$ (blue curve) and the contaminated density of $P_{0.15}$ with $Q = \mathsf{N}(0,3^2)$ (orange curve). The contamination is not easy to recognise from the density of P_ε.

As ε departs from 0 and 1, we no longer deal with a normal distribution and starting from about $\varepsilon \approx 0.1$ to $\varepsilon \approx 0.8$, the sample median is more efficient.

It might seem that the level of contamination $\varepsilon = 0.1$ needed for the median to be more efficient (corresponding to about 10% of bad observations) is high. Taking the contaminating measure $Q = P_1 = \mathsf{N}(0,5^2)$ this threshold reduces already to $\varepsilon \approx 0.026$, and as the variance of Q goes to infinity it drops to $\varepsilon \to 0$ too. In the extreme case of Q without a second moment (e.g., the Student distribution with two degrees of freedom), any contamination $\varepsilon > 0$ leads to $\mathsf{ARE}(\varepsilon) = \infty$; a single observation can break down the sample mean completely, for any sample size n. △

An even more extreme example of the same phenomenon occurs when one estimates the variance.

Example 1.3. Take a similar setup as in Example 1.2. We are interested in estimating $\sigma > 0$ in the model $P = \mathsf{N}(\mu, \sigma^2)$, with $\mu \in \mathbb{R}$ unknown. In the early 20th century, there was a dispute about whether a better estimator of σ is the square root of the sample variance

$$S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2},$$

or an estimator coming from the mean absolute deviation

$$D_n = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - \bar{X}_n\right|.$$

The estimator S_n was preferred by Fisher, a variant of D_n by Eddington. We intend to compare the estimators in terms of their asymptotic relative efficiency again. Direct computations are again possible, but already more complicated than in Example 1.2. First, one needs to realise that to compare S_n with D_n, they need to estimate the same quantity consistently. While, in our setup, S_n clearly estimates $\sqrt{\mathsf{var}\, X_1} = \sigma$, the quantity D_n converges in probability to

$$\mathsf{E}\ |X_1 - \mathsf{E}\, X_1| = \sigma\, \mathsf{E}\ |Z| = \sigma\sqrt{\frac{2}{\pi}}, \quad \text{where } Z \sim \mathsf{N}(0,1).$$

The estimators to compare are, therefore, S_n and $\widetilde{D}_n = \sqrt{\pi/2}\, D_n$. In the model given by $P = \mathsf{N}(\mu, \sigma^2)$, it can be computed that $\mathsf{ARE} \approx 0.876$, and the Fisher's estimator S_n is better. However, taking into account possible contamination in model P_ε as in (3) with $Q = \mathsf{N}(\mu, 9\sigma^2)$ we obtain

$$\mathsf{ARE}(\varepsilon) = \frac{\frac{1}{4}\left(\frac{3(1+80\varepsilon)}{(1+8\varepsilon)^2} - 1\right)}{\frac{\pi(1+8\varepsilon)}{2(1+2\varepsilon)^2} - 1}.$$

This function is displayed in Figure 3. It grows quite fast: already at $\varepsilon = 1.75 \cdot 10^{-3}$, it reaches above the threshold value 1. It grows with a maximum value greater than 2 at around $\varepsilon = 0.055$.

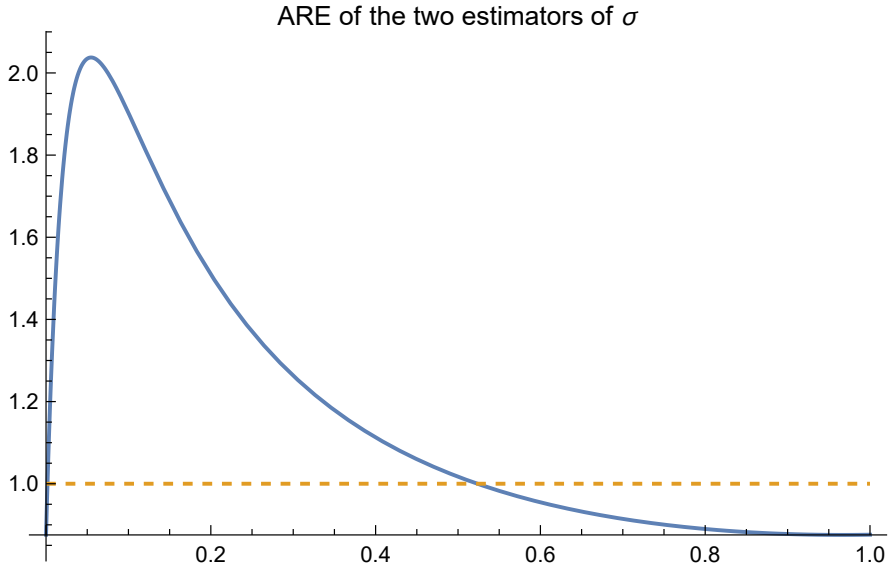

Figure 3: Example 1.3: Asymptotic relative efficiency of the estimators S_n and $\widetilde{D}_n$ of σ in a contamination model.

The phenomenon we see is quite similar to that from Example 1.2, but much more pronounced. Already for the simple contamination by another normal distribution, just 2 in $n = 1000$ observations are enough to disrupt the estimator S_n so that it loses its edge of

12.3% in terms of efficiency. Just 5% of contaminating observations make the Eddington estimator $\widetilde{D}_n$ twice more efficient than S_n. △

Our result in Example 1.3, of course, does not mean that S_n is a bad estimator of σ. Also, we do not claim that $\widetilde{D}_n$ should be preferred; there are certainly better estimators. The example is only intended to illustrate that even very small departures from the ideal distribution may cause the optimal statistical procedures to break down completely. It is the objective of robust statistical methods to understand these issues, and to develop methods to deal with them.

1.2 Statistical functionals

We need to study the continuity of statistical procedures in the argument of the underlying measures. Thus, for an appropriate mathematical formulation, we cannot avoid a certain level of abstraction. The statistical procedure will be represented by a quantity $T_n = T_n(X_1, \ldots, X_n)$. This can be, for example, a point estimator or a test statistic. Further, $X_1, \ldots, X_n$ is typically a random sample from a measure $P \in \mathcal{P}(\mathcal{X})$, and the space $\mathcal{X}$ can be either $\mathbb{R}$ or $\mathbb{R}^k$ with $k > 1$. The random variable $T_n(X_1, \ldots, X_n)$ estimates a quantity that depends on the population distribution P, that is, the true distribution from which $X_1, \ldots, X_n$ was sampled.

To unify the exposition at the sample and the population level, we will not formally work with the random sample $X_1, \ldots, X_n$ itself, but rather represent it by its empirical measure

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} \in \mathcal{P}(\mathcal{X}). \tag{5}$$

Here, $\delta_x \in \mathcal{P}(\mathcal{X})$ is the Dirac measure at the point $x \in \mathcal{X}$, i.e.

$$\delta_x(B) = \mathbb{I}(x \in B) = \begin{cases} 1 & \text{if } x \in B, \\ 0 & \text{if } x \notin B, \end{cases} \quad \text{for all } B \subseteq \mathcal{X} \text{ Borel}, \tag{6}$$

and the algebraic operations in (5) are interpreted as a mixture (4). For the simple case $\mathcal{X} = \mathbb{R}$ we could also represent P_n by its distribution function

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x) \quad \text{for } x \in \mathbb{R},$$

which is the empirical cumulative distribution function of our random sample, or equivalently the distribution function of (5).

Remark 1. In what follows, for $P \in \mathcal{P}(\mathcal{X})$ given, we will need to distinguish between usual sequences of measures $\{P_n\}_{n=1}^{\infty} \subset \mathcal{P}(\mathcal{X})$ (say, any sequences such that $P_n \xrightarrow[n\to\infty]{\text{w}} P$), and

the special sequence of empirical measures of random samples $X_1, \ldots, X_n$ from P, defined as in (5). For that, we will sometimes, for clarity, add to empirical measures P_n an argument $\omega \in \Omega$, to emphasise that $P_n = P_n(\omega) \in \mathcal{P}(\mathcal{X})$ is, in fact, a random measure depending on the random sample (and thus on the random element ω). In case when no confusion can arise about whether we use an ordinary sequence of measures P_n or the empirical measures $P_n(\omega)$, we will drop the argument ω from the latter; in that notation, we understand the empirical measure P_n from (5) as a random measure in $\mathcal{P}(\mathcal{X})$.

In addition, when working with measures P (or P_n) and their corresponding distribution functions F (or F_n), we will frequently freely exchange P with F and P_n with F_n as they both represent the same thing.

Representing the random sample $X_1, \ldots, X_n$ by its empirical measure P_n from (5) is not without loss of generality. In (5), we lose information about e.g. the order of the variables $X_1, \ldots, X_n$ in $\mathbb{R}$. Thus, this representation is not appropriate in problems where $X_1, \ldots, X_n$ come from a time series or when X_i are not identically distributed. Under our assumption of $X_1, \ldots, X_n$ a random sample, however, the distribution of the random vector $(X_1, \ldots, X_n)^\mathsf{T}$ is the same for any permutation of its elements X_i, $i = 1, \ldots, n$. Thus, the random empirical measure P_n (or F_n for $k = 1$) is a sufficient statistic in our model, and the Rao-Blackwell theorem [19, Theorem 16] gives that we do not lose any Fisher information by considering only inference based on P_n (or F_n).

Our statistical procedure will be formally represented as a statistical functional, in the sense of the following definition.

Definition 1. A mapping $T\colon \mathcal{P} \to \mathbb{R}^p$ with $\mathcal{P} \subseteq \mathcal{P}(\mathcal{X})$ is called a *statistical functional* whose domain is $\mathcal{P}$.

In words, a statistical functional is any map from the space of measures, typically into the parameter space $\Theta \subseteq \mathbb{R}^p$. For now, we do not require any measurability of T since, for that, we would need to define a topology on $\mathcal{P}(\mathcal{X})$. That will be discussed in Section 2. Simple examples of statistical functionals are (for $\mathcal{X} = \mathbb{R}$)

- the mean $T(P) = \int_{\mathbb{R}} x \,\mathrm{d}\, P(x) = \mathsf{E}\, X$ with $X \sim P$, defined for $P \in \mathcal{P}_1(\mathbb{R})$. Here

$$\mathcal{P}_s(\mathcal{X}) = \left\{ P \in \mathcal{P}(\mathcal{X}) : \int_{\mathcal{X}} \|x\|^s \,\mathrm{d}\, P(x) < \infty \right\} \quad \text{for } s \in \mathbb{R}, \tag{7}$$

  where we assume that $\mathcal{X}$ is a normed space with norm $\|\cdot\|$;

- the median $T(P) = \inf\{x \in \mathbb{R}\colon F(x) \geq 1/2\} = \mathsf{med}\,(X)$ defined for all $P \in \mathcal{P}(\mathbb{R})$, where F is the distribution function of P; or

- the variance $T(P) = \int_{\mathbb{R}} x^2 \,\mathrm{d}\, P(x) - \left(\int_{\mathbb{R}} x \,\mathrm{d}\, P(x)\right)^2 = \mathsf{E}\, X^2 - (\mathsf{E}\, X)^2$ defined with $X \sim P$ and for $P \in \mathcal{P}_2(\mathbb{R})$ from (7).

- the Neyman-Pearson test statistic [19, Theorem 26], given for two densities $p_1, p_2 \colon \mathcal{X} \to [0, \infty)$ by

$$T(P) = \mathsf{E}_P \, \log\left(p_1(X)/p_0(X)\right).$$

- the maximum likelihood estimator T of $\theta \in \mathbb{R}^p$, defined as

$$T(P) = \arg\max_{\theta \in \Theta} \mathsf{E}_{\theta_X} \, \log f(X, \theta), \tag{8}$$

  with $X \sim P_{\theta_X}$ and the true parameter value $\theta_X \in \Theta$. The maximum likelihood estimator can be defined also implicitly as the solution to the equation

$$\int_{\mathcal{X}} \psi(x, \theta) \,\mathrm{d}\, P(x) = 0,$$

  in θ, where

$$\psi(x, \theta) = \frac{\partial}{\partial \theta} \log f(x, \theta),$$

  for $\{f(x, \theta) \colon \theta \in \Theta\}$ a system of densities of P_θ, $\theta \in \Theta$.

Observe that each of these functionals is well defined for empirical measures given by random samples $X_1, \ldots, X_n$. For the mean functional, we get

$$T(P_n) = \int_{\mathbb{R}} x \,\mathrm{d}\, P_n(x) = \int_{\mathbb{R}} x \,\mathrm{d} \left(\frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}\right)(x) = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}_n,$$

and its empirical version is the sample mean. For the median, we have

$$T(P_n) = \inf\{x \in \mathbb{R} \colon F_n(x) \geq 1/2\} = \mathsf{med}\,(X_1, \ldots, X_n),$$

the sample median. Finally, for the variance

$$T(P_n) = \int_{\mathbb{R}} x^2 \,\mathrm{d}\, P_n(x) - \left(\int_{\mathbb{R}} x \,\mathrm{d}\, P_n(x)\right)^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \left(\bar{X}_n\right)^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

Remark 2. Throughout this text, we will always assume that statistical functionals T are defined at least for all empirical measures $P_n(\omega)$, and for the true measure $P \in \mathcal{P}(\mathcal{X})$ from which we sample. This means we always assume that $T(P_n)$ is a well-defined estimator of $T(P)$ based on the random sample $X_1, \ldots, X_n$ from P. In what follows, we often abuse the notation and say simply that a statistical functional is a map $T \colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}^p$; by this, we mean that the domain of T is an appropriate subset of all measures $\mathcal{P}(\mathcal{X})$ that makes $T(P)$ well defined.

Suppose now that we have a parametric model $\{P_\theta \colon \theta \in \Theta\} \subset \mathcal{P}(\mathcal{X})$. For a random sample $X_1, \ldots, X_n$ from some P_θ, we represent an estimator $T_n = T_n(X_1, \ldots, X_n)$ of θ as a statistical functional

$$T_n(X_1, \ldots, X_n) = T(P_n), \tag{9}$$

for some choice of T. Of course, there is not a unique choice of T; many different functionals can represent the same estimators. The first requirement on T is that, if an empirical measure $P_n(\omega)$ is replaced by the true distribution P_θ in (9) from which we sampled, we have that $T(P) = \theta$. That means that we indeed estimate the quantity of interest.

Definition 2. In a parametric model $\{P_\theta \colon \theta \in \Theta\} \subset \mathcal{P}(\mathcal{X})$, a statistical functional T is said to be *Fisher consistent* for θ if $T(P_\theta) = \theta$ for all $\theta \in \Theta$.

The maximum likelihood estimator from (8) is Fisher consistent if the statistical model is correctly specified (that is, if the true distribution P_{θ_X} from which we sample is an element of the statistical model).

Example 1.4. Let $\theta_X \in \Theta$ be the true value of the parameter from which the random sample $X_1, \ldots, X_n$ is generated. Suppose that the support of $f(x, \theta)$ is the same for each $\theta \in \Theta$, as is always assumed for maximum likelihood estimation. The true value θ_X is fixed. Thus, also $\mathsf{E}_{\theta_X} \log(f(X, \theta_X))$ is given and fixed, and we can write the maximum likelihood estimator as

$$\begin{aligned}\arg\max_{\theta\in\Theta} \left(\mathsf{E}_{\theta_X} \log(f(X,\theta))\right) &= \arg\max_{\theta\in\Theta} \left(\mathsf{E}_{\theta_X} \log(f(X,\theta)) - \mathsf{E}_{\theta_X} \log(f(X,\theta_X))\right) \\ &= \arg\max_{\theta\in\Theta} \left(\mathsf{E}_{\theta_X} \log\left(\frac{f(X,\theta)}{f(X,\theta_X)}\right)\right).\end{aligned}$$

Then, for $\mathcal{X}$ the sample space of X and μ the σ-finite measure defining the densities $f(\cdot, \theta)$, Jensen's inequality gives

$$\mathsf{E}_{\theta_X} \log\left(\frac{f(X,\theta)}{f(X,\theta_X)}\right) \le \log\left(\mathsf{E}_{\theta_X} \frac{f(X,\theta)}{f(X,\theta_X)}\right) = \log\left(\int_{\mathcal{X}} \frac{f(x,\theta)}{f(x,\theta_X)} f(x,\theta_X)\,\mathrm{d}\,\mu(x)\right) = \log(1) = 0,$$

with equality if $\theta = \theta_X$. In particular, the maximum likelihood estimator is Fisher consistent, at least if the true distribution P of X indeed lies in the assumed parametric model. △

If P does not correspond to any parameter $\theta \in \Theta$, there is an interesting connection of maximum likelihood with the Kullback-Leibler divergence. For details see [21, Example 39].

The Fisher consistency is, strictly speaking, a property of the functional T, not a property of the estimator T_n.

Example 1.5. The mean functional $T(P) = \int_{\mathbb{R}} x\,\mathrm{d}\,P(x)$ is clearly Fisher consistent for $\theta = \mathsf{E}\,X$, $X \sim P \in \mathcal{P}_1(\mathbb{R})$. The sample mean $T_n(X_1, \ldots, X_n)$ can, however, be represented

also by the functional

$$\widetilde{T}(P) = \begin{cases} \bar{X}_n & \text{if the measure } P \text{ can be written as (5)}, \\ 0 & \text{elsewhere.} \end{cases}$$

More precisely, in the first case, we suppose that there exists $n = 1, 2, \ldots$ and points $x_1, \ldots, x_n \in \mathbb{R}$ such that P can be written using (5) with $x_i = X_i$, and this n is the smallest possible. The functional $\widetilde{T}$ is defined for all $P \in \mathcal{P}(\mathbb{R})$. For any P_n empirical it takes the value $\widetilde{T}(P_n) = \bar{X}_n = T(P_n)$. Nevertheless, it is not Fisher consistent for $\theta = \mathsf{E}\, X$ with $X \sim P \in \mathcal{P}_1(\mathbb{R})$. △

1.3 Qualitative robustness

In our analysis, estimators $T_n(X_1, \ldots, X_n)$ are replaced by some Fisher consistent statistical functionals T. In that representation, robustness of our procedure at $P \in \mathcal{P}(\mathcal{X})$ can be interpreted as the stability, or continuity, of the functional T in some neighbourhood of P in the space of probability measures $\mathcal{P}(\mathcal{X})$.

To formulate what is meant by "continuity" of T, we need a topology in $\mathcal{P}(\mathcal{X})$. A natural choice is the weak topology (also called *weak-star* topology in functional analysis [26, Section 3.1.11]). Recall that a sequence of measures $\{P_n\}_{n=1}^{\infty} \subset \mathcal{P}(\mathcal{X})$ converges weakly to $P \in \mathcal{P}(\mathcal{X})$ if for all bounded continuous functions $f\colon \mathcal{X} \to \mathbb{R}$ we have

$$\int_{\mathcal{X}} f(x)\, \mathrm{d}\, P_n(x) \xrightarrow[n\to\infty]{} \int_{\mathcal{X}} f(x)\, \mathrm{d}\, P(x).$$

If this is true, we also write $P_n \xrightarrow[n\to\infty]{\mathrm{w}} P$. This convergence is the same as the convergence in distribution of $X_n \sim P_n$ towards $X \sim P$, that is

$$X_n \xrightarrow[n\to\infty]{\mathrm{d}} X \quad \text{if and only if} \quad P_n \xrightarrow[n\to\infty]{\mathrm{w}} P.$$

One natural definition of robustness of a statistical functional T at a measure $P \in \mathcal{P}(\mathcal{X})$ is therefore the requirement of (weak) continuity of T at P

$$P_n \xrightarrow[n\to\infty]{\mathrm{w}} P \quad \text{implies} \quad T(P_n) \to T(P) \quad \text{as } n \to \infty. \tag{10}$$

When working with statistical functionals defined on $\mathcal{P}(\mathcal{X})$, we use "continuous" and "weakly continuous" synonymously. The formal definition of qualitative robustness is, however, more involved; we will want to ensure that the convergence in (10) is uniform. To formulate that, we need, in addition, a metric on $\mathcal{P}(\mathcal{X})$. Take for now any metric $d_*\colon \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to [0, \infty)$ that metrizes the weak topology in $\mathcal{P}(\mathcal{X})$ (that, is $d_*(P_n, P) \to 0$ as $n \to \infty$ if and only if $P_n \xrightarrow[n\to\infty]{\mathrm{w}} P$ in $\mathcal{P}(\mathcal{X})$; details will be given in Section 2).

In the following definition, we write $\mathcal{L}_P(f(X))$ for the law (that is, the distribution) of the random variable $f(X)$ if $X \sim P$. By $\mathcal{L}_P(T_n)$ we mean the law of the estimator $T_n(X_1, \ldots, X_n)$ for $X_1, \ldots, X_n$ sampled independently from P. Likewise, in the sequel, when we write $\mathcal{L}_P(f(P_n))$ we mean the law of $f(P_n)$ when P_n is an empirical measure (5) sampled from P.

Definition 3 (Qualitative robustness)**.** Let $T_n = T_n(X_1, \ldots, X_n) \in \mathbb{R}^p$ for $n = 1, 2, \ldots$ be a sequence of estimators or test statistics based on a random sample $X_1, \ldots, X_n$ from some $P \in \mathcal{P}(\mathcal{X})$. The sequence $\{T_n\}_{n=1}^{\infty}$ is called *qualitatively robust* at $P_0 \in \mathcal{P}(\mathcal{X})$ if the sequence of maps $\{\xi_n\}_{n=1}^{\infty}$ defined by

$$\xi_n \colon \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathbb{R}^p) : P \mapsto \mathcal{L}_P(T_n) \tag{11}$$

is asymptotically equicontinuous at P_0. That means, for each $\varepsilon > 0$ there exists $\delta > 0$ and $n_0 \geq 1$ such that for all $P \in \mathcal{P}(\mathcal{X})$ and $n \geq n_0$ we have

$$d_*(P_0, P) \leq \delta \quad \text{implies} \quad d_*(\mathcal{L}_{P_0}(T_n), \mathcal{L}_P(T_n)) = d_*(\xi_n(P_0), \xi_n(P)) \leq \varepsilon. \tag{12}$$

Definition 3 is very general. If each of the individual maps ξ_n is continuous, the definition of qualitative robustness is, in fact, just the equicontinuity of the set of maps $\{\xi_n\}_{n=1}^{\infty}$ from (11), see [22, Definition 13.4.3]. Indeed, in that case, write $(X_1, d_1) = (\mathcal{P}(\mathcal{X}), d_*)$ and $(X_2, d_2) = (\mathcal{P}(\mathbb{R}^p), d_*)$ for the domain and the codomain metric spaces of ξ_n, respectively. The formula (12) can then be rewritten as the requirement that for a point $x \in X_1$ given (here, $x = P_0$), for each $\varepsilon > 0$ there exists $\delta > 0$ such that

$$d_1(x, y) \leq \delta \quad \text{implies} \quad \sup_{n=1,2,\ldots} d_2(\xi_n(x), \xi_n(y)) \leq \varepsilon.$$

This is precisely the equicontinuity of $\{\xi_n\}_{n=1}^{\infty}$ at $x \in X_1$. The additional requirement of $n \geq n_0$ is in (12) only to allow some finite number of ξ_n not to be weakly continuous.

In the common situation when each $T_n(X_1, \ldots, X_n)$ can be represented by a single statistical functional, we will see in Section 2.3 that the definition of qualitative robustness simplifies substantially. Essentially, it is simply the weak continuity of the underlying functional T. To prove that, we however need some additional facts about the metrics d_* that will be shown in Section 2.

The condition in Definition 3 is called qualitative robustness because it asserts continuity qualitatively, that is, without stating explicitly how much robust the estimators T_n are.

The end of lecture 1 (22.02.2024)

1.4 Quantitative robustness

The notion of qualitative robustness is useful as a minimal criterion of the robustness of a functional — any reasonably robust method must be qualitatively robust. To determine the

degree of robustness of a functional T, we need to introduce additional measures of stability of T. Many numerical characteristics of robustness have been developed in the literature. Possibly the simplest are the maximum bias and maximum variance.

Suppose that the estimators or test statistics $T_n = T_n(X_1, \ldots, X_n)$ are obtained from a single statistical functional T, and take values in $\mathbb{R}$. A reasonable procedure should be consistent, meaning

$$T(P_n) \xrightarrow[n\to\infty]{\mathsf{P}} T(P)$$

for each $P \in \mathcal{P}(\mathcal{X})$ on which T is defined, and P_n an empirical measure (5) sampled from P. In addition, T_n is frequently asymptotically normal, meaning that

$$\sqrt{n}\,(T(P_n) - T(P)) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}(0, A(P,T)), \tag{13}$$

with asymptotic variance $A(P,T) > 0$.

To quantify the robustness of T at $P_0 \in \mathcal{P}(\mathcal{X})$ properly, we will consider the asymptotic bias $|T(P) - T(P_0)|$ and asymptotic variance $A(P,T)$ when $P \in \mathcal{P}(\mathcal{X})$ is taken from a small neighbourhood of P_0. A natural choice for such a neighbourhood would be using the weak topology of the space of measures $\mathcal{P}(\mathcal{X})$; this will be detailed in Section 2. An even simpler idea is to consider the *contamination neighbourhood* of P_0 defined for $\varepsilon > 0$ by

$$\mathcal{P}_\varepsilon(P_0) = \{P \in \mathcal{P}(\mathcal{X}) : P = (1-\varepsilon)\,P_0 + \varepsilon\,Q, \text{ for any } Q \in \mathcal{P}(\mathcal{X})\}. \tag{14}$$

This set consists of all contaminated versions of the ideal measure P_0 by other measures Q, where the maximum contamination allowed is $\varepsilon > 0$. We already encountered this neighbourhood in Examples 1.2 and 1.3. Note that (14) is not a neighbourhood in the sense of the weak topology on $\mathcal{P}(\mathcal{X})$.

Definition 4. For a statistical functional $T\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$, a measure $P_0 \in \mathcal{P}(\mathcal{X})$ and some system of neighbourhoods $\mathcal{P}_\varepsilon(P_0)$ of P_0 with $\varepsilon > 0$ we define the *maximum bias* of T at P_0 as

$$b(\varepsilon, P_0, T) = \sup_{P \in \mathcal{P}_\varepsilon(P_0)} |T(P) - T(P_0)|. \tag{15}$$

The *maximum variance* of T satisfying (13) at P_0 is

$$v(\varepsilon, P_0, T) = \sup_{P \in \mathcal{P}_\varepsilon(P_0)} A(P,T).$$

Of course, both suprema above are taken only over those $P \in \mathcal{P}_\varepsilon(P_0)$ such that $T(P)$ is well-defined, if the domain of T is not the full space $\mathcal{P}(\mathcal{X})$.

A good robust estimator T_n is expected to have low maximum bias and maximum variance. Nevertheless, it is important to realise that both quantities in Definition 4 are asymptotic.

They naturally deal with the functional T, and not with the finite sample situation and T_n. Thus, for example, minimising the asymptotic bias, we handle an expression of the type

$$\sup_{P\in\mathcal{P}_\varepsilon(P_0)} \lim_{n\to\infty} |T(P_n) - T(P_0)|$$

with P_n the empirical measure (5) from P (supposing that $T(P_n)$ is a consistent estimator of $T(P)$). A more natural approach would be to control an expression of the type

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_\varepsilon(P_0)} |T(P_n) - T(P_0)|\,. \tag{16}$$

Since now we take the limit of the supremum, the quantity (16) is larger than $b(\varepsilon, P_0, T)$. It is, however, much more difficult to control (16) because it is necessarily random (it depends on the empirical measure $P_n = P_n(\omega)$). In our analysis, we therefore work with the indices from Definition 4.

An interesting complementary characteristic of robustness is the smallest amount of contamination $\varepsilon > 0$ in maximum bias that completely breaks the functional T down.

Definition 5. The *asymptotic breakdown point* of a statistical functional $T\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ at $P_0 \in \mathcal{P}(\mathcal{X})$ is

$$\varepsilon^*(P_0, T) = \sup\left\{\varepsilon > 0\colon b(\varepsilon, P_0, T) < \lim_{t\to\infty} b(t, P_0, T) = \sup_{P\in\mathcal{P}(\mathcal{X})} |T(P) - T(P_0)|\right\}.$$

For the contamination neighbourhood, note that for $\varepsilon = 1$ we get $\mathcal{P}_\varepsilon(P_0) = \mathcal{P}(\mathcal{X})$.[1] Thus, instead of $t \to \infty$, only $t = 1$ can be taken in Definition 5 for the contamination neighbourhood.

The asymptotic breakdown point can be interpreted as the smallest amount of contamination of P_0 that makes T differ from $T(P_0)$ as extremely as it does for an arbitrary measure P.

Example 1.6. For $T(P) = \int_{\mathbb{R}} x \,\mathrm{d}\, P(x)$ the mean functional and the contamination neighbourhood (14), we have for any $P_0 \in \mathcal{P}_1(\mathbb{R})$ from (7) and $Q \in \mathcal{P}_1(\mathbb{R})$ that $(1-\varepsilon)\,P_0 + \varepsilon\, Q \in \mathcal{P}_\varepsilon(P_0)$ and

$$T((1-\varepsilon)\,P_0 + \varepsilon\, Q) = (1-\varepsilon)T(P_0) + \varepsilon\, T(Q).$$

Since $Q \in \mathcal{P}_1(\mathbb{R})$ can be arbitrary, we can take $T(Q)$ to be arbitrarily large. For T the mean, for any $P_0 \in \mathcal{P}_1(\mathbb{R})$ and $\varepsilon > 0$ we thus have

$$b(\varepsilon, P_0, T) = \infty \quad \text{and} \quad \varepsilon^*(P_0, T) = 0.$$

[1] Observe the subtle difference in notation: $\mathcal{P}_1(\mathcal{X})$ for a normed space $\mathcal{X}$ is the set of $P \in \mathcal{P}(\mathcal{X})$ such that $\int_{\mathcal{X}} \|x\| \,\mathrm{d}\, P(x) < \infty$, and $\mathcal{P}_1(P_0)$ is a 1-neighbourhood of $P_0 \in \mathcal{P}(\mathcal{X})$.

We again see that the mean is very non-robust, having an asymptotic breakdown point equal to 0 and maximum bias equal to infinity for any $\varepsilon > 0$. △

As with the definition of the maximum bias, also $\varepsilon^*(P_0, T)$ is an asymptotic quantity. It is, however, possible to consider also directly the finite sample version of the breakdown point from Definition 5. To avoid problems with the stochastic nature of the empirical measure (5), in this setup, we consider a fixed random sample consisting of points $x_1, \dots, x_n \in \mathcal{X}$. This sample is contaminated by adding $m \geq 0$ additional points $y_1, \dots, y_m$ to the sample; the amount of contamination is controlled by $m/(m+n) \leq \varepsilon$. The corresponding maximum bias of T is the maximum deviation of T when applied to the contaminated dataset and compared with the situation without contamination. The breakdown point is the smallest fraction ε that causes T to break down completely.

Definition 6. Suppose we have a set $\boldsymbol{X}$ of points $x_1, \dots, x_n \in \mathcal{X}$ and $\varepsilon \in (0, 1)$. Let $\{T_n\}_{n=1}^{\infty}$ be a sequence of estimators represented by a statistical functional T. The *finite sample maximum bias* of T at $\boldsymbol{X}$ is

$$b(\varepsilon, \boldsymbol{X}, T) = \sup_{y_1, \dots, y_m \in \mathcal{X}} \left| T_{n+m}(x_1, \dots, x_n, y_1, \dots, y_m) - T_n(x_1, \dots, x_n) \right|, \tag{17}$$

where m is any integer such that $m/(m+n) \leq \varepsilon$.

The *finite sample breakdown point* of T at $\boldsymbol{X}$ is

$$\varepsilon^*(\boldsymbol{X}, T) = \inf \left\{ \varepsilon > 0 \colon b(\varepsilon, \boldsymbol{X}, T) = \infty \right\}. \tag{18}$$

Naturally, if T takes values only in a bounded subset S of $\mathbb{R}$, one modifies the definition in (18) to the smallest ε-contamination of $\boldsymbol{X}$ that drags T to the boundary of S. In the situation when T takes values in $\mathbb{R}^p$ or a general normed space $\mathcal{X}$, one replaces the absolute value in (17) by the (Euclidean) norm.

Very often, the (asymptotic or finite sample) breakdown point of T does not depend on $\boldsymbol{X}$ or P_0. Also, under natural conditions, it can be expected that if $\boldsymbol{X}$ is sampled from P, the finite sample breakdown point of T converges to its asymptotic counterpart as $n \to \infty$.

The finite sample breakdown point of any "reasonable" functional T is naturally bounded from above by $\varepsilon^*(\boldsymbol{X}, T) \leq 1/2$. This can be seen directly because if $\varepsilon > 1/2$, then $m > n$, and in the finite sample maximum bias (17) the contaminating observations $y_1, \dots, y_m$ are already the majority of the data. A formal justification for this phenomenon with additional insights into the maximum finite sample breakdown point can be found in [6, 5].

2 The space of measures and generalised derivatives

We work in the space of Borel probability measures $\mathcal{P}(\mathcal{X})$, where $\mathcal{X}$ is typically the Euclidean space $\mathbb{R}^k$ for some $k \geq 1$. More generally, we could take $\mathcal{X}$ to be any Polish space, that is any topological space $\mathcal{X}$ whose topology is metrizable by a metric $d\colon \mathcal{X} \times \mathcal{X} \to [0, \infty)$, such that $\mathcal{X}$ is complete (that is, each Cauchy sequence in $\mathcal{X}$ has a limit in $\mathcal{X}$) and separable (that is, the space $\mathcal{X}$ contains a countable dense subset).

The space $\mathcal{P}(\mathcal{X})$ is equipped with the weak(-star) topology, that is the weakest topology that makes the map

$$\mathcal{P}(\mathcal{X}) \to \mathbb{R}\colon P \mapsto \int_{\mathcal{X}} f(x) \,\mathrm{d}\, P(x)$$

continuous for every bounded continuous function $f\colon \mathcal{X} \to \mathbb{R}$. Convergence of measures in the weak topology is precisely the weak convergence of measures: $P_n \xrightarrow[n\to\infty]{\mathrm{w}} P$ in $\mathcal{P}(\mathcal{X})$ if and only if

$$\int_{\mathcal{X}} f(x) \,\mathrm{d}\, P_n(x) \xrightarrow[n\to\infty]{} \int_{\mathcal{X}} f(x) \,\mathrm{d}\, P(x).$$

We already know a lot about the weak topology of measures:

- Weak convergence is characterised by the portmanteau theorem [16, Theorem 13.2];
- for $\mathcal{X} = \mathbb{R}$, weak convergence to P is equivalent to point-wise convergence of distribution functions at each point of continuity of the distribution function F of P [16, Theorem 13.12]; and
- by the Prokhorov theorem [16, Theorem 12.8], we know that a set of measures $\mathcal{S} \subset \mathcal{P}(\mathcal{X})$ contains a weakly convergent subsequence if and only if $\mathcal{S}$ is tight, meaning that for every $\varepsilon > 0$ there exists a compact set $K \subset \mathcal{X}$ such that for all $P \in \mathcal{S}$ we have $P(K) \geq 1 - \varepsilon$.

To quantify the robustness of statistical functionals appropriately, we need a metric on the space of measures, and the associated notion of a neighbourhood. Many metrics with different properties have been defined in $\mathcal{P}(\mathcal{X})$. For example, we are already familiar with the Kolmogorov distance for $\mathcal{X} = \mathbb{R}$ given by

$$d_K(P, Q) = \sup_{x \in \mathbb{R}} |F(x) - G(x)|, \tag{19}$$

where F and G are the distribution functions of P and Q, respectively. The Kolmogorov distance is, however, not compatible with weak convergence in $\mathcal{P}(\mathbb{R})$; for $\delta_{1/n}$ the Dirac measure at $1/n \in \mathbb{R}$, we have $\delta_{1/n} \xrightarrow[n\to\infty]{\mathrm{w}} \delta_0$ in $\mathcal{P}(\mathbb{R})$, but $d_K(\delta_{1/n}, \delta_0) = 1$ for each $n = 1, 2, \ldots$.

We will work with two metrics that do metrize weak convergence: in $\mathcal{X} = \mathbb{R}$ we introduce the Lévy metric, and in a general Polish space $\mathcal{X}$ we use the Prokhorov metric.

Definition 7. The *Lévy distance* between P and Q in $\mathcal{P}(\mathbb{R})$ is

$$d_L(P,Q) = \inf\{\varepsilon > 0 \colon F(x-\varepsilon) - \varepsilon \leq G(x) \leq F(x+\varepsilon) + \varepsilon \text{ for all } x \in \mathbb{R}\}, \tag{20}$$

where F is the distribution function of P and G is the distribution function of Q.

As we explained in Remark 1, it will often be more convenient to write directly $d_L(F,G)$ instead of $d_L(P,Q)$.

The Lévy distance d_L is somewhat similar to the Kolmogorov distance d_K. In d_K in (19), we evaluate the greatest difference between F and G in terms of their vertical distance. In d_L in (20), one measures the discrepancy between F and G in terms of both vertical and horizontal shifts; see Figure 4.

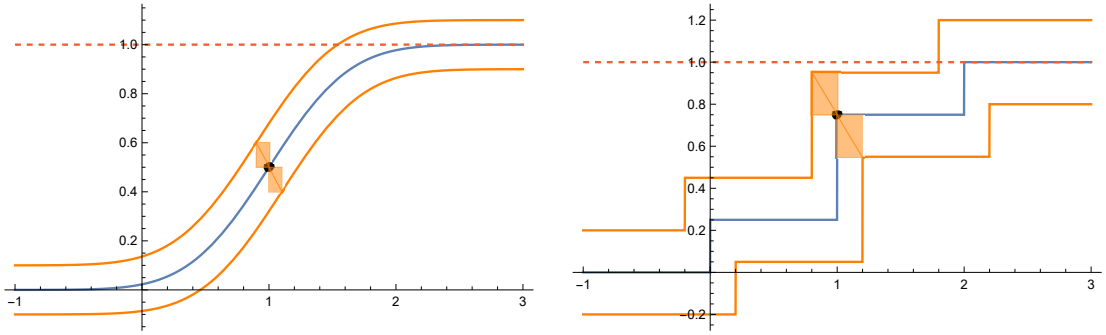

Figure 4: The Lévy metric: Two distribution functions in blue (left panel: a continuous distribution, right panel: a discrete distribution) with their Lévy neighbourhoods with boundaries in orange. The Lévy distance can be interpreted as drawing squares of side length ε with one corner at each $(x, F(x))^\mathsf{T} \in \mathbb{R}^2$ (orange squares for $x = 1$ in both figures). The Lévy ε-neighbourhood of F is the region in $\mathbb{R}^2$ covered by the union of all such squares; each distribution function G lying in this set completely is of Lévy distance at most ε from F.

Theorem 1. *The Lévy distance is a metric that metrizes the weak topology in* $\mathcal{P}(\mathbb{R})$.

Proof. To show that d_L is a metric, we need to establish that for all $P, Q, R \in \mathcal{P}(\mathbb{R})$: (i) $d_L(P,Q) \geq 0$ and $d_L(P,Q) = 0$ if and only if $P = Q$; (ii) $d_L(P,Q) = d_L(Q,P)$; and (iii) $d_L(P,R) \leq d_L(P,Q) + d_L(Q,R)$. All these follow directly from the definition of the Lévy distance (20).

Now we prove that d_L metrizes the weak topology. To prove the first implication, assume that $d_L(F_n, F) \to 0$ as $n \to \infty$, and take $x \in \mathbb{R}$ that is a point of continuity of F. Then $F(x+\varepsilon) + \varepsilon \to F(x)$ and $F(x-\varepsilon) - \varepsilon \to F(x)$ as $\varepsilon \to 0$ because of the continuity of F at x, which means that in the definition of d_L we get $F_n(x) \to F(x)$. This means that $P_n \xrightarrow[n\to\infty]{\mathrm{w}} P$.

For the other implication, suppose that $P_n \xrightarrow[n\to\infty]{\text{w}} P$, meaning that $F_n(x) \to F(x)$ as $n \to \infty$ for each continuity point x of F. Take $\varepsilon > 0$ and find a sequence $x_0 < x_1 < \cdots < x_N$ of continuity points of F such that $F(x_0) < \varepsilon/2$, $F(x_N) > 1 - \varepsilon/2$, and $x_{i+1} - x_i < \varepsilon$ for each $i = 1, \ldots, N$. This is possible since any F is non-decreasing and thus has at most countably many points of discontinuity. Take $n_0 \geq 1$ large enough so that for all $i = 0, 1, \ldots, N$ and $n \geq n_0$ we have $|F_n(x_i) - F(x_i)| < \varepsilon/2$. For any $x \in [x_{i-1}, x_i]$ we then have

$$F_n(x) \leq F_n(x_i) < F(x_i) + \varepsilon/2 \leq F(x + \varepsilon) + \varepsilon.$$

The last inequality comes from $x_i \leq x + \varepsilon$, which is true because $x_{i-1} \leq x \leq x_i < x_{i-1} + \varepsilon$. For $x < x_0$ we have similarly

$$F_n(x) \leq F_n(x_0) < F(x_0) + \varepsilon/2 \leq \varepsilon \leq F(x + \varepsilon) + \varepsilon,$$

and for $x > x_N$

$$F_n(x) \leq 1 < 1 + \varepsilon/2 < F(x_N) + \varepsilon \leq F(x + \varepsilon) + \varepsilon.$$

To get a bound on $F_n(x)$ from below, we write analogously for $x \in [x_{i-1}, x_i]$

$$F_n(x) \geq F_n(x_{i-1}) \geq F(x_{i-1}) - \varepsilon/2 \geq F(x - \varepsilon) - \varepsilon.$$

For $x > x_N$ we have

$$F_n(x) \geq F_n(x_N) \geq F(x_N) - \varepsilon/2 \geq 1 - \varepsilon \geq F(x - \varepsilon) - \varepsilon,$$

and finally for $x < x_0$ we write

$$F_n(x) \geq 0 \geq F(x_0) - \varepsilon \geq F(x - \varepsilon) - \varepsilon.$$

Overall, we obtain that for any $\varepsilon > 0$ we can find $n_0 \geq 1$ such that for all $n \geq n_0$ we have $d_L(P_n, P) \leq \varepsilon$, as we wanted to show. □

The Lévy distance is applicable only for $\mathcal{X} = \mathbb{R}$; the Prokhorov distance can be used in any Polish space $\mathcal{X}$. To define it, we need the concept of a δ-neighbourhood of a set $A \subseteq \mathcal{X}$

$$A^\delta = \left\{ x \in \mathcal{X} \colon \inf_{y \in A} d(x, y) \leq \delta \right\}.$$

The set A^δ is clearly always closed in $\mathcal{X}$.

Definition 8. The *Prokhorov distance* between P and Q in $\mathcal{P}(\mathcal{X})$ is

$$d_P(P, Q) = \inf \{ \varepsilon > 0 \colon P(A) \leq Q(A^\varepsilon) + \varepsilon \text{ for all } A \subseteq \mathcal{X} \text{ Borel} \}. \tag{21}$$

Just as the Lévy metric in $\mathcal{P}(\mathbb{R})$, the Prokhorov metric metrizes the weak convergence in $\mathcal{P}(\mathcal{X})$. It even makes the space of measures $\mathcal{P}(\mathcal{X})$ Polish.

Theorem 2. *For any Polish space $\mathcal{X}$, the Prokhorov distance is a metric that metrizes the weak topology in $\mathcal{P}(\mathcal{X})$. In addition, the space $\mathcal{P}(\mathcal{X})$ equipped with weak topology is Polish.*

Proof. The proof is not difficult. But it is somewhat technical and requires a certain amount of topology. Thus, we omit it; it can be found in [12, Lemma 2.12, Theorems 2.14 and 2.15]. □

The following inequalities between probability metrics will be useful. A much more detailed discussion on relations between metrics for probability measures can be found in [9].

Lemma 1. *For any $P, Q \in \mathcal{P}(\mathbb{R})$ we have*

$$d_L(P,Q) \le d_P(P,Q) \quad \text{and} \quad d_L(P,Q) \le d_K(P,Q).$$

Proof. The definition of the Lévy distance (20) is basically just the Prokhorov distance (21) with sets A of the form $(-\infty, x]$ or $[x, \infty)$ with $x \in \mathbb{R}$. We obtain that there will be more constants $\varepsilon > 0$ satisfying the condition for d_L than for d_P, and consequently $d_L(P,Q) \le d_P(P,Q)$. The inequality for d_K follows directly from the definition or by inspecting Figure 4. □

We have seen in Theorems 1 and 2 that both the Lévy and Prokhorov distance metrize the weak topology in $\mathcal{P}(\mathbb{R})$. It means that $d_L(P_n, P) \to 0$ if and only if $d_P(P_n, P) \to 0$ as $n \to \infty$. This, however, does not mean that the two metrics are equivalent in the sense that there exists a constant $c > 0$ such that $d_P(P,Q) \le c\, d_L(P,Q)$ for all $P, Q \in \mathcal{P}(\mathbb{R})$. In fact, a stronger claim can be shown. It is possible to find two sequences of measures $\{P_n\}_{n=1}^{\infty}$, $\{Q_n\}_{n=1}^{\infty}$ in $\mathcal{P}(\mathbb{R})$ such that $d_L(P_n, Q_n) \to 0$ as $n \to \infty$, but $d_P(P_n, Q_n) \ge 1/2$ for all $n = 1, 2, \ldots$, see [7, Problem 8 in Section 11.3].

The end of lecture 2 (29.02.2024)

2.1 Derivatives in the space of measures

The space of measures $\mathcal{P}(\mathcal{X})$ is convex, but not linear: a convex combination $P_t = t\,P + (1-t)\,Q$ from (4) belongs to $\mathcal{P}(\mathcal{X})$ for $t \in [0,1]$, but not for $t < 0$. When we need to impose a linear structure on $\mathcal{P}(\mathcal{X})$, we embed this space into the linear space of all signed measures on $\mathcal{X}$, defined precisely as the linear space generated by all finite linear combinations of elements from $\mathcal{P}(\mathcal{X})$. A signed measure can attain both positive and negative values. We write $\mathcal{P}'(\mathcal{X})$ for the space of all signed measures on $\mathcal{X}$.

We now explore the possibilities to define derivatives of functionals T from $\mathcal{P}(\mathcal{X})$. For the following definition, let d_* be a metric on $\mathcal{P}(\mathcal{X})$ that metrizes weak topology. We further

require that d_* is compatible with the affine structure on $\mathcal{P}(\mathcal{X})$, meaning that for any $F_t = (1-t)\,F + t\,G$, $t \in [0,1]$ it satisfies

$$d_*(F_t, F_s) \leq |t-s|\,. \tag{22}$$

For the Prokhorov metric, this is true since for any $A \subseteq \mathcal{X}$ Borel we have

$$\begin{aligned} |F_t(A) - F_s(A)| &= |(1-t)\,F(A) + t\,G(A) - (1-s)\,F(A) - s\,G(A)| \\ &= |t-s|\,|G(A) - F(A)| \leq |t-s|\,, \end{aligned}$$

and immediately from (21) we get $d_P(F_t, F_s) \leq |t-s|$. The Lévy distance verifies (22) by Lemma 1. Overall, for the metric d_*, we can take either d_L or d_P.

Definition 9 (Fréchet derivative)**.** A statistical functional $T\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ is said to be *Fréchet differentiable* (with respect to d_*) at $P \in \mathcal{P}(\mathcal{X})$ if there exists a continuous linear functional $L = L_P\colon \mathcal{P}'(\mathcal{X}) \to \mathbb{R}$ (depending on P) such that we can write

$$\lim_{d_*(P,Q)\to 0} \frac{|T(Q) - T(P) - L(Q-P)|}{d_*(P,Q)} = 0, \tag{23}$$

where the limit is taken over all (sequences of) measures $Q \in \mathcal{P}(\mathcal{X})$ converging weakly to P. The functional $L = L_P$ is called the *Fréchet derivative* of T at P.

Although this concept is customarily called the Fréchet derivative, it is, strictly speaking, not precisely the Fréchet derivative in topological vector spaces known from functional analysis [26, Section 7.1.1]. The reason is that in (23), we do not consider the limit in the linear space of signed measures $\mathcal{P}'(\mathcal{X})$. Instead, we approach P only in the space of probability measures $\mathcal{P}(\mathcal{X})$. This is obvious once one realises that the functional T is not even defined for general signed measures on $\mathcal{X}$. For this reason, one has to be careful with applying functional analytic results to our concept of the Fréchet derivative.

It is important to note that different authors may define Fréchet derivatives slightly differently. In [12, Section 2.5], for example, it is not assumed that the linear functional L is continuous (or, equivalently, bounded, see [26, Theorem 1.1.28]). Note that in infinite-dimensional spaces, such as $\mathcal{P}(\mathcal{X})$ is, there exist discontinuous linear functionals. We follow the convention from [27, Section 20.2] and [28, Section 3.9], and assume that L must also be continuous, which then implies that if Fréchet derivative of T at P exists, then T must be (weakly) continuous at P.

Theorem 3. *Let $T\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ be a statistical functional that Fréchet differentiable at P. Then*

- *T is (weakly) continuous in at $P \in \mathcal{P}(\mathcal{X})$,*

- *the Fréchet derivative $L = L_P$ of T at P can be represented as*

$$L_P(Q - P) = \int_{\mathcal{X}} \psi_P(x) \, \mathrm{d}\, Q(x) \tag{24}$$

with $\psi_P \colon \mathcal{X} \to \mathbb{R}$ bounded and continuous, and

- $\int_{\mathcal{X}} \psi_P(x) \, \mathrm{d}\, P(x) = L_P(0) = 0$.

Proof. Let F be the distribution function of P and G the distribution function of Q. First, we prove that L_P is essentially unique. Suppose that both L_1 and L_2 satisfy (23). Take $F_t = (1-t)\, F + t\, G$ with $t \in (0,1)$. Then we have by (22) that $d_*(F_t, F) \leq t$, and we can write

$$\lim_{t \to 0} \frac{|T(F_t) - T(F) - L_i(F_t - F)|}{d_*(F_t, F)} = 0 \quad \text{for } i = 1, 2.$$

This gives

$$\begin{aligned} 0 &\leq \lim_{t \to 0} \frac{|L_1(F_t - F) - L_2(F_t - F)|}{d_*(F_t, F)} \\ &\leq \lim_{t \to 0} \left(\frac{|T(F_t) - T(F) - L_1(F_t - F)|}{d_*(F_t, F)} + \frac{|T(F_t) - T(F) - L_2(F_t - F)|}{d_*(F_t, F)} \right) = 0, \end{aligned}$$

but also $F_t - F = t\,(G - F)$, and by the linearity of L_i we have

$$|L_1(F_t - F) - L_2(F_t - F)| = |t| \, |(L_1 - L_2)(G - F)| \, .$$

Together we get, using again (22), that

$$0 = \lim_{t \to 0} \frac{|t| \, |(L_1 - L_2)(G - F)|}{d_*(F_t, F)} \geq \lim_{t \to 0} \frac{|t| \, |(L_1 - L_2)(G - F)|}{t} = |(L_1 - L_2)(G - F)| \, ,$$

that is $L_1(G - F) = L_2(G - F)$ for all $G \in \mathcal{P}(\mathcal{X})$. This, of course, does not mean that $L_1(G) = L_2(G)$ for all $G \in \mathcal{P}(\mathcal{X})$. But, since F is fixed we can standardise L by, say, assuming $L(F) = 0$, which then gives $L_1(G) = L_1(G - F) = L_2(G - F) = L_2(G)$ for all $G \in \mathcal{P}(\mathcal{X})$ as we wanted to show. Thus, fixing $L(F) = 0$, the linear functional L in (23) is unique.

Next we show that T is continuous at P. To see that, let $d_*(Q_n, P) \to 0$ as $n \to \infty$. We can write

$$\begin{aligned} 0 \leq \limsup_{n \to \infty} |T(Q_n) - T(P)| &\leq \lim_{d_*(Q,P) \to 0} d_*(Q, P) \frac{|T(Q) - T(P) - L(Q - P) + L(Q - P)|}{d_*(Q, P)} \\ &\leq \lim_{d_*(Q,P) \to 0} d_*(Q, P) \frac{|T(Q) - T(P) - L(Q - P)|}{d_*(Q, P)} + \lim_{d_*(Q,P) \to 0} |L(Q - P)| = 0, \end{aligned}$$

because of the Fréchet differentiability (23) of T, and the continuity of the linear functional L. We see that T must be continuous at P.

Now we want to show (24); we only sketch this part of the proof. The function ψ_P is at $x \in \mathcal{X}$ defined by means of taking the Dirac measure $\delta_x \in \mathcal{P}(\mathcal{X})$, and setting

$$L(\delta_x) = \psi_P(x) = \int_{\mathcal{X}} \psi_P(y) \,\mathrm{d}\, \delta_x(y).$$

Because L is linear, for any measure $Q = \sum_{i=1}^m \alpha_i \, \delta_{x_i} \in \mathcal{P}(\mathcal{X})$ with $\alpha_i > 0$, $i = 1, \dots, m$, supported in a finite number of points $x_1, \dots, x_m \in \mathcal{X}$, we can extend the previous formula to

$$L(Q) = \sum_{i=1}^m \alpha_i \, L(\delta_{x_i}) = \sum_{i=1}^m \alpha_i \, \psi_P(x_i) = \int_{\mathcal{X}} \psi_P(y) \,\mathrm{d}\, Q(y).$$

The formula for a general measure $Q \in \mathcal{P}(\mathcal{X})$ is obtained by approximating Q using finitely supported measures, continuity of L, and weak continuity of T; for details, see [12, Section 2.5].

For the final assertion, by (24) we have $\int_{\mathcal{X}} \psi_P(x) \,\mathrm{d}\, P(x) = L_P(P - P) = 0$. □

Remark 3. We saw in the proof above that the functional L_P is defined uniquely up to the (arbitrary) choice of $L_P(P) \in \mathbb{R}$. From now on, we thus make a convention, and will always consider functional derivatives L_P that satisfy $L_P(P) = 0$, without loss of generality.

The concept of Fréchet differentiability of a statistical functional is very strong. Using a Taylor expansion, it immediately gives an asymptotic normality result. In the proof of this theorem, we will need the famous Varadarajan theorem that states that empirical measures $P_n = P_n(\omega)$ from (5) converge to the true sampling distribution $P \in \mathcal{P}(\mathcal{X})$ weakly (as measures in $\mathcal{P}(\mathcal{X})$) almost surely (in the random element $\omega \in \Omega$). This is an interesting complement to the Glivenko-Cantelli theorem (Theorem 6 below in Section 2.3). Varadarajan's theorem can be applied in any separable metric space $\mathcal{X}$ (in particular, it works in Polish spaces $\mathcal{X}$, as we assume throughout this text).

Theorem 4 (Varadarajan)**.** *Let $P \in \mathcal{P}(\mathcal{X})$. Then the empirical measures $P_n = P_n(\omega) \in \mathcal{P}(\mathcal{X})$ from (5) converge to P weakly almost surely, that is*

$$\mathsf{P}\left(\left\{\omega \in \Omega \colon P_n(\omega) \xrightarrow[n\to\infty]{\mathrm{w}} P\right\}\right) = 1.$$

Proof. We give only a sketch of the proof. For any $f \colon \mathcal{X} \to \mathbb{R}$ bounded and continuous we have

$$\int_{\mathcal{X}} f(x) \,\mathrm{d}\, P_n(x) = \frac{1}{n} \sum_{i=1}^n f(X_i) \to \mathsf{E}_P \, f(X_1) = \int_{\mathcal{X}} f(x) \,\mathrm{d}\, P(x) \quad \text{almost surely,} \tag{25}$$

because of the usual law of large numbers. The proof is completed by observing that the space of bounded continuous functions on a Polish space $\mathcal{X}$ is separable (with respect to an appropriately chosen metric). We apply (25) to each element of a countable dense subset of this space, and continuity of the functions f allows to expand (25) also to all other bounded continuous functions. For details, we refer to [7, Theorem 11.4.1]. □

Observe that as an immediate consequence of the Varadarajan theorem 4 we obtain that any weakly continuous statistical functional $T\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}^p$ induces strongly consistent estimators

$$\mathsf{P}\left(\left\{\omega \in \Omega\colon T(P_n(\omega)) \xrightarrow[n\to\infty]{} T(P)\right\}\right) = 1,$$

that is

$$T(P_n) \to T(P) \quad \text{almost surely as } n \to \infty.$$

We are now ready to state a central limit theorem for Fréchet differentiable statistical functionals. In this result, we use some basic stochastic O notation, see [20, Definition 1].

Theorem 5. *Let $P_n = P_n(\omega) \in \mathcal{P}(\mathcal{X})$ be a sequence of empirical measures (5) corresponding to $X_1, X_2, \ldots$ sampled independently from $P \in \mathcal{P}(\mathcal{X})$. Let the metric d_* satisfy*

$$d_*(P, P_n) = O_{\mathsf{P}}(n^{-1/2}), \tag{26}$$

meaning that the sequence of laws $\mathcal{L}_P(\sqrt{n}\, d_*(P, P_n))$ is tight. Suppose that a statistical functional T has a Fréchet derivative (with respect to d_*) at P represented by a function ψ_P as in Theorem 3. Then

$$\sqrt{n}\,(T(P_n) - T(P)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_P(X_i) + o_{\mathsf{P}}(1), \tag{27}$$

where $o_{\mathsf{P}}(1)$ stands for a remainder term that vanishes in probability as $n \to \infty$.

If, in addition, $A(P,T)$ defined as $A(P,T) = \int_{\mathcal{X}} (\psi_P(x))^2 \,\mathrm{d}\, P(x)$ is non-zero and finite, then we can write

$$\sqrt{n}\,(T(P_n) - T(P)) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}(0, A(P,T)).$$

Proof. From (23) and Theorem 4 we have for (P-almost) any fixed $\omega \in \Omega$

$$\lim_{n\to\infty} \frac{|T(P_n(\omega)) - T(P) - L(P_n(\omega) - P)|}{d_*(P, P_n(\omega))} = 0,$$

which means precisely

$$T(P_n(\omega)) - T(P) - L(P_n(\omega) - P) = o(d_*(P, P_n(\omega))).$$

Rearranging the terms in the previous formula and using $L(P) = 0$ gives

$$\begin{aligned}
\sqrt{n}\,(T(P_n(\omega)) - T(P)) &= \sqrt{n}\, L(P_n(\omega)) + o(\sqrt{n}\, d_*(P, P_n(\omega))) \\
&= \frac{\sqrt{n}}{n} \sum_{i=1}^{n} L(\delta_{X_i(\omega)}) + o(\sqrt{n}\, d_*(P, P_n(\omega))) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_P(X_i(\omega)) + o(\sqrt{n}\, d_*(P, P_n(\omega))).
\end{aligned}$$

We used $P_n(\omega) = \sum_{i=1}^n \delta_{X_i(\omega)}/n$ which follows from the definition of an empirical measure (5), and (24). Dropping $\omega \in \Omega$ and considering P_n as a random measure again, this can be rewritten to

$$\sqrt{n}\,(T(P_n) - T(P)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_P(X_i) + o(1)\, O_{\mathsf{P}}(1),$$

where we used our assumption that $\sqrt{n}\, d_*(P, P_n(\omega)) = O_{\mathsf{P}}(1)$. Finally, using $o(1)\, O_{\mathsf{P}}(1) = o_{\mathsf{P}}(1)$ [20, Remark 3], we obtain the first part of our theorem.

For our second claim, note that $\psi_P(X_1), \psi_P(X_2), \ldots$ is a sequence of independent identically distributed random variables with

$$\mathsf{E}\, \psi_P(X_1) = \int_{\mathcal{X}} \psi_P(x)\, \mathrm{d}\, P(x) = L(P) = 0$$

using Theorem 3. Thus, we can use the ordinary central limit theorem on the right-hand side of (27) with

$$\mathsf{var}\, \psi_P(X_1) = \mathsf{E}\, (\psi_P(X_1))^2 = A(P, T),$$

as we wanted to show. □

Theorem 5 says that, if T is Fréchet differentiable and if the metric d_* is chosen well so that (26) is true, we obtain an immediate asymptotic normality result. The condition (26) is not terribly strict (at least if $\mathcal{X} = \mathbb{R}$). It is true for, e.g., the Kolmogorov distance d_K from (19) in $\mathcal{X} = \mathbb{R}$, as for d_K the condition (26) follows from an asymptotic normality result for the Kolmogorov-Smirnov test statistic [15, Theorem 5.2]

$$\sqrt{n}\, d_K(P, P_n) = \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{\mathrm{d}} Z,$$

for Z with the Kolmogorov distribution. Lemma 1 then gives (26) also for the Lévy metric. It is also interesting to note that (26) does not hold true for $\mathcal{X} = \mathbb{R}$ and the Prokhorov metric; for details see [12, page 40] and the references therein.

The difficult part of applying Theorem 5 is the requirement of Fréchet differentiability of T, which may be hard to verify. As argued in [8, Example 2.3.2], already functionals T assigning to P its quantiles fail to be Fréchet differentiable. Several other types of functionals have been found to be Fréchet differentiable, under various technical conditions. For an overview of some of these results see [3].

There exist concepts of differentiability in $\mathcal{P}(\mathcal{X})$ that are weaker than the Fréchet derivative. The weakest one is the Gâteaux differentiation, which is simply the directional derivative along a line segment.

Definition 10 (Gâteaux derivative)**.** A statistical functional $T\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ is said to be *Gâteau differentiable* at $P \in \mathcal{P}(\mathcal{X})$ if there exists a measurable function $\psi_P\colon \mathcal{X} \to \mathbb{R}$ satisfying $\int_{\mathcal{X}} \psi_P(x)\,\mathrm{d}\,P(x) = 0$ such that for all $Q \in \mathcal{P}(\mathcal{X})$ we can write

$$\lim_{t\to 0} \frac{T(P_t) - T(P)}{t} = \int_{\mathcal{X}} \psi_P(x)\,\mathrm{d}\,Q(x), \tag{28}$$

where $P_t = (1-t)\,P + t\,Q$ for $t \in [0,1]$. The linear functional

$$L_P\colon Q \mapsto \int_{\mathcal{X}} \psi_P(x)\,\mathrm{d}\,Q(x)$$

is called the *Gâteaux derivative* of the functional T at P.

Observe that in Definition 10, we assume that the linear functional L_P can be represented by a function ψ_P; for Fréchet derivatives, this was not assumed but was proved to be true in Theorem 3. For Gâteaux derivatives, this assumption must be made. The difference between Gâteaux and Fréchet derivatives is similar to the difference between directional derivatives and the total differential for functions from $\mathbb{R}^k$ [22, Section 11.1]. In particular, if the Fréchet derivative exists, then so does the Gâteaux derivative, and the two derivatives are equal. On the other hand, unlike the Fréchet differentiability, the Gâteaux differentiability of T at P does not guarantee weak continuity of T at P; see [26, Remark 7.1.2(c)] for an example in $\mathbb{R}^2$.

Note that on the left-hand side of (28), we have simply the standard derivative of a real function $t \mapsto T(P_t)$ defined in $[0,1] \subset \mathbb{R}$. Its value can therefore be interpreted as the directional derivative of T at P in direction Q.

The notion of Gâteaux differentiability is sufficient for our purposes, but it is too weak to be used as a derivative in, for example, asymptotic expansions of functionals. Results such as our Theorem 5 are not valid under Gâteaux differentiability. On the other hand, establishing Fréchet differentiability of functionals is very technical and usually complicated. As a compromise, another notion of a derivative, called Hadamard (or compact) differentiability, is often useful.

Definition 11 (Hadamard derivative)**.** A statistical functional $T\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ is said to be *Hadamard differentiable* at $P \in \mathcal{P}(\mathcal{X})$ if there exists a continuous linear functional $L = L_P\colon \mathcal{P}'(\mathcal{X}) \to \mathbb{R}$ (depending on P) such that

$$\lim_{t_n\to 0} \lim_{Q_n \xrightarrow[n\to\infty]{\mathrm{w}} Q} \frac{T(P + t_n\,Q_n) - T(P)}{t_n} = L(Q - P), \tag{29}$$

where the limit is taken for any sequences $t_n \to 0$ and $Q_n \xrightarrow[n\to\infty]{\mathrm{w}} Q$ in $\mathcal{P}(\mathcal{X})$, for any $Q \in \mathcal{P}(\mathcal{X})$ given. The linear functional $L = L_P$ is called the *Hadamard derivative* of the functional T at P.

If in the limit (29) we require only $Q, Q_n \in \mathcal{P}_0(\mathcal{X})$ for each n for some subset $\mathcal{P}_0(\mathcal{X})$ of $\mathcal{P}_0(\mathcal{X})$, then we say that T is Hadamard differentiable at P *tangentially* to $\mathcal{P}_0(\mathcal{X})$.

The difference between the Gâteaux and Hadamard differentiability is that for the Gâteaux derivative, in (28) we approach P only from the fixed direction of Q. On the other hand, for the Hadamard differentiability, we require the limit to be valid also if the directions of Q_n change, and only converge to Q. Finally, for the Fréchet derivative, we require the convergence to be also uniform in Q_n and Q, with respect to the metric d_*. Thus, roughly speaking, Fréchet $\implies$ Hadamard $\implies$ Gâteaux differentiability. To sum things up, it is useful to interpret the three notions of differentiability in a single framework. For $L\colon \mathcal{P}'(\mathcal{X}) \to \mathbb{R}$ a continuous linear functional, define a *remainder of a functional* $T\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ at $Q \in \mathcal{P}(\mathcal{X})$ as

$$R(P + t\,Q) = T(P + t\,Q) - T(P) - L(t\,Q) \quad \text{for } t \geq 0,$$

and suppose that

$$\frac{R(P + t\,Q)}{t} \to 0 \quad \text{as } t \to 0. \tag{30}$$

Then we say that

- T is Gâteaux differentiable at P if (30) is true for each $Q \in \mathcal{P}(\mathcal{X})$,
- T is Hadamard differentiable at P if (30) is true uniformly in Q in any compact subset of $\mathcal{P}(\mathcal{X})$, and
- T is Fréchet differentiable at P if (30) is true uniformly in Q in any bounded subset of $\mathcal{P}(\mathcal{X})$.

In any of these situations, we then say that L is the respective derivative of T at P.

In a finite-dimensional space, a bounded closed set is compact. Thus, in finite-dimensional spaces, Hadamard and Fréchet differentiability are the same. In our setup of an infinite-dimensional space $\mathcal{P}(\mathcal{X})$, this is not true. It turns out that Hadamard differentiability is easier to prove, holds true for large classes of common statistical functionals, and is enough for asymptotic expansions of T and inference. The notion of tangential Hadamard derivative makes this even easier, as under some conditions, we are allowed to prove differentiability only in subspaces $\mathcal{P}_0(\mathcal{X})$ of $\mathcal{P}(\mathcal{X})$. For those reasons, Hadamard differentiability proved to be quite useful in asymptotic statistics. For some applications such as the delta method for statistical functionals see [27, Section 20.2] and [28, Section 3.9]. As we will see, however, for (most of) our purposes, the Gâteaux differentiability will be sufficient.

2.2 Influence function

The simplest Gâteaux derivative is obtained when Q is taken to be the Dirac measure $\delta_x \in \mathcal{P}(\mathcal{X})$. In the limit in (28), we obtain the important influence function introduced by Hampel.

Definition 12 (Influence function)**.** The *influence function* of a statistical functional $T\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ at $P \in \mathcal{P}(\mathcal{X})$ is defined as

$$\mathsf{IF}(x, P, T) = \lim_{s \to 0} \frac{T((1-s)\,P + s\,\delta_x) - T(P)}{s}. \tag{31}$$

The existence of the influence function of T is even weaker than the existence of the Gâteaux derivative of T; it only concerns specific directions from P to δ_x. But, again — if the Gâteaux derivative of T at P exists, then (28) must be true also for $Q = \delta_x$, and we obtain

$$\mathsf{IF}(x, P, T) = \int_{\mathcal{X}} \psi_P(y)\,\mathrm{d}\,\delta_x(y) = \psi_P(x) \quad \text{for all } x \in \mathcal{X}. \tag{32}$$

Thus, the influence function does determine the Gâteaux derivative uniquely. In turn, if the Fréchet derivative (or the Hadamard derivative) of T at P exists, it must also be equal to

$$L_P(Q) = \int_{\mathcal{X}} \psi_P(x)\,\mathrm{d}\,Q(x) = \int_{\mathcal{X}} \mathsf{IF}(x, P, T)\,\mathrm{d}\,Q(x) \tag{33}$$

by Theorem 3. In particular, we can rewrite the claim of Theorem 5 into

$$\sqrt{n}\,(T(P_n) - T(P)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathsf{IF}(X_i, P, T) + o_{\mathsf{P}}(1). \tag{34}$$

The expression above can be found in, e.g., [20, Section 4]. Even if the Fréchet differentiability of T is not true, or is hard to prove, formula (34) provides a useful heuristic on how can the asymptotic expansion of a statistical functional look like. In particular, (34) gives an impression of the importance of the influence function in what follows.

Let us compute the influence functions of several simple statistical functionals.

Example 2.1. Take $A \subset \mathcal{X}$ a fixed measurable set and let $T(P) = P(A)$. Then for $P_t = (1-t)\,P + t\,Q$ we have

$$T(P_t) = (1-t)\,P(A) + t\,Q(A),$$

and

$$\lim_{t \to 0} \frac{T(P_t) - T(P)}{t} = T(Q) - T(P) = \int_{\mathcal{X}} (\mathbb{I}\,(x \in A) - T(P))\,\mathrm{d}\,Q(x).$$

We have $\psi_P(x) = \mathbb{I}\,(x \in A) - P(A)$ for $x \in \mathcal{X}$, and the corresponding influence function is

$$\mathsf{IF}(x, P, T) = \mathbb{I}\,(x \in A) - P(A).$$

Theorem 5 gives

$$\sqrt{n}\,(T(P_n) - T(P)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbb{I}\,(X_i \in A) - P(A)) + o_{\mathsf{P}}(1),$$

which is true even without the remainder term $o_{\mathsf{P}}(1)$, and

$$\sqrt{n}\,(T(P_n) - T(P)) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}(0, A(P,T))$$

with $A(P,T) = \mathsf{E}\,(\mathsf{IF}(X_1, P, T))^2 = P(A)(1-P(A))$. This is also true due to the central limit theorem. △

Example 2.2. Let $T(P) = \int_{\mathcal{X}} x \,\mathrm{d}\, P(x)$ be the mean functional defined on $\mathcal{P}_1(\mathcal{X})$. We have for $P_t = (1-t)\,P + t\,Q$ that

$$T(P_t) = \int_{\mathcal{X}} x \,\mathrm{d}\,((1-t)\,P + t\,Q)(x) = (1-t)\,T(P) + t\,T(Q).$$

We get

$$\lim_{t\to 0} \frac{T(P_t) - T(P)}{t} = T(Q) - T(P) = \int_{\mathcal{X}} (x - T(P)) \,\mathrm{d}\, Q(x),$$

that is $\psi_P(x) = x - \int_{\mathcal{X}} y \,\mathrm{d}\, P(y)$ for all $x \in \mathcal{X}$, and T is differentiable at any P and in all directions $Q \in \mathcal{P}_1(\mathcal{X})$ where T is defined. For $Q = \delta_x$ we obtain the influence function of the mean

$$\mathsf{IF}(x, P, T) = x - \int_{\mathcal{X}} y \,\mathrm{d}\, P(y) = x - \mathsf{E}_P\, X. \tag{35}$$

Theorem 5 gives

$$\sqrt{n}\,(T(P_n) - T(P)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mathsf{E}\, X_i) + o_{\mathsf{P}}(1),$$

which is always trivially true even without the remainder term $o_{\mathsf{P}}(1)$, and

$$\sqrt{n}\,(T(P_n) - T(P)) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}(0, A(P,T))$$

with $A(P,T) = \mathsf{E}\,(\mathsf{IF}(X_1, P, T))^2 = \mathsf{var}\, X_1$. This is the usual central limit theorem for sample averages. Note, however, that the Fréchet differentiability of T is not true and strictly speaking, Theorem 5 cannot be applied directly in our situation. △

Example 2.3. Let

$$T(P) = \mathsf{var}\, X = \int_{\mathbb{R}} x^2 \,\mathrm{d}\, P(x) - \left(\int_{\mathbb{R}} x \,\mathrm{d}\, P(x) \right)^2$$

with $X \sim P \in \mathcal{P}_2(\mathbb{R})$. Then, for P_t as in Example 2.2 we have

$$\begin{aligned}
T(P_t) &= \int_{\mathbb{R}} x^2 \,\mathrm{d}\left((1-t)\,P + t\,Q\right)(x) - \left(\int_{\mathbb{R}} x \,\mathrm{d}\left((1-t)\,P + t\,Q\right)(x)\right)^2 \\
&= (1-t)\int_{\mathbb{R}} x^2 \,\mathrm{d}\,P(x) + t\int_{\mathbb{R}} x^2 \,\mathrm{d}\,Q(x) - (1-t)^2\left(\int_{\mathbb{R}} x \,\mathrm{d}\,P(x)\right)^2 \\
&\quad - t^2\left(\int_{\mathbb{R}} x \,\mathrm{d}\,Q(x)\right)^2 - 2\,(1-t)\,t\int_{\mathbb{R}} x \,\mathrm{d}\,P(x)\int_{\mathbb{R}} x \,\mathrm{d}\,Q(x) \\
&= (1-t)\,\mathsf{E}_P\,X^2 + t\,\mathsf{E}_Q\,X^2 - (1-t)^2\,(\mathsf{E}_P\,X)^2 - t^2\,(\mathsf{E}_Q\,X)^2 \\
&\quad - 2\,t\,(1-t)\mathsf{E}_P\,X\,\mathsf{E}_Q\,X,
\end{aligned}$$

where by $\mathsf{E}_Q\,X$ we mean $\int_{\mathbb{R}} x \,\mathrm{d}\,Q(x)$. A simple computation gives

$$\lim_{t\to 0}\frac{T(P_t) - T(P)}{t} = \mathsf{E}_Q\,X^2 - \mathsf{E}_P\,X^2 - 2\,\mathsf{E}_P\,X\,\mathsf{E}_Q\,X + 2\,(\mathsf{E}_P\,X)^2\,.$$

For $Q = \delta_x$ we get the influence function of the variance

$$\mathsf{IF}(x, P, T) = x^2 - \mathsf{E}_P\,X^2 - 2\,x\,\mathsf{E}_P\,X + 2\,(\mathsf{E}_P\,X)^2 = (x - \mathsf{E}_P\,X)^2 - \mathsf{var}_P\,X. \tag{36}$$

Theorem 5 now gives that for

$$S_n^2 = T(P_n) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$$

we can write

$$\sqrt{n}\left(S_n^2 - \mathsf{var}_P\,X\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big((X_i - \mathsf{E}_P\,X)^2 - \mathsf{var}_P\,X\Big) + o_{\mathsf{P}}(1),$$

and the central limit theorem holds true

$$\sqrt{n}\left(S_n^2 - \mathsf{var}_P\,X\right) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}(0, A(P,T))$$

with

$$\begin{aligned}
A(P,T) &= \mathsf{E}_P\,(\mathsf{IF}(X_1, P, T))^2 = \mathsf{E}_P\Big((X_i - \mathsf{E}_P\,X)^2 - \mathsf{var}_P\,X\Big)^2 \\
&= \mathsf{E}_P\,(X_i - \mathsf{E}_P\,X)^4 + (\mathsf{var}_P\,X)^2 - 2\,(\mathsf{var}_P\,X)\,\mathsf{E}_P\,(X_i - \mathsf{E}_P\,X)^2 \\
&= \mathsf{E}_P\,(X_i - \mathsf{E}_P\,X)^4 - (\mathsf{var}_P\,X)^2.
\end{aligned}$$

This is the same result as [15, Theorem 2.6]. △

The influence function has an interesting heuristic interpretation. It can be seen as the infinitesimal effect of adding a single new observation at $x \in \mathcal{X}$ to a very large random sample from P on the estimator T. For the sample mean and the sample variance in Examples 2.2 and 2.3 we see that the influence functions $\mathsf{IF}(x, P, T)$ are unbounded in x. This means that

by adding a single contaminating observation $x \in \mathcal{X}$ far away from $\mathsf{E}_P X$ (see (35) and (36)), the functional T will change its value drastically. In other words, we see again that neither the sample mean nor the sample variance are robust estimators.

As a quantitative measure of robustness based on the influence function, Hampel suggests the following.

Definition 13. For a statistical functional $T\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ with influence function $\mathsf{IF}(\cdot, P, T)$, the *gross error sensitivity* of T at $P \in \mathcal{P}(\mathcal{X})$ is

$$\gamma^*(P,T) = \sup_{x\in\mathcal{X}} |\mathsf{IF}(x,P,T)|\,. \tag{37}$$

The gross error sensitivity is, in a way, an approach similar to the maximum bias from Definition 4. Using (28) and (32) we have the approximate relation

$$T(P_t) - T(P) \approx t \int_{\mathcal{X}} \mathsf{IF}(x,P,T)\,\mathrm{d}\,Q(x) \tag{38}$$

for any $Q \in \mathcal{P}(\mathcal{X})$ and $P_t = (1-t)\,P + t\,Q$, at least if the Gâteaux derivative of T at P exists. Consider now the contamination neighbourhood (14) of P in the definition of maximum bias. Then a distribution $Q \in \mathcal{P}_\varepsilon(P)$ that maximises the difference $|T(P) - T(Q)|$ is by (38) approximately the one that maximises $\left|\int_{\mathcal{X}} \mathsf{IF}(x,P,T)\,\mathrm{d}\,Q(x)\right|$ over $Q \in \mathcal{P}(\mathcal{X})$. We also have

$$\left|\int_{\mathcal{X}} \mathsf{IF}(x,P,T)\,\mathrm{d}\,Q(x)\right| \le \sup_{x\in\mathcal{X}} |\mathsf{IF}(x,P,T)|\,,$$

with equality attained (in an appropriate limit if necessary) by using $Q = \delta_y$ for some $y \in \mathcal{X}$ at the argument of maxima on the right hand side above. Thus, we may hope that

$$b(\varepsilon,P,T) = \sup_{Q\in\mathcal{P}_\varepsilon(P)} |T(Q) - T(P)| \approx \varepsilon \sup_{x\in\mathcal{X}} |\mathsf{IF}(x,P,T)| = \varepsilon\,\gamma^*(P,T). \tag{39}$$

While this is true under additional conditions, there are examples of functionals T that are robust only either in maximum bias, or only in terms of their gross error sensitivity. We will see several examples later in the course.

Further, note that formula (38) is quite interesting also on its own, as it quantifies the approximate difference of T when applied to P and Q (with $t = 1$).

Compared to the other quantities used for assessment robustness like the maximum bias or maximum variance from Definition 4 in Section 1.4, the influence function — being a derivative itself — does not consider ε-neighbourhoods of the true (ideal, or assumed) distribution P as we did with e.g. the maximum variance. It is possible to search for estimators and tests that minimise, say, the asymptotic variance under the condition of bounded gross error sensitivity (37). But, that approach is inherently infinitesimal and different from optimality criteria based on the quantitative robustness from Section 1.4.

We conclude this section by a visual approximation to the influence function based on a random sample of points.

Definition 14. For a set $\boldsymbol{X}$ of points $x_1, \ldots, x_n \in \mathcal{X}$ corresponding to an empirical measure P_n and a statistical functional $T\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$, the *sensitivity curve* of T at $\boldsymbol{X}$ is

$$\begin{aligned} \mathsf{SC}_n(x, \boldsymbol{X}, T) &= \frac{T\left(\frac{n}{n+1}P_n + \frac{1}{n+1}\delta_x\right) - T(P_n)}{1/(n+1)} \\ &= (n+1)\left(T_{n+1}(x_1, \ldots, x_n, x) - T_n(x_1, \ldots, x_n)\right) \quad \text{for } x \in \mathcal{X}. \end{aligned} \tag{40}$$

The sensitivity curve is an empirical version of the influence function. It is obtained by replacing P in (31) by its empirical counterpart P_n and taking $s = 1/(n+1)$. As $n \to \infty$, it can be expected that under reasonable conditions,

$$\mathsf{SC}_n(x, \boldsymbol{X}, T) \xrightarrow[n\to\infty]{\mathsf{P}} \mathsf{IF}(x, P, T),$$

if $\boldsymbol{X}$ corresponds to a random sample from P.

The sensitivity curve can also be used for visualisation. For a random sample $\boldsymbol{X}$ of points $X_1, \ldots, X_{10}$ from $\mathsf{N}(0,1)$, the sensitivity curves of the two estimators of the centre of symmetry from Example 1.2 and the two estimators of the standard deviation from Example 1.3 are displayed in Figure 5.

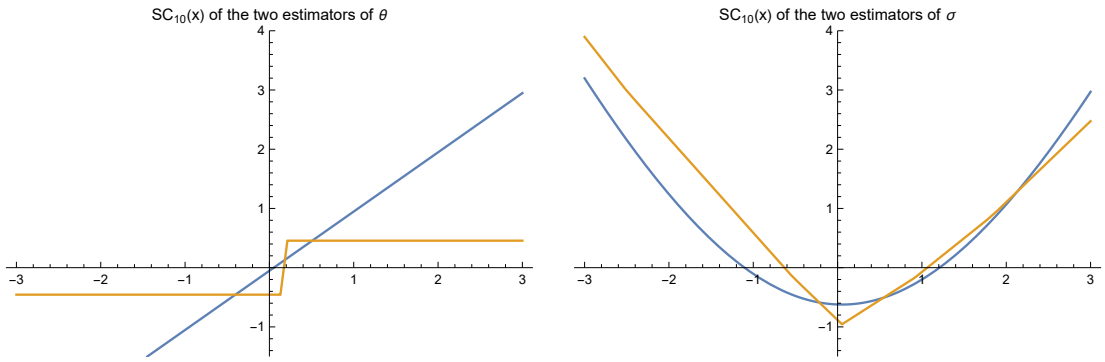

Figure 5: The sensitivity curves (40) for $\boldsymbol{X}$ a random sample from $\mathsf{N}(0,1)$. Left: The sample mean $\bar{X}_{10}$ (blue) and the sample median $\mathsf{med}(X_1, \ldots, X_{10})$ (orange) from Example 1.2; right: the square-root of the sample variance (blue) and the Eddington's estimator $\widetilde{D}_n$ (orange) from Example 1.3.

2.3 Hampel's theorem

In this section we show that the somewhat complicated notion of qualitative robustness of a sequence of estimators from Definition 3 simplifies substantially if one can assume that all T_n

come from a single statistical functional. We will show that for $\{T_n\}_{n=1}^{\infty}$ to be qualitatively robust at $P \in \mathcal{P}(\mathbb{R})$, it is necessary only that T is continuous in the weak topology at P. The main reason why the standard (weak) continuity of T is enough to obtain the asymptotic equicontinuity result in Definition 3 is the Glivenko-Cantelli theorem.

Theorem 6 (Glivenko-Cantelli)**.** *For $P \in \mathcal{P}(\mathbb{R})$ denote by $P_n = P_n(\omega) \in \mathcal{P}(\mathbb{R})$ an empirical measure (5) based on a random sample $X_1, \ldots, X_n$ from P. Then for every $\varepsilon > 0$ and $\delta > 0$ there exists $n_0 \geq 1$ such that for all $n \geq n_0$*

$$\sup_{P \in \mathcal{P}(\mathbb{R})} \mathsf{P}_P\left(\{\omega \in \Omega \colon d_K(P, P_n(\omega)) \leq \delta\}\right) \geq 1 - \varepsilon. \tag{41}$$

Proof. The standard Glivenko-Cantelli theorem without the supremum in (41) of the form

$$\mathsf{P}\left(\{\omega \in \Omega \colon d_K(P, P_n(\omega)) \to 0 \text{ as } n \to \infty\}\right) = 1$$

is similar to, e.g., the Varadarajan Theorem 4. See also [15, Theorem 3.3(v)]. Its uniform extension from (41) can be found in [28, Section 2.8.1]. □

We are now ready to prove our main result on qualitative robustness.

Theorem 7 (Hampel)**.** *Assume that the sequence of estimators $\{T_n\}_{n=1}^{\infty}$ is represented by a single statistical functional $T \colon \mathcal{P}(\mathbb{R}) \to \mathbb{R}^p$. If T is continuous at $P_0 \in \mathcal{P}(\mathbb{R})$, then $\{T_n\}_{n=1}^{\infty}$ is qualitatively robust at P_0.*

Proof. Suppose that T is continuous at P_0. We take $\varepsilon > 0$ and want to prove (12), i.e. that there exists $\delta > 0$ and $n_0 \geq 1$ such that for all $P \in \mathcal{P}(\mathcal{X})$ and $n \geq n_0$ we have

$$d_L(P_0, P) \leq \delta \quad \text{implies} \quad d_P(\mathcal{L}_{P_0}(T_n), \mathcal{L}_P(T_n)) \leq \varepsilon. \tag{42}$$

By the triangle inequality for d_P from Theorem 2 we have

$$d_P(\mathcal{L}_{P_0}(T_n), \mathcal{L}_P(T_n)) \leq d_P(\mathcal{L}_{P_0}(T_n), \delta_{T(P_0)}) + d_P(\delta_{T(P_0)}, \mathcal{L}_P(T_n)).$$

Here, $\delta_{T(P_0)}$ is the Dirac measure (6) concentrated at the point $T(P_0) \in \mathbb{R}^p$. To prove (42), it is enough to show that there exist $\delta > 0$ and $n_0 \geq 1$ such that $d_L(P_0, P) \leq \delta$ implies

$$d_P(\delta_{T(P_0)}, \mathcal{L}_P(T_n)) = d_P(\delta_{T(P_0)}, \mathcal{L}_P(T(P_n))) \leq \varepsilon/2 \quad \text{for all } n \geq n_0. \tag{43}$$

Here, of course, $\mathcal{L}_P(T(P_n))$ stands for the law of $T(P_n)$ with $P_n = P_n(\omega)$ an empirical measure drawn from P. We obtain the bound (43) on the Prokhorov distance by showing

$$\mathsf{P}_P\left(d(T(P_0), T(P_n)) \leq \varepsilon/2\right) \geq 1 - \varepsilon/2. \tag{44}$$

Then for any fixed random element $\omega \in \Omega$ we can write for any $A \subseteq \mathbb{R}^p$ Borel that $T(P_0) \in A$ implies that either $T(P_n(\omega)) \in A^{\varepsilon/2}$ or $d(T(P_0), T(P_n(\omega))) > \varepsilon/2$, meaning that, using (44), we get

$$\begin{aligned}
\mathsf{P}_P\left(T(P_0) \in A\right) &\leq \mathsf{P}_P\left(T(P_n) \in A^{\varepsilon/2} \text{ or } d(T(P_0), T(P_n)) > \varepsilon/2\right) \\
&\leq \mathsf{P}_P\left(T(P_n) \in A^{\varepsilon/2}\right) + \mathsf{P}_P\left(d(T(P_0), T(P_n)) > \varepsilon/2\right) \\
&\leq \mathsf{P}_P\left(T(P_n) \in A^{\varepsilon/2}\right) + \varepsilon/2,
\end{aligned}$$

which is precisely the definition of the Prokhorov distance in (43).

Now we want to show (44). Because T is continuous at P_0, we know that for our $\varepsilon > 0$ there exists $\delta > 0$ such that $d_L(P_0, P) \leq 2\,\delta$ implies $d(T(P_0), T(P)) \leq \varepsilon/2$. We get

$$\mathsf{P}_P\left(d_L(P_0, P_n) \leq 2\,\delta\right) \leq \mathsf{P}_P\left(d(T(P_0), T(P_n)) \leq \varepsilon/2\right),$$

and we obtain (44) by showing that

$$\mathsf{P}_P\left(d_L(P_0, P_n) \leq 2\,\delta\right) \geq 1 - \varepsilon/2.$$

For any $P \in \mathcal{P}(\mathbb{R})$ we have $d_L(P_0, P_n) \leq d_L(P_0, P) + d_L(P_n, P)$ by Theorem 1. If also $d_L(P_0, P) \leq \delta$, then we can write

$$\mathsf{P}_P\left(d_L(P_0, P_n) \leq 2\,\delta\right) \geq \mathsf{P}_P\left(d_L(P_0, P) + d_L(P_n, P) \leq 2\,\delta\right) \geq \mathsf{P}_P\left(d_L(P_n, P) \leq \delta\right),$$

and the Glivenko-Cantelli Theorem 6 now gives that for our $\varepsilon > 0$ and $\delta > 0$ there exists $n_0 \geq 1$ such that for all $n \geq n_0$ we can write

$$\mathsf{P}_P\left(d_L(P_n, P) \leq \delta\right) \geq \mathsf{P}_P\left(d_K(P_n, P) \leq \delta\right) \geq 1 - \varepsilon/2,$$

uniformly in all $P \in \mathcal{P}(\mathbb{R})$. In the last formula, we used the inequality between d_L and d_K from Lemma 1. □

In [12, Theorem 2.21] also a converse to Theorem 7 is shown: If the estimators T_n and consistent in a neighbourhood of P_0 and $\{T_n\}_{n=1}^{\infty}$ is qualitatively robust, then T already must be continuous. It is, of course, possible to prove more general versions of Theorem 7, not only for T defined on $\mathcal{P}(\mathbb{R})$ but more generally on $\mathcal{P}(\mathcal{X})$; one such statement can be found in [18].

3 Families of estimators and their robustness

With respect to their statistical properties, there are three major types of statistical estimators that have been extensively covered in robust statistics. They are:

- The *M-estimators* obtained by minimisation of an objective function; a prime example is the collection of estimators based on maximum likelihood;

- The *L-estimators* formed as linear combinations of order statistics; and
- The *R-estimators* based on ranks, and derived from the nonparametric theory of rank tests.

To a certain extent, each of these classes can be studied together. We will see that each collection contains both robust and non-robust estimators. We will derive the basic robustness characteristics for these estimators, and state properties that pertain to their robustness and optimality.

In the present section we deal with the situation of $p = 1$, that is, a one-dimensional parameter of interest without nuisance. Estimation in more complex models will be considered separately.

In addition to M, L, and R-estimators, many other classes of estimators have been proposed in the literature; for a brief overview see [10, Section 2.3d]. We will, however, see that our three seminal groups of estimators already do cover a majority of commonly used procedures.

3.1 M-estimators: Minimising an objective function

We are given a random sample $X_1, \ldots, X_n$ from a distribution $P \in \mathcal{P}(\mathcal{X})$, which depends on a parameter of interest $\theta \in \Theta \subseteq \mathbb{R}$.

The collection of M-estimators is obtained by minimising an objective function of the form

$$\sum_{i=1}^{n} \rho(X_i, T_n) \tag{45}$$

in the argument $T_n \in \Theta$, that is

$$T_n(X_1, \ldots, X_n) = \arg\min_{t \in \Theta} \sum_{i=1}^{n} \rho(X_i, t). \tag{46}$$

In these formulas, $\rho\colon \mathcal{X} \times \Theta \to \mathbb{R}$ can be any function; if ρ has a partial derivative in its second argument

$$\psi(x, t) = \frac{\partial}{\partial t} \rho(x, t) \quad \text{for all } x \in \mathcal{X},$$

one can also define an M-estimator as a solution (in $T_n \in \Theta$) to the equation

$$\sum_{i=1}^{n} \psi(X_i, T_n) = 0. \tag{47}$$

An estimator of the type (47) is in [20] called a Z-estimator. Of course, the problems of minimising (45) and solving (47) are not always equivalent. It is, however, convenient to consider these two types of estimators together.

Example 3.1. Take a parametric model $\mathcal{F} = \{P_\theta \colon \theta \in \Theta\} \subset \mathcal{P}(\mathcal{X})$ with $\Theta \subseteq \mathbb{R}$ such that each P_θ has a density $f(\cdot, \theta)$ with respect to a given σ-finite measure μ in $\mathcal{X}$. Setting

$$\rho(x, \theta) = -\log(f(x, \theta)),$$

or, provided that $f(x, \theta)$ is differentiable in θ,

$$\psi(x, \theta) = \frac{\partial}{\partial \theta} \log(f(x, \theta)),$$

we obtain the maximum likelihood estimators as a special collection of M-estimators. In particular, the sample mean $\bar{X}_n$ is an M-estimator in the model $\{\mathsf{N}(\theta, 1) \colon \theta \in \mathbb{R}\}$. △

Example 3.2. For $\mathcal{X} = \mathbb{R}$, each sample α-quantile with $\alpha \in (0, 1)$ is an M-estimator if one considers

$$\rho_\alpha(x, t) = \xi_\alpha(x - t)$$

with the "check function"

$$\xi_\alpha(u) = u\,(\alpha - \mathbb{I}\,(u < 0)), \tag{48}$$

see Figure 6 and [15, Lemma 3.4]. The function ξ_α fails to be differentiable at $u = 0$, but elsewhere its derivative is

$$\psi_\alpha(x, t) = \frac{\partial}{\partial t} \rho_\alpha(x, t) = \begin{cases} -\alpha & \text{for } x > t, \\ 1 - \alpha & \text{for } x < t. \end{cases}$$

For $\alpha = 1/2$ and the sample median we obtain $\xi_{1/2}(u) = |u|\,/2$. △

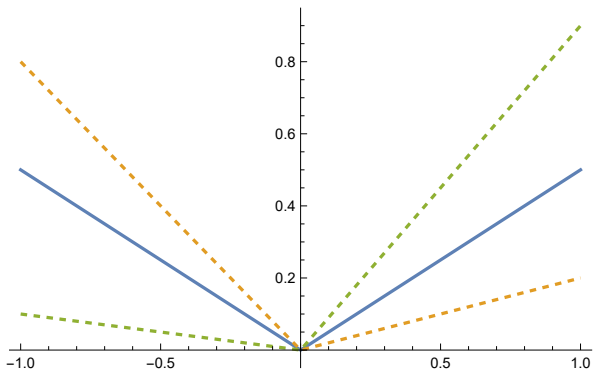

Figure 6: The "check function" from (48) for the α-quantile with $\alpha = 1/2$ (blue), $\alpha = 1/5$ (orange), and $\alpha = 9/10$ (green).

M-estimators can be easily represented by a statistical functional. A natural population analogue of (46) is the functional

$$T(P) = \arg\min_{t \in \Theta} \int_{\mathcal{X}} \rho(x, t)\, \mathrm{d}\, P(x).$$

For an empirical measure $P_n \in \mathcal{P}(\mathcal{X})$ we recover (46) with a factor $1/n$ in front of the objective function ρ, but this does not affect the estimator T_n. A statistical estimator representing the Z-estimators from (47) is $T(P)$ defined implicitly as a solution to

$$\int_{\mathcal{X}} \psi(x, T(P)) \, \mathrm{d}\, P(x) = 0. \tag{49}$$

Example 3.3. Returning to Example 3.2, a statistical functional representing the sample median is

$$\widetilde{T}(P) = \arg\min_{t\in\mathbb{R}} \int_{\mathbb{R}} |x - t| \, \mathrm{d}\, P(x). \tag{50}$$

The objective function of $\widetilde{T}$ is, however, finite only if $P \in \mathcal{P}_1(\mathbb{R})$. Using (50), the median could not be defined if P does not possess the first moment. The problem (50) is however equivalent to

$$\arg\min_{t\in\mathbb{R}} \int_{\mathbb{R}} (|x - t| - |x|) \, \mathrm{d}\, P(x) = \arg\min_{t\in\mathbb{R}} \int_{\mathbb{R}} |x - t| \, \mathrm{d}\, P(x) - \int_{\mathbb{R}} |x| \, \mathrm{d}\, P(x)$$

whenever $P \in \mathcal{P}_1(\mathbb{R})$. At the same time, for any $P \in \mathcal{P}(\mathbb{R})$,

$$0 \leq \left| \int_{\mathbb{R}} (|x - t| - |x|) \, \mathrm{d}\, P(x) \right| \leq \int_{\mathbb{R}} ||x - t| - |x|| \, \mathrm{d}\, P(x) \leq \int_{\mathbb{R}} |t| \, \mathrm{d}\, P(x) = |t|,$$

where we used Jensen's inequality and the reverse triangle inequality. This shows that if we re-define the statistical functional (50) to

$$T(P) = \arg\min_{t\in\mathbb{R}} \int_{\mathbb{R}} (|x - t| - |x|) \, \mathrm{d}\, P(x), \tag{51}$$

we obtain a (different) statistical functional representing the sample median, this time well-defined for all $P \in \mathcal{P}(\mathbb{R})$. Interestingly, we see that $\widetilde{T}$ from (50) fails to be Fisher consistent for θ the median of $P \in \mathcal{P}(\mathbb{R})$, but T from (51) is Fisher consistent. △

We saw in Example 3.2 that the sample median (or empirical α-quantiles) can be represented as M-estimators. They can also be represented as Z-estimators (47).

Example 3.4. The problem with the objective function $\rho_\alpha(x, t) = \xi_\alpha(x - t)$ from (48) is that it fails to be differentiable at $t = x$. Consider, for simplicity, only the case of the median $\alpha = 1/2$. We have

$$\frac{\partial}{\partial t}\rho_{1/2}(x, t) = \begin{cases} -1/2 & \text{for } t < x, \\ 1/2 & \text{for } t > x, \end{cases}$$

and the derivative does not exist at $t = x$. This suggests that one could take $\psi(x, t) = -\operatorname{sign}(x - t)$ in (47) to get the median. Selecting this function, we solve in (49) the equation

$$\begin{aligned} 0 = -\int_{\mathbb{R}} \operatorname{sign}(x - t) \, \mathrm{d}\, P(x) &= \int_{(-\infty,t)} 1 \, \mathrm{d}\, P(x) - \int_{(t,\infty)} 1 \, \mathrm{d}\, P(x) \\ &= P((-\infty, t)) - P((t, \infty)), \end{aligned}$$

which gives $T(P)$ such that

$$P\left((-\infty, T(P))\right) = P\left((T(P), \infty)\right). \tag{52}$$

If $P(\{T(P)\}) = 0$, the previous formula defines t as the median of P. This is, however, not necessarily true if P can put positive mass on $T(P)$. Take, for example, $P \in \mathcal{P}(\mathbb{R})$ supported in three points $P(\{-1\}) = 0.4$, $P(\{0\}) = 0.4$, and $P(\{1\}) = 0.3$. Certainly, the median of P is at $t = 0$, but (52) is not valid. In fact, it is easy to see that no point $t \in \mathbb{R}$ satisfies (49) for our P and $\psi(x, t) = \operatorname{sign}(x - t)$, so the corresponding Z-estimator does not exist.

At least for $P \in \mathcal{P}(\mathbb{R})$ with continuous distribution function, the median (or more generally α-quantiles) can be represented as Z-estimators with

$$\psi_\alpha(x, t) = \begin{cases} 1 - \alpha & \text{for } x \leq t, \\ -\alpha & \text{for } x > t. \end{cases} \tag{53}$$

△

3.1.1 Influence function of M-estimators

The computation of the influence function of M-estimators is straightforward. In the statement of the following theorem, we do not specifically give a list of the needed regularity conditions. They are all quite mild, and obviously follow from the consecutive proof. They all involve statements such as the possibility to interchange derivatives and integrals, the fact that all denominators must be non-zero, or the existence of appropriate derivatives.

Theorem 8. *Under mild regularity conditions, the statistical functional T defined by (49) has at $P \in \mathcal{P}(\mathcal{X})$ a directional (Gâteaux) derivative (28) of the form*

$$L_P \colon Q \mapsto -\frac{\int_{\mathcal{X}} \psi(x, T(P)) \,\mathrm{d}\, Q(x)}{\int_{\mathcal{X}} \psi'(x, T(P)) \,\mathrm{d}\, P(x)},$$

where

$$\psi'(x, s) = \frac{\partial}{\partial s} \psi(x, s). \tag{54}$$

The influence function of T is

$$\mathsf{IF}(x, P, T) = -\frac{\psi(x, T(P))}{\int_{\mathcal{X}} \psi'(y, T(P)) \,\mathrm{d}\, P(y)} \quad \text{for } x \in \mathcal{X}.$$

Proof. Take $Q \in \mathcal{P}(\mathcal{X})$ and $P_t = (1-t)\,P + t\,Q$. We first differentiate the defining formula (49)

applied to P_t in $t \in (0,1)$ to get

$$\begin{aligned}
0 &= \frac{\partial}{\partial t}\left(\int_{\mathcal{X}} \psi(x, T(P_t))\,\mathrm{d}\,P_t(x)\right) \\
&= \frac{\partial}{\partial t}\left((1-t)\int_{\mathcal{X}} \psi(x, T(P_t))\,\mathrm{d}\,P(x) + t\int_{\mathcal{X}} \psi(x, T(P_t))\,\mathrm{d}\,Q(x)\right) \\
&= -\int_{\mathcal{X}} \psi(x, T(P_t))\,\mathrm{d}\,P(x) + (1-t)\left(\frac{\partial}{\partial t}T(P_t)\right)\int_{\mathcal{X}} \frac{\partial}{\partial s}\left[\psi(x,s)\right]_{s=T(P_t)}\,\mathrm{d}\,P(x) \\
&\quad + \int_{\mathcal{X}} \psi(x, T(P_t))\,\mathrm{d}\,Q(x) + t\left(\frac{\partial}{\partial t}T(P_t)\right)\int_{\mathcal{X}} \frac{\partial}{\partial s}\left[\psi(x,s)\right]_{s=T(P_t)}\,\mathrm{d}\,Q(x).
\end{aligned} \tag{55}$$

Here and elsewhere in the text, we write

$$\frac{\partial}{\partial y}\left[f(x,y)\right]_{y=0}$$

for the partial derivative of the function $f(x,y)$ in y evaluated at the point $y=0$. To obtain a directional (Gâteaux) derivative of T we need to take $t \to 0$ in the previous formula. That gives

$$0 = \int_{\mathcal{X}} \psi(x, T(P))\,\mathrm{d}\,(Q(x) - P(x)) + \left(\frac{\partial}{\partial t}\left[T(P_t)\right]_{t=0}\right)\int_{\mathcal{X}} \frac{\partial}{\partial s}\left[\psi(x,s)\right]_{s=T(P)}\,\mathrm{d}\,P(x),$$

and

$$\frac{\partial}{\partial t}\left[T(P_t)\right]_{t=0} = \frac{\int_{\mathcal{X}} \psi(x, T(P))\,\mathrm{d}\,(P(x) - Q(x))}{\int_{\mathcal{X}} \frac{\partial}{\partial s}\left[\psi(x,s)\right]_{s=T(P)}\,\mathrm{d}\,P(x)}.$$

This expression still simplifies. By (49) we know that

$$\int_{\mathcal{X}} \psi(x, T(P))\,\mathrm{d}\,P(x) = 0,$$

and using (54), we reduce the previous expression to

$$\frac{\partial}{\partial t}\left[T(P_t)\right]_{t=0} = -\frac{\int_{\mathcal{X}} \psi(x, T(P))\,\mathrm{d}\,Q(x)}{\int_{\mathcal{X}} \psi'(x, T(P))\,\mathrm{d}\,P(x)}.$$

This gives the expression for the Gâteaux derivative of T in direction Q. The influence function (31) is obtained by taking $Q = \delta_x$ for $x \in \mathcal{X}$

$$\mathsf{IF}(x, P, T) = -\frac{\psi(x, T(P))}{\int_{\mathcal{X}} \psi'(y, T(P))\,\mathrm{d}\,P(y)}.$$

□

As an illustration, we compute the influence function of the maximum likelihood estimators introduced in Example 3.1.

Example 3.5. For the maximum likelihood estimators in a parametric model $\mathcal{F} = \{P_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$ we have

$$\psi(x, \theta) = \frac{\partial}{\partial \theta} \log(f(x, \theta)).$$

First note that under the standard regularity conditions on maximum likelihood estimation [19, Theorem 23], the M-functional corresponding to the maximum likelihood estimator is Fisher consistent for θ. We saw this already in Example 1.4, but it can also be seen as follows. Writing

$$f'(x, \theta) = \frac{\partial}{\partial \theta} f(x, \theta)$$

and $M = \{x \in \mathcal{X} : f(x, \theta) > 0\}$ for the common support of all densities $f(x, \theta)$ with respect to a σ-finite measure μ on $\mathcal{X}$, we have

$$\begin{aligned} \int_{\mathcal{X}} \psi(x, \theta) \, \mathrm{d} \, P_\theta(x) &= \int_M \psi(x, \theta) \, \mathrm{d} \, P_\theta(x) = \int_M \frac{f'(x, \theta)}{f(x, \theta)} f(x, \theta) \, \mathrm{d} \, \mu(x) \\ &= \int_M f'(x, \theta) \, \mathrm{d} \, \mu(x) = 0, \end{aligned}$$

the last equality following by the requirement from [19, Definition 4]. Thus $\theta = T(P_\theta)$ solves (49), and T is Fisher consistent.

For the influence function of the induced functional T we have, by Theorem 8,

$$\begin{aligned} \mathsf{IF}(x, P_\theta, T) &= -\frac{\psi(x, T(P_\theta))}{\int_{\mathcal{X}} \frac{\partial^2}{\partial \theta^2} \log(f(y, T(P_\theta)) \, \mathrm{d} \, P_\theta(y)} \\ &= -\frac{\psi(x, \theta)}{\int_{\mathcal{X}} \frac{\partial^2}{\partial \theta^2} \log(f(y, \theta)) \, \mathrm{d} \, P_\theta(y)} = \frac{\psi(x, \theta)}{J(\theta)}, \end{aligned}$$

where $J(\theta)$ is the Fisher information in θ; we used [19, Theorem 1]. We see that the robustness of a maximum likelihood estimator depends on the score function $\psi(x, \theta)$; if the score function is bounded, the gross error sensitivity of a maximum likelihood estimator is bounded too.

The asymptotic normality result from Theorem 5 now gives

$$\sqrt{n}(T(P_n) - T(P_\theta)) \xrightarrow[n \to \infty]{\mathrm{d}} \mathsf{N}(0, A(P_\theta, T)),$$

where $P_n \in \mathcal{P}(\mathcal{X})$ is the empirical distribution sampled from P_θ, and

$$A(P_\theta, T) = \int_{\mathcal{X}} (\mathsf{IF}(x, P_\theta, T))^2 \, \mathrm{d} \, P_\theta(x) = \frac{1}{(J(\theta))^2} \int_{\mathcal{X}} \left(\frac{f'(x, \theta)}{f(x, \theta)} \right)^2 \mathrm{d} \, P_\theta(x) = \frac{1}{J(\theta)}. \tag{56}$$

This is in accordance with [19, Theorem 23]. △

Example 3.6. The function $\psi_\alpha(x, t)$ for the α-quantile from Example 3.4 fails to be differentiable at $x = t$. Thus, one cannot apply Theorem 8 immediately. Nevertheless, returning

to the proof of Theorem 8, we can see that the differentiability of ψ was used only for convenience. Modifying the proof slightly, also for our function ψ_α, we can obtain the influence function. Indeed, in formula (55), we used that the partial derivative $\partial/(\partial t)$ can be exchanged with the integral. Not doing this simplification, we would obtain in the right-hand side in (55) instead of

$$\int_{\mathcal{X}} \frac{\partial}{\partial s} \left[\psi(x,s)\right]_{s=T(P_t)} \mathrm{d}\, P(x)$$

the expression

$$\frac{\partial}{\partial s} \left[\int_{\mathcal{X}} \psi(x,s)\, \mathrm{d}\, P(x)\right]_{s=T(P_t)}.$$

The latter expression can be evaluated for $\psi(x,t) = \psi_\alpha(x,t)$ from (53). We get, for P with a continuous distribution function F,

$$\int_{\mathbb{R}} \psi_\alpha(x,t)\, \mathrm{d}\, P(x) = (1-\alpha) \int_{-\infty}^{t} 1\, \mathrm{d}\, P(x) - \alpha \int_{t}^{\infty} 1\, \mathrm{d}\, P(x) = (1-\alpha)F(t) - \alpha(1-F(t)).$$

Taking a derivative of this in t and writing f for the density of P with respect to the Lebesgue measure in $\mathbb{R}$ we get

$$\frac{\partial}{\partial t} \left[\int_{\mathbb{R}} \psi_\alpha(x,t)\, \mathrm{d}\, P(x)\right]_{t=T_\alpha(P)} = f(T_\alpha(P)) = f(F^{-1}(\alpha)).$$

Plugging this expression into the proof of Theorem 8, we obtain the influence function of the α-quantile T_α of P

$$\mathsf{IF}(x, T_\alpha, P) = -\frac{\psi_\alpha(x, F^{-1}(\alpha))}{f(F^{-1}(\alpha))} = \begin{cases} \frac{\alpha-1}{f(F^{-1}(\alpha))} & \text{if } x \le F^{-1}(\alpha), \\ \frac{\alpha}{f(F^{-1}(\alpha))} & \text{if } x > F^{-1}(\alpha), \end{cases} \quad \text{for } x \in \mathbb{R}. \tag{57}$$

The discontinuity of the influence function at $x = F^{-1}(\alpha)$ makes sense — if we contaminate the distribution F by a point on the right in $\mathbb{R}$ from $F^{-1}(\alpha)$, we perturb the α-quantile to the positive side, which corresponds to the positive influence with numerator $\alpha > 0$. If we add a point x smaller than $F^{-1}(\alpha)$, we increase the distribution function in a neighbourhood of $F^{-1}(\alpha)$, meaning that the new α-quantile will be lower. This corresponds to the negative influence and numerator $\alpha - 1 < 0$. Note that the influence function of the α-quantile functional can also be computed directly, without using Theorem 8. That is performed in [12, Section 3.3.1].

We found that the influence function of the α-quantile T_α is (57). If T is Fréchet differentiable, Theorem 5 would give for T_α that

$$\sqrt{n}\,(T_\alpha(P_n) - T_\alpha(P)) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}(0, A(P, T_\alpha(P))),$$

where

$$\begin{aligned}
A(P, T_\alpha) &= \int_{\mathbb{R}} (\mathsf{IF}(x, P, T_\alpha))^2 \, \mathrm{d}\, P(x) \\
&= \int_{-\infty}^{F^{-1}(\alpha)} \left(\frac{\alpha - 1}{f(F^{-1}(\alpha))} \right)^2 \mathrm{d}\, P(x) + \int_{F^{-1}(\alpha)}^{\infty} \left(\frac{\alpha}{f(F^{-1}(\alpha))} \right)^2 \mathrm{d}\, P(x) \\
&= \frac{1}{(f(F^{-1}(\alpha)))^2} \left((\alpha - 1)^2 \int_{-\infty}^{F^{-1}(\alpha)} 1 \, \mathrm{d}\, P(x) + \alpha^2 \int_{F^{-1}(\alpha)}^{\infty} 1 \, \mathrm{d}\, P(x) \right) \\
&= \frac{1}{(f(F^{-1}(\alpha)))^2} \left((\alpha - 1)^2 F(F^{-1}(\alpha)) + \alpha^2 (1 - F(F^{-1}(\alpha))) \right) \\
&= \frac{1}{(f(F^{-1}(\alpha)))^2} \left((\alpha - 1)^2 \alpha + \alpha^2 (1 - \alpha) \right) = \frac{\alpha(1 - \alpha)}{(f(F^{-1}(\alpha)))^2}.
\end{aligned}$$

This result is true; it can be shown for any $P \in \mathcal{P}(\mathbb{R})$ that has a density f (with respect to the Lebesgue measure) in a neighbourhood of $F^{-1}(\alpha)$ such that f is positive and continuous at $F^{-1}(\alpha)$, see [25, Corollary B in Section 2.3.3]. In particular, for $\alpha = 1/2$ and the median functional, we obtain the asymptotic result (1) that we used in our motivating Example 1.2. △

3.1.2 Distributional properties of M-estimators

A relatively simple situation arises in the common setup when the function $\psi(x, t)$ in (49) is monotone in t. We assume that ψ is non-increasing in t for each $x \in \mathcal{X}$, and takes both positive and negative values. We do not assume the continuity of ψ. Each $\psi(X_i, t)$ in (47) is then a non-increasing function in t, and so is $\sum_{i=1}^n \psi(X_i, t)$. Naturally, the estimator T_n solving (47) can then be taken as (any) number in the interval $[T_n^*, T_n^{**}] \subset \mathbb{R}$, where

$$\begin{aligned}
T_n^* &= \sup \left\{ t \in \mathbb{R} \colon \sum_{i=1}^n \psi(X_i, t) > 0 \right\}, \\
T_n^{**} &= \inf \left\{ t \in \mathbb{R} \colon \sum_{i=1}^n \psi(X_i, t) < 0 \right\}.
\end{aligned} \tag{58}$$

Strictly speaking, since ψ can be discontinuous, it is not necessarily true that T_n then solves (47), see Figure 7. However, because $\sum_{i=1}^n \psi(X_i, t)$ is a non-increasing function, it is natural to consider this minor modification to (49) in what follows. We thus, for ψ monotone as above, define the statistical functional T implicitly as any value in the interval $[T^*(P), T^{**}(P)]$, where

$$\begin{aligned}
T^*(P) &= \sup \left\{ t \in \mathbb{R} \colon \int_{\mathcal{X}} \psi(x, t) \, \mathrm{d}\, P(x) > 0 \right\}, \\
T^{**}(P) &= \inf \left\{ t \in \mathbb{R} \colon \int_{\mathcal{X}} \psi(x, t) \, \mathrm{d}\, P(x) < 0 \right\}.
\end{aligned} \tag{59}$$

Certainly, for any $P \in \mathcal{P}(\mathcal{X})$ we have $-\infty < T^*(P) \leq T^{**}(P) < \infty$, and our choice of T_n^* and T_n^{**} is consistent with this.

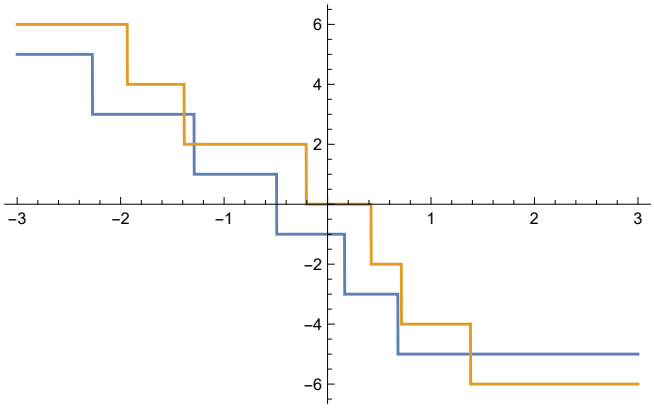

Figure 7: The function $t \mapsto \sum_{i=1}^n \psi(X_i, t)$ for $\psi(x,t) = \text{sign}(x-t)$ with $n = 5$ (in blue) and $n = 6$ (in orange). This corresponds to T the median from Example 3.4. The unique median for $n = 5$ is $\theta \approx -0.49$, but because ψ is discontinuous in θ, this value does not solve (47). For $n = 6$, we have $T_n^* \approx -0.21$ and $T_n^{**} \approx 0.42$; the whole interval $[T_n^*, T_n^{**}]$ is a solution to (47).

Immediately from (58) we see that if $T_n^* < t$, then $\sum_{i=1}^n \psi(X_i, t) \leq 0$, which in turn implies $T_n^* \leq t$. We get that, in terms of random events, we can write

$$\begin{aligned} [T_n^* < t] &\subseteq \left[\sum_{i=1}^n \psi(X_i, t) \leq 0\right] \subseteq [T_n^* \leq t]\,, \\ [T_n^{**} < t] &\subseteq \left[\sum_{i=1}^n \psi(X_i, t) < 0\right] \subseteq [T_n^{**} \leq t]\,, \end{aligned} \tag{60}$$

for any $t \in \mathbb{R}$. This gives

$$\begin{aligned} \mathsf{P}\left(T_n^* < t\right) &\leq \mathsf{P}\left(\sum_{i=1}^n \psi(X_i, t) \leq 0\right) \leq \mathsf{P}\left(T_n^* \leq t\right), \\ \mathsf{P}\left(T_n^{**} < t\right) &\leq \mathsf{P}\left(\sum_{i=1}^n \psi(X_i, t) < 0\right) \leq \mathsf{P}\left(T_n^{**} \leq t\right). \end{aligned}$$

At each continuity point $t \in \mathbb{R}$, we can express the distribution functions of both T_n^* and T_n^{**} directly in terms of the distribution of $\sum_{i=1}^n \psi(X_i, t)$.

The easiest way to select a unique estimator T_n from the interval $[T_n^*, T_n^{**}]$ is to take T_n at random, either T_n^* or T_n^{**} with equal probabilities. This gives a simple expression for the exact distribution of T_n given by

$$\mathsf{P}\left(T_n \leq t\right) = \frac{1}{2}\mathsf{P}\left(\sum_{i=1}^n \psi(X_i, t) \leq 0\right) + \frac{1}{2}\mathsf{P}\left(\sum_{i=1}^n \psi(X_i, t) < 0\right)$$

for any $t \in \mathbb{R}$ where this distribution function is continuous. Interestingly, this gives the exact distribution of T_n only in terms of the convolution powers of the random variable $\psi(X_1, t)$. The latter convolution powers are, however, usually not very easy to compute.

From our analysis, we obtain a very simple sample consistency result for M-estimators.

Theorem 9. *Suppose that $\psi(x,t)$ is non-increasing and taking both positive and negative values in $t \in \mathbb{R}$ for each $x \in \mathcal{X}$. Let $P \in \mathcal{P}(\mathcal{X})$ be such that $T^*(P) = T^{**}(P)$ in (59). Then all T_n^*, T_n^{**}, and T_n converge to $T(P) = T^*(P) = T^{**}(P)$ in probability as $n \to \infty$.*

Proof. We prove the result for T_n^*; the other two estimators are treated analogously. Take any $\varepsilon > 0$ and bound, using (58) and (60),

$$\begin{aligned} &\mathsf{P}\left(|T_n^* - T(P)| \geq \varepsilon\right) \leq \mathsf{P}\left(T_n^* \geq T(P) + \varepsilon\right) + \mathsf{P}\left(T_n^* < T(P) - \varepsilon/2\right) \\ &\quad \leq \mathsf{P}\left(\sum_{i=1}^n \psi(X_i, T(P) + \varepsilon/2) > 0\right) + \mathsf{P}\left(\sum_{i=1}^n \psi(X_i, T(P) - \varepsilon/2) \leq 0\right) \\ &\quad = \mathsf{P}\left(\frac{1}{n}\sum_{i=1}^n \psi(X_i, T(P) + \varepsilon/2) > 0\right) + \mathsf{P}\left(\frac{1}{n}\sum_{i=1}^n \psi(X_i, T(P) - \varepsilon/2) \leq 0\right). \end{aligned} \tag{61}$$

On the right-hand side, we have two averages of independent identically distributed random variables with expectations

$$\mathsf{E}\,\psi(X_1, T(P) + \varepsilon/2) = \int_{\mathbb{R}} \psi(x, T(P) + \varepsilon/2)\,\mathrm{d}\,P(x) < 0$$

and

$$\mathsf{E}\,\psi(X_1, T(P) - \varepsilon/2) = \int_{\mathbb{R}} \psi(x, T(P) - \varepsilon/2)\,\mathrm{d}\,P(x) > 0,$$

respectively. We used the assumption that $T(P) = T^*(P) = T^{**}(P)$. The law of large numbers thus gives that the right-hand side of (61) vanishes as $n \to \infty$. □

If $T(P)$ is not defined uniquely, one cannot expect the sample M-estimators to be consistent.

Example 3.7. Take $\mathcal{X} = \mathbb{R}$ and let $P \in \mathcal{P}(\mathbb{R})$ be a distribution with density $f(x) = \mathbb{I}(x \in [-2,-1] \cup [1,2])/2$. The median $T(P)$ can be any point in the interval between $T^*(P) = -1$ and $T^{**}(P) = 1$. For $X_1, \ldots, X_n$ a random sample from P with n odd, there will be at least $n/2$ observations either in the interval $[-2,-1]$, or in the interval $[1,2]$. In the first case, the sample median T_n is in the interval $[-2,-1]$; in the second case, it falls into $[1,2]$. As $n \to \infty$, both situations occur infinitely many times, almost surely. Thus, the sample median T_n does not converge to any point, almost surely. △

M-estimators can also be proved to be asymptotically normal under reasonably mild assumptions. Such a result can be found in, e.g., [20, Sections 3.2 and 3.4]. In fact, under a

collection of technical assumptions, in [1, 2, 3] it is proved that M-estimators are Fréchet differentiable statistical functionals. In particular, our stochastic representation from Theorem 5 applies, and using Theorem 8 we get

$$\sqrt{n}\,(T(P_n) - T(P)) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}(0, A(P,T))$$

where

$$A(P,T) = \frac{\int_{\mathcal{X}} \psi(x, T(P))^2 \,\mathrm{d}\, P(x)}{\left(\int_{\mathcal{X}} \psi'(y, T(P)) \,\mathrm{d}\, P(y)\right)^2}.$$

3.1.3 Robustness of M-estimators of location

From the expression for the influence function of an M-estimator in Theorem 8 we see that the gross error sensitivity (37) of an M-estimator is

$$\gamma^*(P,T) = \frac{\sup_{x\in\mathcal{X}} |\psi(x, T(P))|}{\left|\int_{\mathcal{X}} \psi'(y, T(P)) \,\mathrm{d}\, P(y)\right|}.$$

It is proportional to the supremum norm of the function $\psi(x, T(P))$. Thus, bounded functions ψ give robust estimators in the sense of having finite gross error sensitivity.

We now explore the maximum bias (15) of M-estimators in the case of a location parameter. Recall that $\mathcal{X} = \mathbb{R}$, $X \sim P \in \mathcal{P}(\mathbb{R})$ and $Y = X + c \sim Q \in \mathcal{P}(\mathbb{R})$ with $c \in \mathbb{R}$, we say that a parameter $T = T(P)$ is a *location parameter* of P if

$$T(Q) = T(P) + c \quad \text{for all } c \in \mathbb{R}. \tag{62}$$

For estimating a location parameter, we thus suppose that $\mathcal{X} = \mathbb{R}$ and $\psi(x,t) = \psi_0(x-t)$ with a monotone non-decreasing function $\psi_0 \colon \mathbb{R} \to \mathbb{R}$. This corresponds to our assumptions from Section 3.1.2, as $\psi(x,t)$ is now non-increasing in t for each $x \in \mathbb{R}$. In the maximum bias (15) we consider the neighbourhoods $\mathcal{P}_\varepsilon(P_0)$ of P_0 in $\mathcal{P}(\mathbb{R})$ induced by the Lévy distance (20).

The function ψ is chosen so that the statistical functional T in (49) is *translation equivariant*, meaning that for any $X \sim P \in \mathcal{P}(\mathbb{R})$ and $Y = X + c \sim Q \in \mathcal{P}(\mathbb{R})$ with $c \in \mathbb{R}$, we have

$$\int_{\mathbb{R}} \psi_0(x-t) \,\mathrm{d}\, P(x) = \int_{\mathbb{R}} \psi_0(x + c - (t+c)) \,\mathrm{d}\, P(x) = \int_{\mathbb{R}} \psi_0(y - (t+c)) \,\mathrm{d}\, Q(y), \tag{63}$$

that is (62) is true for T. Thus, without loss of generality, we can suppose that $T(P_0) = 0$; otherwise, we would just use the distribution of the random variable $X - T(P_0)$ instead of $X \sim P_0$ in our analysis.

The maximum bias of T can be written as

$$b(\varepsilon, P_0, T) = \max\left\{b_+(\varepsilon), -b_-(\varepsilon)\right\}, \tag{64}$$

where

$$\begin{aligned} b_+(\varepsilon) &= \sup\{T(P)\colon d_L(P_0, P) \le \varepsilon\}, \\ b_-(\varepsilon) &= \inf\{T(P)\colon d_L(P_0, P) \le \varepsilon\}. \end{aligned} \tag{65}$$

Denote

$$\lambda(t, P) = \int_{\mathbb{R}} \psi_0(x - t)\,\mathrm{d}\,P(x). \tag{66}$$

As in (59), we see that λ is a non-increasing function in $t \in \mathbb{R}$, and $T(P)$ is a solution to $\lambda(T(P), P) = 0$. Any such solution lies in the interval $[T^*(P), T^{**}(P)]$ from (59). In the argument of $P \in \mathcal{P}(\mathbb{R})$, the function (66) is non-decreasing in the sense of stochastic ordering in $\mathcal{P}(\mathbb{R})$. For that, recall the following defintion.

Definition 15 (stochastic ordering)**.** If F and G are distribution functions of P and $Q \in \mathcal{P}(\mathbb{R})$, respectively, such that $F(x) \le G(x)$ for all $x \in \mathbb{R}$, then we say that P is *stochastically larger* than Q.

Lemma 2. *For any $t \in \mathbb{R}$ is the function (66) non-decreasing in $P \in \mathcal{P}(\mathbb{R})$ in the sense of stochastic ordering.*

Proof. We have distribution functions F and G of P and $Q \in \mathcal{P}(\mathbb{R})$, respectively, such that $F(x) \le G(x)$ for all $x \in \mathbb{R}$. We want to show that $\lambda(t, P) \ge \lambda(t, Q)$. First, note that $\xi(x) = \psi_0(x - t)$ is a non-decreasing function of $x \in \mathbb{R}$. We approximate ξ by a sequence of non-decreasing step functions $\xi_n(x) = b_n + \sum_{j=1}^n c_{j,n}\,\mathbb{I}(x \in (a_{j,n}, \infty))$ for $c_{j,n} \ge 0$, $b_n \in \mathbb{R}$, and $a_{j_n} \in \mathbb{R}$. The term $b_n \in \mathbb{R}$ is included because ξ may also take negative values. For each function ξ_n we have

$$\begin{aligned} \int_{\mathbb{R}} \xi_n(x)\,\mathrm{d}\,P(x) &= b_n + \sum_{j=1}^n c_{j,n} P((a_{j,n}, \infty)) = b_n + \sum_{j=1}^n c_{j,n}(1 - F(a_{j,n})) \\ &\ge b_n + \sum_{j=1}^n c_{j,n}(1 - G(a_{j,n})) = \int_{\mathbb{R}} \xi_n(x)\,\mathrm{d}\,G(x). \end{aligned}$$

Since this is true for each such sequence of functions, taking the appropriate limit $\xi_n \to \xi$ as $n \to \infty$ guarantees that the inequality must be true also for ξ. □

Thanks to Lemma 2, if we find a distribution $P \in \mathcal{P}_\varepsilon(P_0)$ that is stochastically larger than any other $Q \in \mathcal{P}_\varepsilon(P_0)$, the function (66) will be maximised in P for all $t \in \mathbb{R}$. The stochastically largest element in $\mathcal{P}_\varepsilon(P_0)$ is clearly the point-wise minimum function inside the neighbourhood band around the distribution function F_0 of P_0 depicted in Figure 4. To make this function a distribution function, we must also bound it from below by zero; we get the function

$$F_1(x) = \max\{0, F_0(x - \varepsilon) - \varepsilon\} \quad \text{for } x \in \mathbb{R}.$$

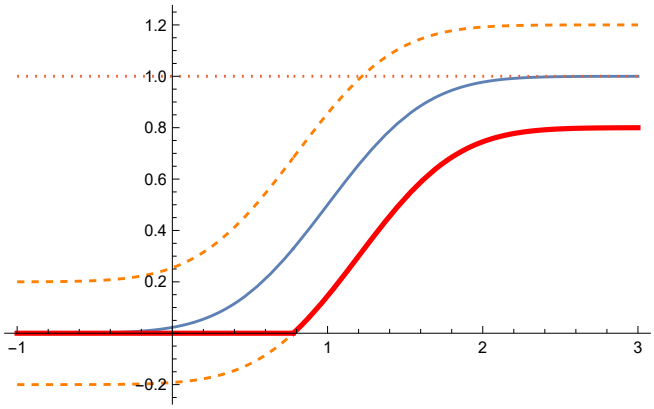

Figure 8: The stochastically largest element of the Lévy neighbourhood of F_0 (blue line) is the function F_1 in thick red. It corresponds to an improper distribution giving mass $\varepsilon > 0$ to a point in positive infinity.

This function is displayed in Figure 8; it can also be written as

$$F_1(x) = \begin{cases} 0 & \text{for } x < x_0 + \varepsilon, \\ F_0(x - \varepsilon) - \varepsilon & \text{for } x \geq x_0 + \varepsilon, \end{cases} \tag{67}$$

where $x_0 = F_0^{-1}(\varepsilon)$ is the ε-quantile of P_0. Function F_1 does not correspond to a proper distribution, as such distribution would give mass ε to a point in the positive infinity. Nevertheless, clearly F_1 can be approximated (from above) by proper distribution functions point-wise, and we can treat F_1 as the limit case for distributions given by

$$F_{1,n}(x) = F_1(x) + \varepsilon\, \mathbb{I}\,(x \geq n)\,. \tag{68}$$

in the neighbourhood $\mathcal{P}_\varepsilon(P_0)$.

Lemma 2 now gives that (interchanging the notation for the (improper) distribution with its distribution function freely)

$$\lambda(t, F) \leq \lambda(t, F_1) = \int_{\mathbb{R}} \psi_0(x - t)\, \mathrm{d}\, F_1(x) = \int_{x_0}^{\infty} \psi_0(x + \varepsilon - t)\, \mathrm{d}\, F_0(x) + \varepsilon\, \psi(\infty), \tag{69}$$

where $\psi_0(\infty) = \lim_{t\to\infty} \psi_0(t)$ and the last expression follows from a change of variables and the properties of the Stieltjes integral.[2] The last result is easiest to see when using an approximation argument, considering instead of the improper distribution function F_1 a

[2]Recall that [7, Note § 3.2], in analogy with the Riemann integral, the Riemann-Stieltjes integral $\int_a^b g(x)\, \mathrm{d}\, F(x)$ can be defined as the limit of sums $\sum_{i=1}^n g(y_i)(F(x_i) - F(x_{i-1}))$ over all partitions $a = x_0 \leq y_1 \leq x_1 \leq y_2 \leq \cdots \leq x_n = b$ as $\max_i(x_i - x_{i-1}) \to 0$ with $n \to \infty$.

sequence of proper distribution functions (68). Indeed

$$\begin{aligned}\lambda(t, F_{1,n}) &= \int_{\mathbb{R}} \psi_0(x-t)\,\mathrm{d}\,\left(F_1(x) + \varepsilon\,\mathbb{I}\,(x \geq n)\right) \\ &= \int_{x_0+\varepsilon}^{\infty} \psi_0(x-t)\,\mathrm{d}\,F_0(x-\varepsilon) + \int_{x_0+\varepsilon}^{\infty} \psi_0(x-t)\,\mathrm{d}\,\left(-\varepsilon + \varepsilon\,\mathbb{I}\,(x \geq n)\right) \\ &= \int_{x_0}^{\infty} \psi_0(s+\varepsilon-t)\,\mathrm{d}\,F_0(s) + \varepsilon\,\psi_0(n-t).\end{aligned} \tag{70}$$

In the last equation, we substituted $s = x - \varepsilon$ and used that the function $-\varepsilon + \varepsilon\,\mathbb{I}\,(x \geq n)$ is constant, except for the single jump of size ε at $x = n$. Taking the limit as $n \to \infty$ gives (69).

The function $t \mapsto \lambda(t, P)$ is maximal for each t when P is F_1. Therefore, $T^{**}(F_1)$ is the maximum positive bias of T in the ε-neighbourhood of P_0, and

$$b_+(\varepsilon) = \inf\left\{t \in \mathbb{R}\colon\, \lambda(t, F_1) < 0\right\}. \tag{71}$$

Analogously, $b_-(\varepsilon)$ is obtained from the "upper envelope" of the band in Figure 8 given by

$$F_2(x) = \min\left\{1, F_0(x+\varepsilon) + \varepsilon\right\} \quad \text{for } x \in \mathbb{R},$$

and

$$b_-(\varepsilon) = \sup\left\{t \in \mathbb{R}\colon\, \lambda(t, F_2) > 0\right\}.$$

Combining our formulas for $b_+(\varepsilon)$ and $b_-(\varepsilon)$ with (64), we obtain the exact expression for the maximum bias of any location M-estimator in $\mathbb{R}$ with a monotone function ψ.

Let us now find the asymptotic breakdown point of T from Definition 5. For simplicity, suppose that F_0 is continuous, strictly increasing and symmetric around the origin (that is, $F_0(x) = 1 - F_0(-x)$ for all $x \in \mathbb{R}$); the general case is completely analogous, yet a bit more cumbersome to write explicitly. From (69) we get that if $\psi_0(\infty) = \infty$, $\lambda(t, F_1) = \infty$ for all $\varepsilon > 0$, and T breaks down immediately. Using our symmetry considerations, the same happens if $\psi_0(-\infty) = -\infty$ for $\psi_0(-\infty) = \lim_{t\to-\infty} \psi_0(t)$. Thus, a necessary condition for a positive asymptotic breakdown point of T is the boundedness of ψ_0. Now, suppose that ψ_0 is bounded. Formulas (69) and (71) give that T breaks down also if $\lambda(t, F_1) \geq 0$ for all $t \in \mathbb{R}$. To prevent this, it must be that as $t \to \infty$, the right-hand side of (69) is negative. That is, it must be that

$$\begin{aligned}0 &> \lim_{t\to\infty} \int_{x_0}^{\infty} \psi_0(x+\varepsilon-t)\,\mathrm{d}\,F_0(x) + \varepsilon\,\psi_0(\infty) \\ &= \int_{x_0}^{\infty} \lim_{t\to\infty} \psi_0(x+\varepsilon-t)\,\mathrm{d}\,F_0(x) + \varepsilon\,\psi_0(\infty) \\ &= \psi_0(-\infty) \int_{x_0}^{\infty} 1\,\mathrm{d}\,F_0(x) + \varepsilon\,\psi_0(\infty) = \psi_0(-\infty)(1 - F_0(x_0)) + \varepsilon\,\psi_0(\infty) \\ &= \psi_0(-\infty)(1-\varepsilon) + \varepsilon\,\psi_0(\infty),\end{aligned}$$

where we used that $x_0 = F_0^{-1}(\varepsilon)$. We get that for $\varepsilon^*(P_0, T)$ to reach $\varepsilon > 0$, it must be that

$$\frac{\varepsilon}{1-\varepsilon} < -\frac{\psi_0(-\infty)}{\psi_0(\infty)}.$$

An analogous analysis of $b_-(\varepsilon)$ gives the second necessary condition

$$\frac{\varepsilon}{1-\varepsilon} < -\frac{\psi_0(\infty)}{\psi_0(-\infty)},$$

which together gives the asymptotic breakdown point of T to be

$$\varepsilon^*(P_0, T) = \frac{\eta}{1+\eta}, \quad \text{where} \quad \eta = \min\left\{-\frac{\psi_0(-\infty)}{\psi_0(\infty)}, -\frac{\psi_0(\infty)}{\psi_0(-\infty)}\right\}. \tag{72}$$

Since $\eta \in (0, 1]$, we have that $\varepsilon^*(P_0, T)$ is maximised if $\eta = 1$ (that is, $-\psi_0(-\infty) = \psi_0(\infty) < \infty$) and takes its maximum value $\varepsilon^*(P_0, T) = 1/2$. If ψ_0 is unbounded, we have $\varepsilon^*(P_0, T) = 0$.

Finally, from (69) we also get for any F in $\mathcal{P}_\varepsilon(F_0)$

$$\int_{\mathbb{R}} \psi_0(x-t)\,\mathrm{d}\,F(x) = \lambda(t, F) \le \int_{x_0}^{\infty} \psi_0(x-(t-\varepsilon))\,\mathrm{d}\,F_0(x) + \varepsilon\,\psi_0(\infty).$$

Because ψ_0 is non-decreasing, we have using $x_0 = F_0^{-1}(\varepsilon)$ that we can write

$$\int_{-\infty}^{x_0} \psi_0(-\infty)\,\mathrm{d}\,F_0(x) = \psi_0(-\infty)\int_{-\infty}^{x_0} 1\,\mathrm{d}\,F_0(x) = \varepsilon\,\psi_0(-\infty) \le \int_{-\infty}^{x_0} \psi_0(x-(t-\varepsilon))\,\mathrm{d}\,F_0(x).$$

Putting the last two formulas together, we obtain

$$\lambda(t, F) \le \int_{\mathbb{R}} \psi_0(x-(t-\varepsilon))\,\mathrm{d}\,F_0(x) - \varepsilon\,\psi_0(-\infty) + \varepsilon\,\psi_0(\infty) = \lambda(t-\varepsilon, F_0) + \varepsilon\,\|\psi_0\|,$$

where we denoted $\|\psi_0\| = \psi_0(\infty) - \psi_0(-\infty)$. An analogous formula holds true also with the other inequality, giving for any $F \in \mathcal{P}_\varepsilon(F_0)$

$$\lambda(t+\varepsilon, F_0) - \varepsilon\,\|\psi_0\| \le \lambda(t, F) \le \lambda(t-\varepsilon, F_0) + \varepsilon\,\|\psi_0\|.$$

Observe the similarity of this expression with the band for the neighbourhood of F_0 in the Lévy distance from Figure 4. It follows that if $T(P_0)$ is unique, necessarily, $T(P) \to T(P_0)$ as $\varepsilon \to 0$. We get that if ψ_0 is bounded and $T(P_0)$ is uniquely defined, the M-estimator T is weakly continuous at P_0. In view of Hampel's Theorem 7, this guarantees qualitative robustness of T.

If ψ_0 is not bounded, we see already from (69) that T does not have to be weakly continuous at any P_0. At the same time, even if ψ_0 is bounded but if $T(P_0)$ is not uniquely defined, T also fails to be weakly continuous. That can be seen, e.g., in Example 3.7. The next theorem summarises our findings.

Theorem 10. *Let $\psi_0 \colon \mathbb{R} \to \mathbb{R}$ be a non-decreasing, but not necessarily continuous, function that takes values of both signs. Then the M-estimator of location T, defined by $\int_{\mathbb{R}} \psi_0(x - T(P)) \,\mathrm{d}\, P(x) = 0$, is translation equivariant. It is weakly continuous and qualitatively robust at $P_0 \in \mathcal{P}(\mathbb{R})$ if and only if ψ_0 is bounded and $T(P_0)$ is unique. The asymptotic breakdown point of T is given by (72). It takes the maximum value $\varepsilon^*(P_0, T) = 1/2$ if and only if $-\psi_0(-\infty) = \psi_0(\infty) < \infty$.*

Example 3.8. The mean functional $T(P) = \int_{\mathbb{R}} x \,\mathrm{d}\, P(x)$ is an M-estimator of location given by $\psi_0(x) = x$. Theorem 10 gives that T fails to be weakly continuous or qualitatively robust, and its asymptotic breakdown point is the minimum possible value $\varepsilon^*(P_0, T) = 0$ at any P_0. △

Example 3.9. The α-quantile functional T from Examples 3.2, 3.4 and 3.6 can be represented as an M-estimator with the function ψ of the form (53). Theorem 10 gives that T is weakly continuous and qualitatively robust at any $P_0 \in \mathcal{P}(\mathbb{R})$ with a uniquely defined $T(P_0)$, which is true if the distribution function F_0 is strictly increasing at $F_0^{-1}(\alpha)$. In particular, if the support of P_0 is connected, the α-quantile is qualitatively robust for any $\alpha \in (0, 1)$. The asymptotic breakdown point of T is $\varepsilon^*(P_0, T) = \min\{\alpha, 1 - \alpha\}$. For $\alpha = 1/2$, we obtain the median functional with an asymptotic breakdown point equal to $1/2$. △

More generally, if ψ_0 is bounded and strictly monotone, the corresponding M-estimator T is always uniquely defined, and Theorem 10 gives that T is then everywhere weakly continuous and qualitatively robust. If ψ_0 is also odd, the asymptotic breakdown point of T is the maximum possible value $1/2$.

Example 3.10. Consider the maximum likelihood estimation from Example 3.1 in the context of location M-estimators. In this case, one takes a distribution function $F_0 \colon \mathbb{R} \to [0, 1]$ with density f_0, and introduces a location parameter $\theta \in \mathbb{R}$ by taking $F_\theta(x) = F_0(x - \theta)$ for $x \in \mathbb{R}$. The density of F_θ is then $f_\theta(x) = f_0(x - \theta)$, $x \in \mathbb{R}$. The parameter θ is estimated using a location M-estimator with

$$\psi(x, t) = -\frac{\partial}{\partial t} \log f_0(x - t).$$

The robustness of such an estimator thus depends on the shape of this (negative) score function. For F_0 the standard normal distribution, we get $\psi_0(x, t) = x - t$; this, of course, corresponds to the non-robust mean functional. For F_0 the standard Cauchy distribution we have

$$\psi(x, t) = \frac{2(x - t)}{1 + (x - t)^2} \quad \text{for } x, t \in \mathbb{R}.$$

This function is bounded and odd in x for each $t \in \mathbb{R}$ given, but it is not monotone. Theorem 10 thus cannot be applied blindly, but it can be shown that the associated M-estimator is translation equivariant, with bounded influence function, and robust with asymptotic breakdown point 1/2. However, it does not have to be defined uniquely.

For F_0 the Laplace distribution

$$F_0(x) = \begin{cases} 1 - \exp(-x)/2 & \text{for } x \geq 0, \\ \exp(x)/2 & \text{for } x < 0, \end{cases} \tag{73}$$

we get $\psi(x,t) = \psi_{1/2}(x,t)$ from (53), and the M-estimator is the median functional. For a plot of all these three score functions see Figure 9. △

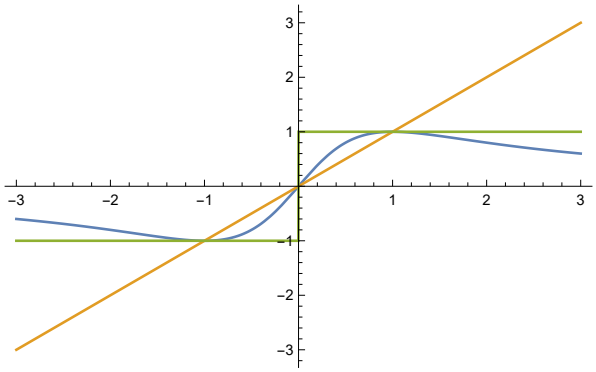

Figure 9: (Negative) score functions $\psi(x,0)$ for the location M-estimators from Example 3.10: F_0 standard normal (orange), F_0 standard Cauchy (blue), and F_0 Laplace (green). The latter two M-functionals are robust, the first one is not.

3.1.4 Robustness of M-estimators of scale

We now briefly deal also with the M-estimation of a scale parameter. Similarly to the location parameter defined in (62), we say that for $\mathcal{X} = \mathbb{R}$, $X \sim P \in \mathcal{P}(\mathbb{R})$ and $Y = a\,X \sim Q \in \mathcal{P}(\mathbb{R})$ with $a > 0$, a parameter $S = S(P)$ is a *scale parameter* of P if

$$S(Q) = a\,S(P) \quad \text{for all } a > 0. \tag{74}$$

A statistical functional S is said to be *scale equivariant* for P if (74) holds true.

M-estimators of scale are naturally defined by the equation

$$\int_{\mathbb{R}} \psi_1\left(\frac{x}{S(P)}\right) \mathrm{d}\,P(x) = 0 \tag{75}$$

that is solved in $S(P) \in (0, \infty)$. The function $\psi(x,t) = \psi_1(x/t)$ is now chosen so that the resulting M-estimator $S(P)$ is scale equivariant, as can be easily seen by verifying (74)

$$\mathsf{E}_Q\, \psi_1\left(\frac{Y}{S(Q)}\right) = \mathsf{E}_P\, \psi_1\left(\frac{a\,X}{S(Q)}\right) = \mathsf{E}_P\, \psi_1\left(\frac{a\,X}{a\,S(P)}\right) = \int_{\mathbb{R}} \psi_1\left(\frac{x}{S(P)}\right) \mathrm{d}\, P(x) = 0.$$

Example 3.11. As for the location case from Example 3.10, also in the scale situation, the simplest M-estimators are the maximum likelihood estimators. Take the scale family of distribution functions $F_\sigma(x) = F_1(x/\sigma)$ for $\sigma > 0$, for $F_1 \colon \mathbb{R} \to [0,1]$ given. If the density of F_1 is f_1 and the derivative of f_1 is denoted by f_1', we get

$$\psi(x,s) = -\frac{\partial}{\partial s} \log\left(\frac{1}{s} f_1\left(\frac{x}{s}\right)\right) = \frac{1}{s} + \frac{f_1'(x/s)}{f_1(x/s)} \frac{x}{s^2}.$$

In particular, for F_1 the distribution function of the standard Gaussian distribution $\mathsf{N}(0,1)$ we get

$$\psi(x,s) = \frac{1}{s}\left(1 - \frac{x^2}{s^2}\right), \quad \text{or alternatively} \quad \psi_1(x) = 1 - x^2,$$

leading to the functional $S(P) = \sqrt{\mathsf{E}_P\, X^2}$ and the estimator

$$S(P_n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}.$$

This is equivalent to taking the even function

$$\psi_1(x) = x^2 - 1 \quad \text{for } x \in \mathbb{R}$$

in the estimating equation (75). △

Similarly as in the example above, the function ψ_1 is usually taken to be even ($\psi_1(-x) = \psi_1(x)$ for all $x \in \mathbb{R}$) and non-decreasing on $[0, \infty)$. For the functional (74) to be well defined, it is also required that $\psi_1(0) < 0$ and $\psi_1(\infty) = \lim_{x\to\infty} \psi_1(x) > 0$. We also assume that ψ_1 is continuous at 0 to avoid trivial degeneracy issues.

The influence function of a scale M-functional S follows directly from Theorem 8. We have

$$\psi'(y,s) = \frac{\partial}{\partial s}\psi_1\left(\frac{y}{s}\right) = -\psi_1'\left(\frac{y}{s}\right)\frac{y}{s^2},$$

for ψ_1' the derivative of ψ_1, and thus

$$\mathsf{IF}(x,P,S) = \frac{\psi_1\left(\frac{x}{s}\right)}{\int_{\mathbb{R}} \psi_1'\left(\frac{y}{s}\right)\frac{y}{s^2}\,\mathrm{d}\,P(y)} = \frac{\psi_1\left(\frac{x}{s}\right)}{\int_{\mathbb{R}} \psi_1'\left(y\right) y\,\mathrm{d}\,P(y)}.$$

In particular, again, the influence function is bounded if and only if both $\psi_1(0) > -\infty$ and $\psi_1(\infty) < \infty$.

Example 3.12. An obvious robust choice of a function ψ_1 for scale M-estimation is

$$\psi_1(x) = \operatorname{sign}(|x| - 1) = \begin{cases} -1 & \text{for } |x| < 1, \\ 0 & \text{for } |x| = 1, \\ 1 & \text{for } |x| > 1. \end{cases}$$

This gives the functional S given as the solution to

$$0 = \int_{\mathbb{R}} \psi_1\left(\frac{x}{s}\right) \mathrm{d}\, P(x) = \mathsf{E}_P \operatorname{sign}(|X| - 1) = -\mathsf{P}(|X| < s) + \mathsf{P}(|X| > s),$$

which is solved by $s = S(P) = \mathsf{med}(|X|)$. This is a functional called the *median absolute deviation from* 0. △

The robustness of scale M-functionals can be explored similarly as for the location M-functionals. Here, we only treat the breakdown point of $S(P_0)$. For simplicity, consider the breakdown point given by the contamination neighbourhood $\mathcal{P}_\varepsilon(P_0)$ from (14). Writing

$$\lambda(s, P) = \int_{\mathbb{R}} \psi_1\left(\frac{x}{s}\right) \mathrm{d}\, P(x)$$

for the left hand side of (75), the functional S breaks down if • $\lambda(s, P) > 0$ for all $s \in \mathbb{R}$, or • $\lambda(s, P) < 0$ for all $s \in \mathbb{R}$, for some $P = (1 - \varepsilon)P_0 + \varepsilon\, Q$. We have

$$\begin{aligned} \lambda(s, P) &= \int_{\mathbb{R}} \left(\psi_1\left(\frac{x}{s}\right) - \psi_1(0)\right) \mathrm{d}\, P(x) + \psi_1(0) \\ &= (1 - \varepsilon) \int_{\mathbb{R}} \left(\psi_1\left(\frac{x}{s}\right) - \psi_1(0)\right) \mathrm{d}\, P_0(x) + \varepsilon \int_{\mathbb{R}} \left(\psi_1\left(\frac{x}{s}\right) - \psi_1(0)\right) \mathrm{d}\, Q(x) + \psi_1(0). \end{aligned}$$

The first two summands are always non-negative because ψ_1 is minimised at 0. The last term $\psi_1(0)$ is negative. Thus, in the first case $\lambda(s, P) > 0$, we want to find $Q \in \mathcal{P}(\mathbb{R})$ so that

$$\int_{\mathbb{R}} \left(\psi_1\left(\frac{x}{s}\right) - \psi_1(0)\right) \mathrm{d}\, Q(x)$$

is as large as possible, which is obviously true if Q is a "mass at infinity", as we also had in the location case in (70), and the maximum value is

$$\sup_{Q \in \mathcal{P}(\mathbb{R})} \int_{\mathbb{R}} \left(\psi_1\left(\frac{x}{s}\right) - \psi_1(0)\right) \mathrm{d}\, Q(x) = \psi_1(\infty) - \psi_1(0).$$

To make $\lambda(s, P) > 0$ for all $s \in (0, \infty)$, we thus need

$$\varepsilon > -\frac{\psi_1(0)}{\psi_1(\infty) - \psi_1(0)}.$$

On the other hand, for any $\varepsilon > 0$ smaller than this constant, we have only that

$$\varepsilon \sup_{Q \in \mathcal{P}(\mathbb{R})} \int_{\mathbb{R}} \left(\psi_1\left(\frac{x}{s}\right) - \psi_1(0)\right) \mathrm{d}\, Q(x) + \psi_1(0) = \varepsilon(\psi_1(\infty) - \psi_1(0)) + \psi_1(0) < 0, \tag{76}$$

which means that taking $s > 0$ finite but large enough in $\lambda(s, P)$, the first term

$$(1-\varepsilon)\int_{\mathbb{R}}\left(\psi_1\left(\frac{x}{s}\right)-\psi_1(0)\right)\,\mathrm{d}\,P_0(x)$$

can be made arbitrarily small (but still positive), and eventually smaller than the negative constant from (76). Observe that in the last argument, we used that ψ_1 is continuous at 0. In conclusion, for $S(P_0)$ to break down "to infinity", we need contamination exactly

$$\varepsilon = -\frac{\psi_1(0)}{\psi_1(\infty)-\psi_1(0)}.$$

In the second breakdown event of $\lambda(s, P) < 0$, a similar argument shows that

$$\inf_{Q\in\mathcal{P}(\mathbb{R})}\int_{\mathbb{R}}\left(\psi_1\left(\frac{x}{s}\right)-\psi_1(0)\right)\,\mathrm{d}\,Q(x) = 0,$$

and this is attained if $Q = \delta_0 \in \mathcal{P}(\mathbb{R})$. For $S(P)$ to break down, we thus need

$$(1-\varepsilon)\sup_{s>0}\int_{\mathbb{R}}\left(\psi_1\left(\frac{x}{s}\right)-\psi_1(0)\right)\,\mathrm{d}\,P_0(x)+\psi_1(0) = (1-\varepsilon)\left(\psi_1(\infty)-\psi_1(0)\right)+\psi_1(0) < 0,$$

which is equivalent with

$$\varepsilon > 1+\frac{\psi_1(0)}{\psi_1(\infty)-\psi_1(0)}.$$

In this case, an "implosion" of $S(P)$ to 0 will happen, which also counts as a breakdown. We have found the following result.

Theorem 11. *Let $\psi_1\colon \mathbb{R}\to\mathbb{R}$ be an even function that is non-decreasing on $[0,\infty)$, continuous at 0, and such that $\psi_1(0) < 0$, $\psi_1(\infty) > 0$. Then the M-estimator of scale S, defined by $\int_{\mathbb{R}}\psi_1(x/S(P))\,\mathrm{d}\,P(x) = 0$, is scale equivariant. The asymptotic breakdown point of S at any $P\in\mathcal{P}(\mathbb{R})$ with respect to the contamination neighbourhood is*

$$\varepsilon^*(P,S) = \min\left\{-\frac{\psi_1(0)}{\psi_1(\infty)-\psi_1(0)}, 1+\frac{\psi_1(0)}{\psi_1(\infty)-\psi_1(0)}\right\}.$$

It takes the maximum value $\varepsilon^*(P_0,T) = 1/2$ if and only if $\psi(\infty) = -\psi(0) < \infty$.

It is not hard to observe that if we consider neighbourhoods given by the Prokhorov distance, the breakdown point in the previous theorem remains the same. However, for neighbourhoods given by the Lévy (or Kolmogorov) distance, the breakdown point halves to

$$\varepsilon^*(P,S) = \frac{1}{2}\min\left\{-\frac{\psi_1(0)}{\psi_1(\infty)-\psi_1(0)}, 1+\frac{\psi_1(0)}{\psi_1(\infty)-\psi_1(0)}\right\}.$$

This is because in the Lévy contamination model, dislocating from P_0 mass $\varepsilon/2$ to $-\infty$ and mass $\varepsilon/2$ to $+\infty$, the Lévy distance between P_0 and the contaminated distribution is only $\varepsilon/2$, while for the Prokhorov distance it would be ε.

3.2 L-estimators: Linear combinations of order statistics

We are given a random sample $X_1, \dots, X_n$ from $P \in \mathcal{P}(\mathbb{R})$ in the space $\mathcal{X} = \mathbb{R}$, and consider the corresponding ordered sample $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$. L-estimators are defined as linear combinations of (a function $h\colon \mathbb{R} \to \mathbb{R}$ of) order statistics

$$T_n(X_1, \dots, X_n) = \sum_{i=1}^{n} a_{n,i}\, h(X_{(i)}). \tag{77}$$

To couple these estimators with a statistical functional T, we assume that the coefficients $a_{n,i} \in \mathbb{R}$ take a specific form

$$a_{n,i} = M\left(\left(\frac{i-1}{n}, \frac{i}{n}\right]\right),$$

for a given signed measure M on $[0,1]$. The measure M is usually absolutely continuous with a density $m\colon [0,1] \to \mathbb{R}$ with respect to the Lebesgue measure. In that case,

$$a_{n,i} = \int_{(i-1)/n}^{i/n} m(x)\,\mathrm{d}\,x,$$

or, if m is continuous, one can also define $a_{n,i} = m((i-1/2)/n)/n$.

The statistical functional T corresponding to (77) is taken to be

$$T(P) = \int_0^1 h\left(F^{-1}(s)\right)\,\mathrm{d}\,M(s), \tag{78}$$

where

$$F^{-1}(s) = \inf\{x \in \mathbb{R}\colon F(x) \ge s\} \quad \text{for } s \in [0,1]$$

is the quantile function associated with the distribution function F of $P \in \mathcal{P}(\mathbb{R})$. For F_n the empirical distribution function of the sample $X_1, \dots, X_n$ we have

$$F_n^{-1}(s) = \begin{cases} -\infty & \text{for } s = 0, \\ X_{(j)} & \text{for } s \in \left(\frac{j-1}{n}, \frac{j}{n}\right] \text{ with } j = 1, 2, \dots, n. \end{cases}$$

Thus, the empirical version of (78) is

$$\begin{aligned} T(F_n) &= \int_0^1 h\left(F_n^{-1}(s)\right)\,\mathrm{d}\,M(s) \\ &= h(-\infty)\,M(\{0\}) + \sum_{i=1}^{n} \int_{((i-1)/n, i/n]} h(X_{(i)})\,\mathrm{d}\,M(s) \\ &= h(-\infty)\,M(\{0\}) + \sum_{i=1}^{n} M\left(\left(\frac{i-1}{n}, \frac{i}{n}\right]\right) h(X_{(i)}), \end{aligned}$$

which coincides with (77) if $h(-\infty) = \lim_{t \to -\infty} h(t)$ is finite and $M(\{0\}) = 0$. The last condition, $M(\{0\}) = 0$, will always be assumed in the section, as otherwise, the associated L-functional faces degeneracy problems.

Example 3.13. Two natural L-estimators of location are

- the α-quantiles obtained from (78) by considering $M = \delta_\alpha$ the Dirac measure at $\alpha \in (0,1)$ and $h(x) = x$; or
- the α-trimmed mean with $\alpha \in (0, 1/2)$ obtained via the density

$$m(x) = \frac{1}{1-2\,\alpha}\mathbb{I}\left(x \in (\alpha, 1-\alpha)\right) \tag{79}$$

of M and $h(x) = x$. The L-estimator customarily associated with the α-trimmed mean is

$$T_n(X_1, \ldots, X_n) = \frac{1}{n - 2\,\lfloor \alpha\, n \rfloor} \sum_{i=\lfloor \alpha\, n \rfloor + 1}^{n - \lfloor \alpha\, n \rfloor} X_{(i)}, \tag{80}$$

where $\lfloor x \rfloor$ is the floor function, that is the largest integer y that satisfies $y \leq x$. Note, however, that the trimmed mean estimator (80) does not equal $T(P_n)$ exactly if $\alpha\, n$ is not an integer. Indeed then, in $T(P_n)$, we need to weight the order statistics slightly differently. We would need to down-weight the extreme statistics $X_{(\lfloor \alpha\, n \rfloor + 1)}$ and $X_{(n - \lfloor \alpha\, n \rfloor)}$. The two estimators $T(P_n)$ and (80) are, nevertheless, asymptotically equivalent.

Among L-estimators, we also find well-known estimators of scale, such as the inter-quartile range obtained using $h(x) = x$ and $M = \delta_{3/4} - \delta_{1/4}$. For the inter-quartile range, M is a proper signed measure. △

3.2.1 Influence function of L-estimators

For the influence function $\mathsf{IF}(x, P, T)$ of the L-functional (78), we need to take $P_t = (1-t)\,P + t\,\delta_x$, plug its distribution function F_t into (78), and compute the difference

$$\begin{aligned} \mathsf{IF}(x, P, T) &= \lim_{t \to 0} \frac{T(P_t) - T(P)}{t} = \int_0^1 \lim_{t \to 0} \frac{h\left(F_t^{-1}(s)\right) - h\left(F^{-1}(s)\right)}{t} \,\mathrm{d}\, M(s) \\ &= \int_0^1 \mathsf{IF}(x, P, h(T_s)) \,\mathrm{d}\, M(s), \end{aligned} \tag{81}$$

supposing that the limit and the integral can be interchanged, where we denote by T_s the statistical functional that assigns to $F \in \mathcal{P}(\mathbb{R})$ the s-quantile $F^{-1}(s)$.

We thus need to find the influence function of the functional $h(T_s)$ for $s \in (0,1)$. Since the influence function is just a derivative of a real function, we can use the chain rule for derivatives. We get that, for h differentiable with derivative h',

$$\mathsf{IF}(x, P, h(T_s)) = \lim_{t \to 0} \frac{h(T_s(P_t)) - h(T_s(P))}{t} = h'(T_s(P))\, \mathsf{IF}(x, P, T_s). \tag{82}$$

It remains to use the expression for the influence function of the s-quantile functional T_s from Example 3.6. To get the final influence function of L-estimators, we put together our formulas (81), (82), and (57). The final expression is given in the following theorem; as for the M-estimators, we gloss over the obvious mild technical conditions in the statement of the theorem.

Theorem 12. *Under mild regularity conditions, the influence function of an L-functional T given by (78) is*

$$\begin{aligned}\mathsf{IF}(x, P, T) &= \int_0^1 \mathsf{IF}(x, P, h(T_s)) \,\mathrm{d}\, M(s) \\ &= \int_0^1 \frac{s\, h'(F^{-1}(s))}{f(F^{-1}(s))} \,\mathrm{d}\, M(s) - \int_{F(x)}^{\infty} \frac{h'(F^{-1}(s))}{f(F^{-1}(s))} \,\mathrm{d}\, M(s).\end{aligned} \tag{83}$$

Proof. The proof follows directly from (81), (82), and (57). It is enough to realise that $x > F^{-1}(s)$ if and only if $F(x) > s$. □

In the common situation when the signed measure M has a density m, we can simplify the influence function in (83) by making a substitution $y = F^{-1}(s)$ to

$$\begin{aligned}\mathsf{IF}(x, P, T) &= \int_0^1 \frac{s\, h'(F^{-1}(s))}{f(F^{-1}(s))} m(s) \,\mathrm{d}\, s - \int_{F(x)}^{\infty} \frac{h'(F^{-1}(s))}{f(F^{-1}(s))} m(s) \,\mathrm{d}\, s \\ &= \int_{\mathbb{R}} \frac{F(y)\, h'(y)}{f(y)} m(F(y)) f(y) \,\mathrm{d}\, y - \int_x^{\infty} \frac{h'(y)}{f(y)} m(F(y)) f(y) \,\mathrm{d}\, y \\ &= \int_{\mathbb{R}} F(y)\, h'(y)\, m(F(y)) \,\mathrm{d}\, y - \int_x^{\infty} h'(y)\, m(F(y)) \,\mathrm{d}\, y \\ &= \int_{-\infty}^{x} h'(y)\, m(F(y)) \,\mathrm{d}\, y - \int_{\mathbb{R}} (1 - F(y))\, h'(y)\, m(F(y)) \,\mathrm{d}\, y.\end{aligned} \tag{84}$$

The second summand is just a constant in $x \in \mathbb{R}$; the influence function is, therefore, relatively simple when written in terms of its derivative

$$\frac{\partial}{\partial x} \mathsf{IF}(x, P, T) = h'(x)\, m(F(x)). \tag{85}$$

Example 3.14. For T the α-trimmed mean from Example 3.13 with $\alpha \in (0, 1/2)$ we have $h(x) = x$ and m of the form (79). Applying this to (85) we get that $\mathsf{IF}(x, P, T)$ is constant for $x \notin [F^{-1}(\alpha), F^{-1}(1-\alpha)]$, and for $x \in [F^{-1}(\alpha), F^{-1}(1-\alpha)]$ we have

$$\mathsf{IF}(x, P, T) = \frac{x}{1 - 2\,\alpha} + c$$

for an appropriate constant $c \in \mathbb{R}$. The exact expression for the influence function of T can be computed from (84), but it is somewhat cumbersome. We perform the computation for F such that F^{-1} does not have discontinuities at α and $1 - \alpha$. That will allow us to write

$F(F^{-1}(\alpha)) = \alpha$ and $F(F^{-1}(1-\alpha)) = 1-\alpha$. We begin by computing the second term on the right-hand side of (84). We integrate by parts the Stieltjes integral[3] [23, Theorem 16.4] to get

$$\begin{aligned}(1-2\,\alpha)\int_{\mathbb{R}}(1-F(y))\,m(F(y))\,\mathrm{d}\,y &= \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)}(1-F(y))\,\mathrm{d}\,y \\ &= F^{-1}(1-\alpha)(1-(1-\alpha)) - F^{-1}(\alpha)(1-\alpha) + \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} y\,\mathrm{d}\,F(y) \\ &= \alpha\,F^{-1}(1-\alpha) + (\alpha-1)F^{-1}(\alpha) + \int_{\alpha}^{1-\alpha} F^{-1}(s)\,\mathrm{d}\,s.\end{aligned}$$

The quantity on the right-hand side is interesting; it can be written as $W(P) - F^{-1}(\alpha)$, where

$$W(P) = \int_{\alpha}^{1-\alpha} F^{-1}(s)\,\mathrm{d}\,s + \alpha\left(F^{-1}(\alpha) + F^{-1}(1-\alpha)\right) \tag{86}$$

is another well-known L-functional called the α-Windsorized mean of $P \in \mathcal{P}(\mathbb{R})$ with distribution function F.

Returning to the influence function of the α-trimmed mean, consider now also the first term in (84). It takes the form

$$\int_{-\infty}^{x} m(F(y))\,\mathrm{d}\,y = \frac{1}{1-2\,\alpha}\int_{-\infty}^{x} \mathbb{I}\left(F^{-1}(\alpha) < y < F^{-1}(1-\alpha)\right)\,\mathrm{d}\,y$$

which is zero if $x \le F^{-1}(\alpha)$. For $x \in (F^{-1}(\alpha), F^{-1}(1-\alpha))$ we get

$$\int_{-\infty}^{x} \mathbb{I}\left(F^{-1}(\alpha) < y < F^{-1}(1-\alpha)\right)\,\mathrm{d}\,y = x - F^{-1}(\alpha),$$

and finally for $x \ge F^{-1}(1-\alpha)$

$$\int_{-\infty}^{x} \mathbb{I}\left(F^{-1}(\alpha) < y < F^{-1}(1-\alpha)\right)\,\mathrm{d}\,y = F^{-1}(1-\alpha) - F^{-1}(\alpha).$$

Putting all these results together, the influence function of the α-trimmed mean is

$$\mathsf{IF}(x,P,T) = \begin{cases} \frac{1}{1-2\,\alpha}\left(F^{-1}(\alpha) - W(P)\right) & \text{for } x \le F^{-1}(\alpha), \\ \frac{1}{1-2\,\alpha}\left(x - W(P)\right) & \text{for } x \in (F^{-1}(\alpha), F^{-1}(1-\alpha)), \\ \frac{1}{1-2\,\alpha}\left(F^{-1}(1-\alpha) - W(P)\right) & \text{for } x \ge F^{-1}(1-\alpha). \end{cases} \tag{87}$$

We see that the influence function is continuous, piecewise linear, and the gross error sensitivity of the α-trimmed mean is bounded. For F symmetric in the sense $F(-x) = 1 - F(x)$ for all $x \in \mathbb{R}$ we certainly have $W(P) = 0$ and

$$\gamma^*(P,T) = \frac{F^{-1}(1-\alpha)}{1-2\,\alpha}.$$

△

[3] For $F, G\colon \mathbb{R} \to [0,1]$ continuous distribution functions and $a < b$ we have $F(b)G(b) - F(a)G(a) = \int_a^b F(x)\,\mathrm{d}\,G(x) + \int_a^b G(x)\,\mathrm{d}\,F(x)$.

Example 3.15. The α-Windsorized mean $W(P)$ from (86) follows an idea similar to the trimmed mean. In the computation of the trimmed mean in (80), we "throw away" all extreme observations $X_{(1)} \leq \cdots \leq X_{(\lfloor \alpha\, n \rfloor)}$ and $X_{(n-\lfloor \alpha\, n \rfloor+1)} \leq \cdots \leq X_{(n)}$. In contrast, in the Windsorized mean, we rather replace all the leftmost observations $X_{(i)}$, $i = 1, \ldots, \lfloor \alpha\, n \rfloor$, by the boundary order statistic $X_{(\lfloor \alpha\, n \rfloor+1)}$, and all the rightmost $X_{(i)}$, $i = (n - \lfloor \alpha\, n \rfloor + 1), \ldots, n$ by $X_{(n-\lfloor \alpha\, n \rfloor)}$. Then we take the usual average of these modified n data points.

The influence function of the Windsorized mean can be computed directly. Because

$$W(P) = (1 - 2\,\alpha)\, T(P) + \alpha\, (T_\alpha(P) + T_{1-\alpha}(P))$$

for T the α-trimmed mean and T_α the α-quantile functionals, we immediately get that

$$\mathsf{IF}(x, W, P) = (1 - 2\,\alpha)\, \mathsf{IF}(x, T, P) + \alpha\, (\mathsf{IF}(x, T_\alpha, P) + \mathsf{IF}(x, T_{1-\alpha}, P))\,.$$

It remains to plug in the expressions for the influence function of the trimmed mean from (87), and the influence function of the quantiles from (57). In Figure 10 we see that the influence function of W is discontinuous at the boundary points $x = F^{-1}(\alpha)$ and $x = F^{-1}(1-\alpha)$. This is because of the mass that the measure M puts in the functional W to the two quantiles at values α and $1-\alpha$. We saw the same phenomenon already in the expression for the influence function of the α-quantile in (57).

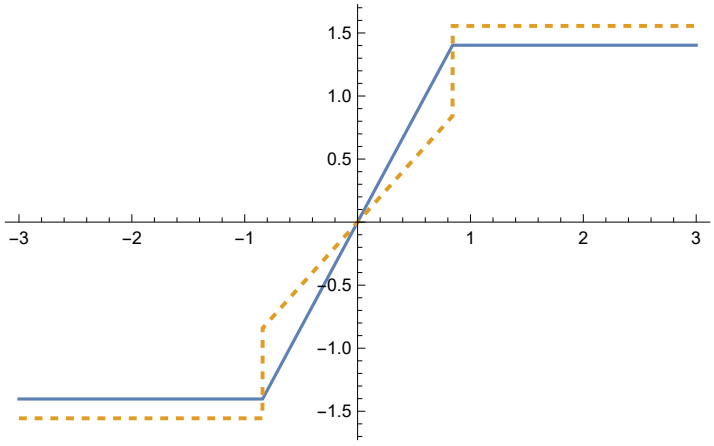

Figure 10: The influence function of the α-trimmed mean (blue) and the α-Windsorized mean (dashed orange) for $\alpha = 0.2$ and P the standard normal distribution.

△

3.2.2 Robustness of L-estimators

We consider the common situation when $h(x) = x$ is the identity function and M is a (non-negative) measure of total mass 1. In this case, we obtain an L-estimator (78) that is both

translation and *scale equivariant*. It means that for any $X \sim P \in \mathcal{P}(\mathbb{R})$ with distribution function F_0 and $Y = aX + b \sim Q \in \mathcal{P}(\mathbb{R})$ with distribution function $F_{a,b}$ with $a > 0$ and $b \in \mathbb{R}$, we have

$$F_{a,b}(x) = \mathsf{P}(aX + b \le x) = F_0\left(\frac{x-b}{a}\right) \quad \text{for all } x \in \mathbb{R},$$

and thus also

$$F_{a,b}^{-1}(\alpha) = a\,F_0^{-1}(\alpha) + b \quad \text{for all } \alpha \in (0,1), \tag{88}$$

and

$$T(Q) = \int_0^1 F_{a,b}^{-1}(s)\,\mathrm{d}\,M(s) = a\int_0^1 F_0^{-1}(s)\,\mathrm{d}\,M(s) + b = a\,T(P) + b. \tag{89}$$

Consider first the situation when the support of M contains one of the endpoints of $[0,1]$, say 0. Then $M([0,1/n)) > 0$ for all $n = 1, 2, \ldots$, and it is not difficult to see that the resulting L-functional cannot be weakly continuous. This is shown in the next example.

Example 3.16. Take $P \in \mathcal{P}(\mathbb{R})$ the uniform distribution on $[-1,1]$. Contaminate P by an appropriate Dirac measure, defining a sequence

$$P_n = \left(1 - \frac{1}{n}\right)P + \frac{1}{n}\delta_{x_n} \quad \text{for } n = 1, 2, \ldots,$$

where $x_n = \min\{-n, -n/(M[0,1/n))\}$. For F_n the distribution function of P_n we have

$$F_n(x_n) = P_n((-\infty, x_n]) \ge \frac{1}{n}\delta_{x_n}((-\infty, x_n]) = \frac{1}{n},$$

which gives

$$F_n^{-1}(s) \le 1 \quad \text{for all } s \in (0,1),$$
$$F_n^{-1}(1/n) \le x_n.$$

Altogether, we have $P_n \xrightarrow[n\to\infty]{\mathrm{w}} P$ and $T(P) \in [-1,1]$, but

$$T(P_n) = \int_0^1 F_n^{-1}(s)\,\mathrm{d}\,M(s) \le x_n\,M([0,1/n)) + 1\,M([1/n,1]) \le -n + 1,$$

and naturally $T(P_n)$ does not converge to $T(P)$. △

Defining $\alpha \in [0, 1/2]$ the largest number such that the interval $[\alpha, 1-\alpha]$ contains the support of M, Example 3.16 shows that if $\alpha = 0$, the L-functional (78) cannot be weakly continuous.

We thus focus on $\alpha \in (0, 1/2]$. In the definition (78) of $T(P)$, we consider only the s-quantiles of P with $s \in [\alpha, 1-\alpha]$. We already saw in Example 3.9 that the asymptotic breakdown point of each such quantile is at least α, for $\alpha \le 1/2$. At the same time, (a direct modification of) the previous Example 3.16 also shows that $\varepsilon^*(P,T) \le \alpha$, and in particular $\varepsilon^*(P,T) = \alpha$. It is actually easy to see more. Just as for the M-functionals in Section 3.1.3, also the L-functional T is non-decreasing in the sense of stochastic ordering, see Lemma 2.

Lemma 3. *Let M be a probability measure on $[0,1]$, and denote by F the distribution function of $P \in \mathcal{P}(\mathbb{R})$. Then the L-functional $T(P) = \int_0^1 F^{-1}(s)\,\mathrm{d}\,M(s)$ is non-decreasing in P in the sense of stochastic ordering.*

Proof. Let $P \in \mathcal{P}(\mathbb{R})$ with distribution function F be stochastically larger than $Q \in \mathcal{P}(\mathbb{R})$ with distribution function G. That means $F(x) \leq G(x)$ for all $x \in \mathbb{R}$. For any $s \in [0,1]$ fixed we thus have that $F(x) \geq s$ implies $G(x) \geq s$, which gives

$$F^{-1}(s) = \inf\{x \in \mathbb{R} \colon F(x) \geq s\} \geq \inf\{x \in \mathbb{R} \colon G(x) \geq s\} = G^{-1}(s).$$

This allows us to write

$$T(P) = \int_0^1 F^{-1}(s)\,\mathrm{d}\,M(s) \geq \int_0^1 G^{-1}(s)\,\mathrm{d}\,M(s) = T(Q),$$

as we needed to show. □

Thanks to Lemma 3, the maximum positive bias $b_+(\varepsilon)$ and the minimum negative bias $b_-(\varepsilon)$ from (65) of an L-estimator T at $P_0 \in \mathcal{P}(\mathbb{R})$ such that $T(P_0) = 0$ (this assumption is without loss of generality, due to (89)) is given by

$$b_+(\varepsilon) = T(F_1) \quad \text{and} \quad b_-(\varepsilon) = T(F_2)$$

with F_1 and F_2 precisely as in Section 3.1.3. To compute $T(F_1)$, we need the expression for the s-quantiles of

$$F_1(x) = \begin{cases} 0 & \text{if } x < F_0^{-1}(\varepsilon) + \varepsilon, \\ F_0(x-\varepsilon) - \varepsilon & \text{if } x \geq F_0^{-1}(\varepsilon) + \varepsilon, \end{cases}$$

see (67). Here, F_0 is the distribution function of P_0. For $s \in (0,1)$ we get

$$\begin{aligned} F_1^{-1}(s) &= \inf\{x \in \mathbb{R} \colon F_0(x-\varepsilon) - \varepsilon \geq s\} \\ &= \inf\{y + \varepsilon \in \mathbb{R} \colon F_0(y) \geq s + \varepsilon\} = \varepsilon + F_0^{-1}(s+\varepsilon). \end{aligned} \tag{90}$$

If $s > 1 - \varepsilon$, we obtain

$$F_1^{-1}(s) = \infty,$$

as expected by considering Figure 8. Thus, if the upper endpoint of the support of M is larger than $1-\varepsilon$ (which corresponds to $\alpha < \varepsilon$), we have $T(F_1) = \infty$. In the other situation ($\varepsilon < \alpha$) we get

$$\begin{aligned} b_+(\varepsilon) = T(F_1) &= \varepsilon + \int_\alpha^{1-\alpha} F_0^{-1}(s+\varepsilon)\,\mathrm{d}\,M(s), \\ b_-(\varepsilon) = T(F_2) &= -\varepsilon + \int_\alpha^{1-\alpha} F_0^{-1}(s-\varepsilon)\,\mathrm{d}\,M(s). \end{aligned} \tag{91}$$

From Example 3.9 we know that the s-quantile functional with $s \in (0,1)$ is weakly continuous at F_0 if and only if F_0^{-1} is continuous at s. If each $s \in (0,1)$ is a point of continuity of F_0^{-1}, we have

$$F_0^{-1}(s+\varepsilon) \to F_0^{-1}(s) \quad \text{as } \varepsilon \to 0, \tag{92}$$

and we get with $\varepsilon \to 0$ that

$$T(F_1) = \varepsilon + \int_\alpha^{1-\alpha} F_0^{-1}(s+\varepsilon)\,\mathrm{d}\,M(s) \to \int_\alpha^{1-\alpha} F_0^{-1}(s)\,\mathrm{d}\,M(s) = T(F_0). \tag{93}$$

Suppose now that F_0^{-1} has points of discontinuity and let M be absolutely continuous with density m. Since F_0^{-1} is non-decreasing, there are only (at most) finitely many points of discontinuity of F_0^{-1}. Thus, for M-almost all $s \in (\alpha, 1-\alpha)$ we still have (92). Consequently, we can interchange the limit and the integral and write (93). It turns out that the only situation when points of discontinuity of F_0^{-1} pose a problem for L-estimators is when F_0^{-1} and the distribution function of M both share the same points of discontinuity. This is exactly what happens when the s-quantile functional happens to be weakly discontinuous, see Example 3.7.

Overall, we have found that as $\varepsilon \to 0$, both $b_+(\varepsilon)$ and $b_-(\varepsilon)$ go to zero if F_0^{-1} and the distribution function of M do not have common discontinuity points in the interval $[\alpha, 1-\alpha]$. Under that assumption, the L-functional T is weakly continuous. We can now summarise all our observations.

Theorem 13. *Let M be a (non-negative) measure on $[0,1]$, and let T be the L-functional defined by $T(F) = \int_0^1 F^{-1}(s)\,\mathrm{d}\,M(s)$. Let $\alpha \geq 0$ be the largest number such that the interval $[\alpha, 1-\alpha]$ contains the support of M. Then T is weakly continuous at F_0 if and only if (i) $\alpha > 0$, and (ii) F_0^{-1} does not share a point of discontinuity with the distribution function of M in $[\alpha, 1-\alpha]$. In addition, $\varepsilon^*(P,T) = \alpha$ for any $P \in \mathcal{P}(\mathbb{R})$.*

If the measure M is signed, one obtains a result similar to Theorem 13 by decomposing M into $M^+ - M^-$, with both M^+ and M^- (non-negative) measures in $[0,1]$. Even though we do not prove the Fréchet differentiability of the L-functionals, the asymptotic representation and the asymptotic normality result from Theorem 5 hold true, under reasonable assumptions. One such result can be found in, e.g., [12, Theorem 3.8].

3.3 R-estimators: Rank-based estimation

The R-estimators are constructed by inverting rank tests. Take the two-sample Wilcoxon test [15, Section 6.4]. We have two independent random samples $X_1, \ldots, X_n$ from $P \in \mathcal{P}(\mathbb{R})$, and $Y_1, \ldots, Y_m$ from $Q \in \mathcal{P}(\mathbb{R})$, where the distributions of P and Q are the same, but possibly

shifted by a constant $\delta \in \mathbb{R}$. We want to test whether $\delta = 0$, meaning that $P = Q$. To do that, we first pool both random samples into a single vector $\boldsymbol{Z} = (Z_1, \ldots, Z_{n+m})^\mathsf{T} = (X_1, \ldots, X_n, Y_1, \ldots, Y_m)^\mathsf{T}$ and then consider T defined as the sum of the ranks

$$R_j = \sum_{i=1}^{n+m} \mathbb{I}\left(Z_i \leq Z_j\right) \tag{94}$$

corresponding to the points X_j from the first random sample, i.e. $T = \sum_{j=1}^{n} R_j$. Take, for simplicity, the case when $n = m$. Then, we cannot reject the null hypothesis of $\delta = 0$ if T is close to

$$\left(\sum_{i=1}^{2\,n} i\right)/2 = 2\,n(2\,n+1)/4 = n(n+1/2) = \sum_{i=1}^{n}(1/2+n).$$

That is equivalent with

$$0 = 2\,n^2 \frac{1}{n} \sum_{i=1}^{n} \frac{R_i - 1/2 - n}{2\,n},$$

or, dividing by $2\,n^2$, also

$$0 = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i - 1/2 - n}{2\,n} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{R_i - 1/2}{2\,n} - \frac{1}{2} \right)$$

whose right-hand side can be written as

$$\frac{1}{n} \sum_{i=1}^{n} \left(\frac{R_i - 1/2}{2\,n} - 1/2 \right) = \frac{1}{n} \sum_{i=1}^{n} a(R_i). \tag{95}$$

Here, we denote

$$a(t) = \frac{t - 1/2}{2\,n} - 1/2 = 2\,n \int_{(t-1)/(2\,n)}^{t/(2\,n)} J(s)\,\mathrm{d}\,s,$$

for

$$J(s) = s - 1/2 \quad \text{for } s \in [0, 1]. \tag{96}$$

The rank (94) of X_i can be written as the value of the empirical distribution function of $\boldsymbol{Z}$ at the point X_i, multiplied by $2\,n$. If we denote by F_n and G_n the empirical distribution functions of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, respectively, we get

$$\frac{R_j}{2\,n} = \frac{1}{2}\left(F_n(X_j) + G_n(X_j)\right) \quad \text{for } j = 1, \ldots, n.$$

The functional analogue to the rank statistic (95) is therefore

$$\int_{\mathbb{R}} J\left(\frac{1}{2}\left(F(x) + G(x)\right)\right) \mathrm{d}\,F(x) \tag{97}$$

for F and G the distribution functions of P and Q, respectively. If the true shift δ is zero, the integral in (97) should be (close to) zero too. Note, however, that (97) does not correspond

to the statistic (95) precisely. If we apply the empirical distributions F_n and G_n to (97) we get

$$\int_{\mathbb{R}} J\left(\frac{1}{2}\left(F_n(x)+G_n(x)\right)\right)\,\mathrm{d}\,F_n(x) = \frac{1}{n}\sum_{j=1}^{n} J\left(\frac{1}{2}\left(F_n(X_j)+G_n(X_j)\right)\right)$$
$$= \frac{1}{n}\sum_{j=1}^{n} J\left(\frac{R_j}{2\,n}\right) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{R_j}{2\,n}-\frac{1}{2}\right).$$

The last expression differs from (95) by $1/(4\,n)$, which is, however, asymptotically negligible.

To estimate the location shift $\delta \in \mathbb{R}$ between P and Q, one can now invert the rank test based on the statistic (95) or the expression (97). The idea is to consider different shifts of the second random sample $Y_1, \ldots, Y_n$ by various $\delta \in \mathbb{R}$. To use the test above, we thus first modify all Y_i to $Y_i + \delta$, and build $\boldsymbol{Z} = (X_1, \ldots, X_n, Y_1+\delta, \ldots, Y_n+\delta)^{\mathsf{T}}$. For this sample, the standard ranks (94) are computed, and the statistic (95) is used. The corresponding functional (97) is

$$\int_{\mathbb{R}} J\left(\frac{1}{2}\left(F(x)+G(x-\delta)\right)\right)\,\mathrm{d}\,F(x). \tag{98}$$

If the true shift between P and Q is δ_0, one then expects that as a function of $\delta \in \mathbb{R}$, the expression (98) will be equal to zero for $\delta = \delta_0$.

We now use this idea to construct an estimator in the situation when only a single sample $X_1, \ldots, X_n$ is at our disposal. One can replace the second sample with the mirror images of X_i around δ

$$Y_i = \delta - (X_i - \delta) = 2\,\delta - X_i \quad \text{for } i = 1, \ldots, n.$$

The distribution function G of Y_i is then

$$G(x) = \mathsf{P}\left(2\,\delta - X_1 \le x\right) = \mathsf{P}\left(2\,\delta - x \le X_1\right) = 1 - F\left(2\,\delta - x\right)$$

if F is continuous. The expression (98) changes to

$$\int_{\mathbb{R}} J\left(\frac{1}{2}\left(F(x)+1-F(2\,\delta-x)\right)\right)\,\mathrm{d}\,F(x), \tag{99}$$

and once again, one searches for $\delta \in \mathbb{R}$ that makes this expression equal to zero. This defines the (one-sample) R-functional as a solution $T(P) = T(F) \in \mathbb{R}$ to the equation

$$\int_{\mathbb{R}} J\left(\frac{1}{2}\left(F(x)+1-F(2\,T(F)-x)\right)\right)\,\mathrm{d}\,F(x) = 0. \tag{100}$$

The function J in (100) and (108), of course, does not have to be only of the form (96). In general, it is assumed that $J\colon [0,1] \to \mathbb{R}$ has the property

$$\int_0^1 J(s)\,\mathrm{d}\,s = 0, \tag{101}$$

which corresponds to the fact that the expected value of the test statistic (95) under the null hypothesis is zero. Indeed, if the distribution function F is symmetric around, say, the origin $\delta = 0$, we then obtain in (99)

$$\int_{\mathbb{R}} J\left(\frac{1}{2}\left(F(x) + 1 - F(-x)\right)\right) \mathrm{d}\, F(x) = \int_{\mathbb{R}} J\left(F(x)\right) \mathrm{d}\, F(x) = \int_0^1 J(s) \,\mathrm{d}\, s = 0, \tag{102}$$

as we wanted.

Example 3.17. The simplest reasonable choice of the function J is

$$J(s) = \begin{cases} -1 & \text{for } s \in [0, 1/2), \\ 0 & \text{for } s = 1/2, \\ 1 & \text{for } s \in (1/2, 1]. \end{cases}$$

Then, in formula (99) with F replaced by the empirical distribution function F_n of $X_1, \ldots, X_n$, we identify that

$$\frac{1}{2}\left(F_n(X_i) + 1 - F_n(2\,\delta - X_i)\right)$$

is the rank of X_i in the pooled sample $\boldsymbol{X}(\delta) = (X_1, \ldots, X_n, 2\,\delta - X_1, \ldots, 2\,\delta - X_n)^\mathsf{T}$ divided by $(2\,n)$. Certainly, the median of the pooled sample $\boldsymbol{X}(\delta)$ is δ. Thus, the sample version of formula (99) counts the number of X_i, $i = 1, \ldots, n$ that lie below δ with a minus sign, and the number of X_i, $i = 1, \ldots, n$ that lie above δ with a plus sign. This sum equals zero if δ is the median of $X_1, \ldots, X_n$, and we obtain the median functional from Example 3.4 as a special case of an R-estimator. △

Example 3.18. Taking J from (96), we obtain an R-estimator that corresponds to the Wilcoxon test. The sample version of this R-estimator is the Hodges-Lehmann estimator, given as the median of the set of n^2 points $(X_i + X_j)/2$, $i, j = 1, \ldots, n$. To see this,[4] note that $T(F)$ is defined as the solution to (100), giving

$$0 = \int_{\mathbb{R}} \left(\frac{1}{2}\left(F(x) + 1 - F(2\,T(F) - x)\right) - \frac{1}{2}\right) \mathrm{d}\, F(x).$$

That is equivalent with

$$\int_{\mathbb{R}} F(x) \,\mathrm{d}\, F(x) = \int_{\mathbb{R}} F(2\,T(F) - x) \,\mathrm{d}\, F(x), \tag{103}$$

with the left-hand side equal to

$$\int_{\mathbb{R}} F(x) \,\mathrm{d}\, F(x) = \int_0^1 s \,\mathrm{d}\, s = 1/2.$$

[4] Derivation thanks to F. Bočinec.

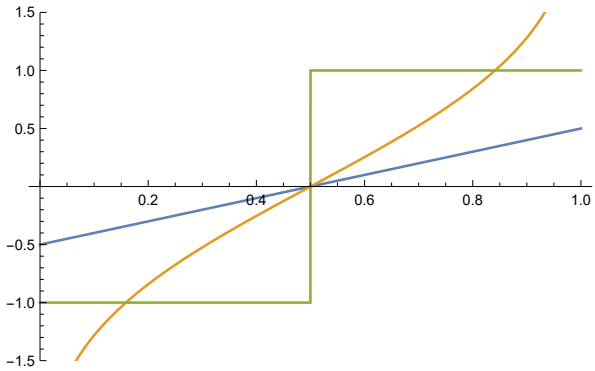

Figure 11: Three commonly used functions J for R-estimators: the Hodges-Lehmann function (96) (blue), the function giving the median from Example 3.17 (green), and the normal scores function $J = \Phi^{-1}$ (orange).

Now, plug the empirical distribution function $F_n(x)$ of $X_1, \dots, X_n$ into the right-hand side of (103) instead of F. We get

$$\begin{aligned}\frac{1}{2} &= \int_{\mathbb{R}} F_n(2\,T(F_n) - x)\,\mathrm{d}\,F_n(x) = \frac{1}{n}\sum_{i=1}^{n} F_n(2\,T(F_n) - X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}\left(X_j \le 2\,T(F_n) - X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{I}\left(\frac{X_j + X_i}{2} \le T(F_n)\right).\end{aligned}$$

Here, the right-hand side expression can be seen as the empirical distribution function G_{n^2} of the n^2 values $Y_{i,j} = (X_i + X_j)/2$ for $i, j = 1, \dots, n$ at the point $T(F_n)$. The value of this distribution function G_{n^2} equals $1/2$ at $T(F_n)$ if $T(F_n)$ is (close to) the median all the values $Y_{i,j}$, $i, j = 1, \dots, n$. The sample version of the R-estimator that corresponds to the Wilcoxon test is thus indeed the Hodges-Lehmann estimator (sometimes also called the pseudo-median, see, e.g. [15, Section 5.4]). △

Example 3.19. Other choices of J have been considered in the literature. Typically, it is assumed that J is symmetric in the sense that

$$J(1-t) = -J(t) \quad \text{for all } t \in [0,1], \tag{104}$$

A common choice is $J = \Phi^{-1}$, which gives the so-called normal scores R-estimator of location. For a plot of commonly used functions J see Figure 11. △

3.3.1 Influence function of R-estimators

The influence function of a general R-estimator can be computed by replacing F by $F_t = (1-t)\,F + t\,\delta_x$ in the formula (100) with $x \in \mathbb{R}$, differentiating with respect to t, and setting

$t = 0$ in the resulting derivative. The computation is, however, rather tedious, and we do not perform it. Under the condition (104) of symmetry of J, it is shown in [12, Section 3.4.1] that

$$\mathsf{IF}(x, P, T) = \frac{U(x) - \int_{\mathbb{R}} U(y)\, f(y)\, \mathrm{d}\, y}{\int_{\mathbb{R}} U'(y)\, f(y)\, \mathrm{d}\, y} = \frac{U(x) - \int_{\mathbb{R}} U(y)\, \mathrm{d}\, F(y)}{\int_{\mathbb{R}} U'(y)\, \mathrm{d}\, F(y)}. \tag{105}$$

Here, U is the primitive function of

$$U'(x) = J'\left(\frac{1}{2}\left(F(x) + 1 - F(2\,T(F) - x)\right)\right) f(2\,T(F) - x),$$

J' is the derivative of J, and f is the density of F. The formula (105) simplifies if F is symmetric, i.e. if $F(-x) = 1 - F(x)$ for all $x \in \mathbb{R}$. In that situation, we have by (101) and (102) that $T(F) = 0$. That gives

$$U'(x) = J'\left(\frac{1}{2}\left(F(x) + 1 - F(2\,T(F) - x)\right)\right) f(2\,T(F) - x) = J'(F(x))\, f(x),$$

whose primitive function is simply

$$U(x) = J(F(x)),$$

Further,

$$\int_{\mathbb{R}} J(F(y))\, f(y)\, \mathrm{d}\, y = \int_{\mathbb{R}} J(F(x))\, \mathrm{d}\, F(x) = \int_0^1 J(s)\, \mathrm{d}\, s = 0.$$

We get that for F symmetric, we can simplify

$$\mathsf{IF}(x, P, T) = \frac{J(F(x)) - \int_{\mathbb{R}} J(F(y))\, f(y)\, \mathrm{d}\, y}{\int_{\mathbb{R}} J'(F(y))\, (f(y))^2\, \mathrm{d}\, y} = \frac{J(F(x))}{\int_{\mathbb{R}} J'(F(y))\, (f(y))^2\, \mathrm{d}\, y}. \tag{106}$$

Example 3.20. The Hodges-Lehmann estimator $T(P) = T(F)$ from Example 3.18 solves the equation

$$\begin{aligned} 0 &= \int_{\mathbb{R}} \left(\frac{1}{2}\left(F(x) + 1 - F(2\,T(F) - x)\right) - \frac{1}{2}\right) \mathrm{d}\, F(x) \\ &= \int_{\mathbb{R}} \frac{F(x)}{2}\, \mathrm{d}\, F(x) - \int_{\mathbb{R}} \frac{F(2\,T(F) - x)}{2}\, \mathrm{d}\, F(x) \\ &= \int_0^1 \frac{s}{2}\, \mathrm{d}\, s - \int_{\mathbb{R}} \frac{F(2\,T(F) - x)}{2}\, \mathrm{d}\, F(x) \\ &= \frac{1}{4} - \int_{\mathbb{R}} \frac{F(2\,T(F) - x)}{2}\, \mathrm{d}\, F(x). \end{aligned} \tag{107}$$

We have $J'(s) = 1$ for all $s \in (0, 1)$, which gives

$$U(x) = -F(2\,T(F) - x) \quad \text{for } x \in \mathbb{R},$$

and the influence function takes the form

$$\begin{aligned} \mathsf{IF}(x, P, T) &= \frac{-F(2\,T(F) - x) + \int_{\mathbb{R}} F(2\,T(F) - y)\, f(y)\, \mathrm{d}\, y}{\int_{\mathbb{R}} f(2\,T(F) - y)\, f(y)\, \mathrm{d}\, y} \\ &= \frac{1/2 - F(2\,T(F) - x)}{\int_{\mathbb{R}} f(2\,T(F) - y)\, f(y)\, \mathrm{d}\, y}. \end{aligned}$$

where the second equality follows from (107). For F symmetric this simplifies to

$$\mathsf{IF}(x, P, T) = \frac{F(x) - 1/2}{\int_{\mathbb{R}} (f(y))^2 \,\mathrm{d}\, y}.$$

Theorem 5 now suggests that for F symmetric, the asymptotic normality result

$$\sqrt{n}(T(P_n) - T(P)) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}(0, A(P, T))$$

might be true for P_n empirical measure sampled from P, where

$$\begin{aligned} A(P, T) &= \int_{\mathbb{R}} (\mathsf{IF}(x, P, T))^2 \,\mathrm{d}\, P(x) = \frac{\int_{\mathbb{R}} (F(x) - 1/2)^2 \,\mathrm{d}\, F(x)}{\left(\int_{\mathbb{R}} (f(y))^2 \,\mathrm{d}\, y\right)^2} \\ &= \frac{\int_0^1 (s - 1/2)^2 \,\mathrm{d}\, s}{\left(\int_{\mathbb{R}} (f(y))^2 \,\mathrm{d}\, y\right)^2} = \frac{1}{12 \left(\int_{\mathbb{R}} (f(y))^2 \,\mathrm{d}\, y\right)^2}. \end{aligned}$$

This result is, indeed, true. Both the asymptotic variance and the influence function suggest that from the viewpoint of infinitesimal robustness, the Hodges-Lehmann estimator is surprisingly non-robust since the integral in the denominator can be arbitrarily small. For example, for f the density of the uniform distribution on the interval $[-M, M]$ for $M > 0$ we have $\int_{\mathbb{R}} (f(y))^2 \,\mathrm{d}\, y = 1/(2\,M)$, and as $M \to \infty$ we get $\gamma^*(P, T) \to \infty$. We saw the same problem with the influence function of the quantiles in (57). △

3.3.2 Robustness of R-estimators

In what follows, it turns out that the R-functional given by (100) is more convenient to work with if we substitute $F(x) = s$ in (100). That gives an alternative expression for $T(F)$ in the form

$$\int_0^1 J\left(\frac{1}{2}\left(s + 1 - F(2\,T(F) - F^{-1}(s))\right)\right) \mathrm{d}\, s = 0. \tag{108}$$

Expressions (100) and (108) are equivalent only if the distribution function F is continuous and strictly increasing; the difference is however asymptotically negligible as $n \to \infty$.

We assume that J is non-decreasing and symmetric as in (104). Under our assumption of integrability of J following from (101), the function

$$\lambda(t, P) = \lambda(t, F) = \int_0^1 J\left(\frac{1}{2}\left(s + 1 - F(2\,t - F^{-1}(s))\right)\right) \mathrm{d}\, s \tag{109}$$

is well-defined. It is also monotone in both $t \in \mathbb{R}$ and $P \in \mathcal{P}(\mathbb{R})$, similarly as the function (66) defining the M-estimators, see Lemma 2.

Lemma 4. *The function* (109) *is*

- *non-increasing in* $t \in \mathbb{R}$, *and*

- *non-decreasing in $P \in \mathcal{P}(\mathbb{R})$ in the sense of stochastic ordering.*

Proof. Take any $t_1 < t_2$ and fix F. Then $F(2\,t_1 - F^{-1}(s)) \leq F(2\,t_2 - F^{-1}(s))$ for all $s \in [0,1]$, and consequently the integrand in (109) is non-increasing in t for each $s \in [0,1]$. Naturally, also $\lambda(t_1, F) \leq \lambda(t_2, F)$.

Now, let $t \in \mathbb{R}$ be fixed, and take F_2 stochastically larger than F_1, i.e. $F_2(x) \leq F_1(x)$ for all $x \in \mathbb{R}$. Then $F_1^{-1}(s) \leq F_2^{-1}(s)$ for all $s \in [0,1]$ as in the proof of Lemma 3, and

$$F_2(2\,t - F_2^{-1}(s)) \leq F_2(2\,t - F_1^{-1}(s)) \leq F_1(2\,t - F_1^{-1}(s)) \quad \text{for all } s \in [0,1].$$

Using the same argument as before, we get $\lambda(t, F_2) \geq \lambda(t, F_1)$. □

Observing that for any $P \in \mathcal{P}(\mathbb{R})$ is (109) non-increasing in $t \in \mathbb{R}$, we see that we are in a situation quite similar to that for M-estimators from Section 3.1. In particular, the solution of $\lambda(t, P) = 0$ in t does not have to be unique, and if J is discontinuous, it does not even have to exist. We could, however, adapt the same approach as for M-estimators in Section 3.1.2 and define the statistical functional $T(P)$ as any value in the interval $[T^*(P), T^{**}(P)]$, where

$$T^*(P) = \sup\{t \in \mathbb{R}\colon \lambda(t, P) > 0\},$$
$$T^{**}(P) = \inf\{t \in \mathbb{R}\colon \lambda(t, P) < 0\}.$$

Exactly the same argumentation as we used in Section 3.1.3 for M-functionals gives that the maximum positive ε-bias $b_+(\varepsilon)$ in the Lévy metric d_L of an R-functional is attained at the improper density F_1 from (67), see also Figure 8. Thus, we must establish $\lambda(t, F_1)$ for

$$F_1(x) = \begin{cases} 0 & \text{for } x < x_0 + \varepsilon, \\ F_0(x - \varepsilon) - \varepsilon & \text{for } x \geq x_0 + \varepsilon, \end{cases}$$

where $x_0 = F_0^{-1}(\varepsilon)$.

First, we know from (90) that the quantiles of F_1 take the form

$$F_1^{-1}(s) = \inf\{x \in \mathbb{R}\colon F_1(x) \geq s\} = \inf\{y + \varepsilon \in \mathbb{R}\colon F_0(y) \geq s + \varepsilon\}$$
$$= \begin{cases} F_0^{-1}(s + \varepsilon) + \varepsilon & \text{if } s \in [0, 1 - \varepsilon], \\ \infty & \text{if } s \in (1 - \varepsilon, 1]. \end{cases}$$

Now, if both $s \leq 1 - \varepsilon$ and

$$2\,t - F_1^{-1}(s) \geq x_0 + \varepsilon, \tag{110}$$

we have that in the integrand of (109) we can write

$$F_1(2\,t - F_1^{-1}(s)) = F_0(2\,t - F_1^{-1}(s) - \varepsilon) - \varepsilon = F_0(2\,t - (F_0^{-1}(s + \varepsilon) + \varepsilon) - \varepsilon) - \varepsilon$$
$$= F_0(2\,(t - \varepsilon) - F_0^{-1}(s + \varepsilon)) - \varepsilon.$$

If $s > 1 - \varepsilon$, we have $F_1^{-1}(s) = \infty$, and

$$F_1(2\,t - F_1^{-1}(s)) = 0. \tag{111}$$

If (110) is not true, then (111) is also valid. It remains to realise that (110) is equivalent with

$$2\,t - x_0 - \varepsilon \geq F_1^{-1}(s) = F_0^{-1}(s + \varepsilon) + \varepsilon,$$

that is (assuming, for simplicity, that F_0 is strictly increasing everywhere)

$$s \leq F_0(2\,t - x_0 - 2\,\varepsilon) - \varepsilon.$$

Since this condition already implies $s \leq 1 - \varepsilon$, we can put all our formulas together and write

$$\begin{aligned}\lambda(t, F_1) = &\int_0^{s_0} J\left(\frac{1}{2}\left(s + 1 + \varepsilon - F_0(2\,(t - \varepsilon) - F_0^{-1}(s + \varepsilon))\right)\right) \mathrm{d}\,s \\ &+ \int_{s_0}^1 J\left(\frac{1}{2}\,(s+1)\right) \mathrm{d}\,s,\end{aligned} \tag{112}$$

where $s_0 = \max\{0, F_0(2\,t - x_0 - 2\,\varepsilon) - \varepsilon\}$. Using the same argument as for M-estimators in Section 3.1.3, we get

$$b_+(\varepsilon) = \inf\{t \in \mathbb{R} \colon \lambda(t, F_1) < 0\},$$

and for F_0 symmetric, this is also the maximum bias $b(\varepsilon, F_0, T)$.

To find the asymptotic breakdown point of T we need to find $\lim_{t\to\infty} \lambda(t, F_1)$. We see in (112) that as $t \to \infty$, then $s_0 \to 1 - \varepsilon$. Further,

$$\begin{aligned}\lim_{t\to\infty} \lambda(t, F_1) &= \int_0^{1-\varepsilon} J\left(\frac{1}{2}\,(s + 1 + \varepsilon - 1)\right) \mathrm{d}\,s + \int_{1-\varepsilon}^1 J\left(\frac{1}{2}\,(s+1)\right) \mathrm{d}\,s \\ &= 2\left(\int_{\varepsilon/2}^{1/2} J(s)\,\mathrm{d}\,s + \int_{1-\varepsilon/2}^1 J(s)\,\mathrm{d}\,s\right).\end{aligned}$$

We now use the symmetry of J from (104) to finalise

$$\lim_{t\to\infty} \lambda(t, F_1) = 2\left(\int_{1-\varepsilon/2}^1 J(s)\,\mathrm{d}\,s - \int_{1/2}^{1-\varepsilon/2} J(s)\,\mathrm{d}\,s\right).$$

We know that T breaks down at F_1 if and only if $\lim_{t\to\infty} \lambda(t, F_1) > 0$. This means that the asymptotic breakdown point of T must be $\varepsilon^* = \varepsilon^*(F_0, T)$ defined by

$$\int_{1-\varepsilon^*/2}^1 J(s)\,\mathrm{d}\,s = \int_{1/2}^{1-\varepsilon^*/2} J(s)\,\mathrm{d}\,s. \tag{113}$$

Example 3.21. For the Hodges-Lehmann estimator T from Example 3.18 we have for $x \in (1/2, 1]$

$$\int_{1/2}^x J(s)\,\mathrm{d}\,s = \frac{(1-2x)^2}{8},$$

which equals $1/16 = (\int_{1/2}^{1} J(s)\,\mathrm{d}\,s)/2$ if $x = 1 - \varepsilon^*/2 = (2+\sqrt{2})/4$, which gives

$$\varepsilon^*(F_0, T) = 1 - \frac{1}{\sqrt{2}} \approx 0.293$$

for any symmetric distribution F_0. This should be compared with the influence function of T, which was shown to be unbounded from above in Example 3.20 if all symmetric distributions are considered. Thus, even though the Hodges-Lehmann estimator is not robust in the sense of its influence function, it possesses a rather high positive breakdown point. △

It remains to inspect the qualitative robustness, or weak continuity, of the R-functionals. We do it again as for L-functionals in Section 3.2.2, and study the maximum bias determined by (112) as $\varepsilon \to 0$. The constant $x_0 = F_0^{-1}(\varepsilon)$ converges, as $\varepsilon \to 0$, to the lower endpoint ℓ of the support of F_0. Then, s_0 converges to 1 if either $\ell = -\infty$, or if F_0 is symmetric. Similarly as we argued for L-estimators, both F_0 and F_0^{-1} have at most countably many points of discontinuity because they are monotone. At the same time, we assumed that also J is monotone, meaning that it also has at most countably many discontinuity points. That means that the integrand function from the integral in (112)

$$J\left(\frac{1}{2}\left(s + 1 + \varepsilon - F_0(2\,(t-\varepsilon) - F_0^{-1}(s+\varepsilon))\right)\right)$$

converges to

$$J\left(\frac{1}{2}\left(s + 1 - F_0(2\,t - F_0^{-1}(s))\right)\right)$$

for almost all points $s \in \mathbb{R}$, for any $t \in \mathbb{R}$, as $\varepsilon \to 0$. The limit function is the integrand of $\lambda(t, F_0)$. By Lemma 4 we also know that this convergence is monotone decreasing as $\varepsilon \to 0$. Thus, the monotone convergence theorem [23, Theorem 4.7] gives that also the integrals converge as $\varepsilon \to 0$, provided that $t \in \mathbb{R}$ is a continuity point of $\lambda(\cdot, F_0)$.

We obtain that the sequence of monotone non-increasing (in t) functions $g(t, \varepsilon) = \lambda(t, F_1)$ converge to $g(t, 0) = \lambda(t, F_0)$ as $\varepsilon \to 0$ at all continuity points t of $g(t, 0)$. We ask if this is enough for the sequence of roots of $g(t, \varepsilon)$ to converge to the root of $g(t, 0)$. Following the same argumentation as for the weak convergence of measures and the corresponding weak continuity of quantiles (see Example 3.9) we get that the R-functionals are weakly continuous at F_0 if and only if $T(F_0)$ is uniquely defined, meaning that $T(F_0)$ is the unique point $t \in \mathbb{R}$ such that $\lambda(t, F_0) = 0$.

We can now summarise all our findings in a theorem.

Theorem 14. *Let $J\colon [0,1] \to \mathbb{R}$ be non-decreasing, symmetric as in (104) and satisfying the integrability condition (101). Suppose that F corresponding to $P \in \mathcal{P}(\mathbb{R})$ is either symmetric or supported in $\mathbb{R}$. If the R-functional $T(P)$ given as a solution to (108) is uniquely defined, then T is weakly continuous at P. The asymptotic breakdown point $\varepsilon^* = \varepsilon^*(P, T)$ of T is given by (113).*

3.4 Asymptotic efficiency of estimators

Suppose that we have a parametric model $\mathcal{F} = \{P_\theta \colon \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$. We are given a statistical functional T that is Fréchet differentiable and Fisher consistent for $\mathcal{F}$. Using Theorem 5, we know that under certain conditions, the estimator $T_n = T(P_n)$ based on an empirical measure $P_n \in \mathcal{P}(\mathcal{X})$ sampled from P_θ is asymptotically normal. We can express

$$\sqrt{n}\,(T(P_n) - T(P_\theta)) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}(0, A(P_\theta, T)), \tag{114}$$

where, by (32), we have

$$A(P_\theta, T) = \int_{\mathcal{X}} (\mathsf{IF}(x, P_\theta, T))^2 \,\mathrm{d}\, P_\theta(x).$$

In [19, Theorem 3], we studied the Rao-Cramér bound, giving a lower estimate for the variance of an unbiased estimator. We have shown that no (regular enough) unbiased estimator of θ can have a variance lower than the inverse Fisher information of θ. As we show in the following theorem, a similar bound also applies to the asymptotic variance $A(P_\theta, T)$ of a statistical functional.

Theorem 15. *Suppose that $\mathcal{F} = \{P_\theta \colon \theta \in \Theta\}$ is a parametric model with $\Theta \subseteq \mathbb{R}$, and let T be a statistical functional. Suppose that the following conditions are true for every $\theta \in \Theta$*

(C_1) The functional T is Fréchet differentiable (with respect to the Prokhorov metric) at P_θ.

(C_2) The functional T is Fisher consistent for θ, i.e. $T(P_\theta) = \theta$.

(C_3) $d_P(P_\theta, P_{\theta+\delta}) = O(\delta)$ as $\delta \to 0$.

(C_4) The model $\mathcal{F}$ corresponds to a regular system of densities $\{f(\cdot, \theta) \colon \theta \in \Theta\}$ with respect to a σ-finite measure μ on $\mathcal{X}$. The common support of all these densities is denoted by $M = \{x \in \mathcal{X} \colon f(x, \theta) > 0\}$, and the Fisher information of the system $\mathcal{F}$ is denoted by $J(\theta) \in (0, \infty)$ for $\theta \in \Theta$.

(C_5) We can write

$$\lim_{\delta\to 0} \int_{\mathcal{X}} \mathsf{IF}(x, P_\theta, T) \left(\frac{f(x, \theta+\delta) - f(x, \theta)}{\delta} \right) \mathrm{d}\,\mu(x) = \int_{\mathcal{X}} \mathsf{IF}(x, P_\theta, T) \frac{\partial}{\partial\theta} f(x, \theta) \,\mathrm{d}\,\mu(x).$$

Then the asymptotic variance $A(P_\theta, T)$ from (114) satisfies

$$A(P_\theta, T) \geq \frac{1}{J(\theta)} \quad \text{for all } \theta \in \Theta, \tag{115}$$

with equality if and only if for μ-almost all $x \in \mathcal{X}$

$$\mathsf{IF}(x, P_\theta, T) = \frac{1}{J(\theta)} \frac{\partial}{\partial\theta} \log(f(x, \theta)). \tag{116}$$

Proof. By the definition (23) of the Fréchet derivative L at P_θ we have as $\delta \to 0$

$$T(P_{\theta+\delta}) - T(P_\theta) - L(P_{\theta+\delta} - P_\theta) = o(d_P(P_{\theta+\delta}, P_\theta)). \tag{117}$$

The Fréchet derivative L can be expressed using Theorem 3 and (33) as

$$\begin{aligned} L(P_{\theta+\delta} - P_\theta) &= \int_{\mathcal{X}} \mathsf{IF}(x, P_\theta, T) \,\mathrm{d}\, (P_{\theta+\delta} - P_\theta)(x) \\ &= \int_{\mathcal{X}} \mathsf{IF}(x, P_\theta, T) \,(f(x, \theta+\delta) - f(x,\theta)) \,\mathrm{d}\,\mu(x). \end{aligned}$$

We plug this into (117) and use conditions (C_2) and (C_3) to get

$$\begin{aligned} o(\delta) &= \delta - \int_{\mathcal{X}} \mathsf{IF}(x, P_\theta, T) \,(f(x, \theta+\delta) - f(x,\theta)) \,\mathrm{d}\,\mu(x) \\ &= \delta - \int_{M} \mathsf{IF}(x, P_\theta, T) \,(f(x, \theta+\delta) - f(x,\theta)) \,\mathrm{d}\,\mu(x). \end{aligned}$$

Divide both sides by $\delta > 0$ to get

$$\int_M \mathsf{IF}(x, P_\theta, T) \frac{f(x,\theta+\delta) - f(x,\theta)}{\delta\, f(x,\theta)} f(x,\theta) \,\mathrm{d}\,\mu(x) = 1 + o(\delta)/\delta.$$

Now, take $\delta \to 0$ and apply condition (C_5) to obtain

$$\begin{aligned} 1 &= \int_M \mathsf{IF}(x, P_\theta, T) \left(\frac{\partial}{\partial\theta} \log(f(x,\theta))\right) f(x,\theta) \,\mathrm{d}\,\mu(x) \\ &= \int_{\mathcal{X}} \mathsf{IF}(x, P_\theta, T) \left(\frac{\partial}{\partial\theta} \log(f(x,\theta))\right) f(x,\theta) \,\mathrm{d}\,\mu(x). \end{aligned}$$

It remains to use the Cauchy-Schwarz inequality [7, Corollary 5.1.4] in the L_2-Hilbert space given by the measure $f(x,\theta)\,\mathrm{d}\,\mu(x)$ to get

$$\begin{aligned} 1 &\le \left(\int_{\mathcal{X}} (\mathsf{IF}(x, P_\theta, T))^2 \, f(x,\theta) \,\mathrm{d}\,\mu(x) \int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta} \log(f(x,\theta))\right)^2 f(x,\theta) \,\mathrm{d}\,\mu(x) \right)^{1/2} \\ &= \sqrt{A(P_\theta, T)\, J(\theta)} \end{aligned}$$

as we wanted to prove. In the Cauchy-Schwarz lemma, we reach equality if and only if the two functions are linearly dependent, meaning that there exists a constant $c(\theta) > 0$ such that

$$\mathsf{IF}(x, P_\theta, T) = c(\theta) \,\frac{\partial}{\partial\theta} \log(f(x,\theta)) \quad \text{for } \mu\text{-almost all } x \in \mathcal{X}. \tag{118}$$

We take a square of the previous formula and integrate both sides with respect to the measure $f(x,\theta)\,\mathrm{d}\,\mu(x)$ to get

$$A(P_\theta, T) = c(\theta)^2 J(\theta).$$

In the case of equality in (115), the left-hand side equals $1/J(\theta)$, which gives $c(\theta) = 1/J(\theta)$ in (118). □

The conditions of Theorem 15 are quite natural. Condition (C_3) gives a connection between the chosen topology on the space of measures and the parametrisation of the model $\mathcal{F}$. Note that if $\mathcal{X} = \mathbb{R}$, we could equally use the Lévy distance instead of the Prokhorov distance in both (C_1) and (C_3). Condition (C_4) only requires that the Fisher information is well-defined, and the technical condition (C_5) concerns the possibility of interchanging a limit with an integral.

Together with Theorem 5, the inequality in Theorem 15 is quite strong. As a special case, it guarantees the asymptotic optimality of the maximum likelihood estimators from Examples 3.1 and 3.5. For them, we know that the optimal bound in (115) is attained by [19, Theorem 23], or our formula (56). More generally, Theorem 15 should be seen as complementary to the standard Rao-Cramér bound from [19, Theorem 3] — even if the estimator T_n fails to be unbiased, under mild conditions its (asymptotic) variance cannot be lower than the Rao-Cramér bound (115).

Theorem 15 does not, however, say that only maximum likelihood estimators are asymptotically efficient. It can be applied to any (Fréchet differentiable) statistical functional T. Formula (116) can be then used to design estimators that are asymptotically optimal at a given parametric model. We apply these results to M, L, and R-estimators:

- **M-estimators:** A general solution to (116) is the M-estimator given by the maximum likelihood equations, that is the M-estimator (49) with

$$\psi(x, \theta) = \frac{\partial}{\partial \theta} \log(f(x, \theta)).$$

  We already saw this in Example 3.5.

- **L-estimators:** We consider only the location model $\mathcal{F}$ defined by a system of densities $\{f(\cdot, \theta) = f_0(\cdot - \theta) \colon \theta \in \mathbb{R}\}$ for f_0 given, and the identity function $h(x) = x$ in the L-functional (78). From the expression (85) for the influence function of T we get

$$\frac{\partial}{\partial x} \mathsf{IF}(x, P_\theta, T) = m(F_\theta(x)),$$

  plugging this into (116) we get

$$m(F_\theta(x)) = \frac{1}{J(\theta)} \frac{\partial^2}{\partial \theta \partial x} \log(f_0(x - \theta)) = -\frac{1}{J(\theta)} \frac{\partial^2}{\partial t^2} \left[\log(f_0(t))\right]_{t=x-\theta},$$

  that is

$$m(F_0(x)) = -\frac{1}{J(0)} \frac{\partial^2}{\partial x^2} \log(f_0(x)). \tag{119}$$

We used that $F_\theta(x) = F_0(x-\theta)$ for all $x \in \mathbb{R}$ and $\theta \in \mathbb{R}$, and

$$\begin{aligned} J(\theta) &= \int_{\mathbb{R}} \left(\frac{\partial}{\partial \theta} \log(f_0(x-\theta)) \right)^2 f_0(x-\theta) \, \mathrm{d}\, \mu(x) \\ &= \int_{\mathbb{R}} \left(\frac{\partial}{\partial t} \left[\log(f_0(t)) \right]_{t=x-\theta} \right)^2 f_0(x-\theta) \, \mathrm{d}\, \mu(x) \\ &= \int_{\mathbb{R}} \left(\frac{\partial}{\partial t} \left[\log(f_0(t)) \right]_{t=x} \right)^2 f_0(x) \, \mathrm{d}\, \mu(x) = J(0), \end{aligned}$$

which is true in location models $\mathcal{F}$.

- **R-estimators:** Consider again only the location model $\mathcal{F}$ as for the L-estimators, and take only F_0 symmetric with density f_0. In that case we have by (106)

$$\mathsf{IF}(x, P_\theta, T) = \frac{J(F_\theta(x))}{\int_{\mathbb{R}} J'(F_\theta(y)) \, (f_\theta(y))^2 \, \mathrm{d}\, y}.$$

The denominator is a constant in $x \in \mathbb{R}$, meaning that if (116) is to be true, we should have

$$J(F_0(x)) = c \frac{\partial}{\partial x} \log(f_0(x))$$

for some constant $c \neq 0$.

We conclude by giving asymptotically optimal M, L, and R-estimators in several important location models.

Example 3.22. For the location model $\mathcal{F} = \{\mathsf{N}(\theta, 1) \colon \theta \in \mathbb{R}\}$ we have

$$f_0(x) = \frac{1}{\sqrt{2\,\pi}} \exp(-x^2/2),$$

and the distribution function $F_0 = \Phi$, which gives the optimal estimators corresponding to

- **M-estimators:**

$$\psi(x, \theta) = x - \theta,$$

which is the mean M-functional;

- **L-estimators:**

$$m(\Phi(x)) = 1 \quad \text{for } x \in \mathbb{R},$$

that is $m(s) = 1$ for $s \in [0, 1]$, and we again get the mean as an L-functional;

- **R-estimators:**

$$J(\Phi(x-\theta)) = x - \theta \quad \text{for } x \in \mathbb{R},$$

that is $J(s) = \Phi^{-1}(s)$ for $s \in [0, 1]$. This R-estimator T corresponds to the normal scores R-estimator from Example 3.19. Unlike the M and L-estimators for the normal model,

this optimal R-estimator is robust. Indeed, applying Theorem 14 and (113), we will calculate in Example 3.25 below that T has a strictly positive asymptotic breakdown point $\varepsilon^*(F,T) = 2\,\Phi(-\sqrt{\log(4)}) \approx 0.239$ for any F symmetric.

△

Example 3.23. Take the location model $\mathcal{F} = \{F_0(\cdot - \theta) \colon \theta \in \mathbb{R}\}$ with the logistic distribution given by its distribution function

$$F_0(x) = \frac{1}{1+\exp(-x)} \quad \text{for } x \in \mathbb{R}.$$

Its inverse is

$$F_0^{-1}(s) = \log\left(\frac{s}{1-s}\right) \quad \text{for } s \in (0,1).$$

The density of F_0 is

$$f_0(x) = \frac{\exp(-x)}{(1+\exp(-x))^2} \quad \text{for } x \in \mathbb{R},$$

and its log-derivatives are for $x \in \mathbb{R}$

$$\begin{aligned} \frac{\partial}{\partial x}\log(f_0(x)) &= \frac{1-\exp(x)}{1+\exp(x)}, \\ \frac{\partial^2}{\partial x^2}\log(f_0(x)) &= \frac{-2\,\exp(x)}{(1+\exp(x))^2}. \end{aligned} \tag{120}$$

The optimal estimators are then given by

- **M-estimators:**
$$\psi(x,\theta) = \frac{1-\exp(x-\theta)}{1+\exp(x-\theta)},$$
which gives a robust M-estimator of location. By Theorem 10 this estimator is always uniquely defined, weakly continuous, and its asymptotic breakdown point is $\varepsilon^*(P,T) = 1/2$ for each $P \in \mathcal{P}(\mathbb{R})$.

- **L-estimators:** We have the Fisher information $J(0) = 1/3$ and
$$m(F_0(x)) = 3\,\frac{2\,\exp(x)}{(1+\exp(x))^2} \quad \text{for } x \in \mathbb{R},$$
that is
$$m(s) = 6\,s\,(1-s) \quad \text{for } s \in [0,1].$$
This generates an L-functional that is not qualitatively robust, by Theorem 13.

- **R-estimators:** For the function J defining an R-functional (100) we have the condition
$$J(F_0(x)) = \frac{1-\exp(x)}{1+\exp(x)} \quad \text{for } x \in \mathbb{R},$$

which gives $J(s) = 2\,s - 1$ for $s \in [0,1]$. After appropriate scaling so that the assumptions of Theorem 14 are met, we obtain

$$J(s) = s - 1/2 \quad \text{for } s \in [0,1].$$

That is the Hodges-Lehmann estimator from Example 3.18.

△

Example 3.24. Take the location model $\mathcal{F} = \{F_0(\cdot - \theta) \colon \theta \in \mathbb{R}\}$ with the Laplace distribution function (73) and density

$$f_0(x) = \exp(-|x|)/2 \quad \text{for } x \in \mathbb{R}.$$

Its log-derivative is

$$\frac{\partial}{\partial x} \log(f_0(x)) = -\operatorname{sign}(x) \quad \text{for } x \neq 0,$$

and is undefined for $x = 0$. The asymptotically efficient estimators are

- **M-estimators:** $\psi(x,\theta) = -\operatorname{sign}(x - \theta)$, which gives the median functional.
- **L-estimators:** The efficient L-functional cannot be defined directly by formula (119) because the second log-derivative of f_0 is constant zero.
- **R-estimators:** $J(F_0(x)) = -\operatorname{sign}(x)$ for $x \neq 0$, which gives the score-generating function J as in Example 3.17, and leads again to the median functional.

△

It is interesting to compare two of the estimators derived above in the following examples.

Example 3.25. By the expression for influence functions of R-estimators from (106) we have for P standard normal that $\mathsf{IF}(x, P, T)$ is proportional to $J(\Phi(x)) = \Phi^{-1}(\Phi(x)) = x$ for $x \in \mathbb{R}$, and $\gamma^*(P,T) = \infty$ because the influence function is unbounded. Let us now compute the asymptotic breakdown point of T. By the expression for the asymptotic breakdown point of R-estimators (113) and Theorem 14 we have that $1 - \varepsilon^*(P,T)/2 = 1 - \varepsilon^*/2$ is given as a solution $x \in [1/2, 1]$ to the equation

$$\int_{1/2}^{x} J(s)\,\mathrm{d}\,s = \int_{x}^{1} J(s)\,\mathrm{d}\,s,$$

which is in our situation

$$\int_{1/2}^{x} \Phi^{-1}(s)\,\mathrm{d}\,s = \int_{x}^{1} \Phi^{-1}(s)\,\mathrm{d}\,s.$$

To get this, we need to find

$$\xi(x) = \int_{1/2}^{x} \Phi^{-1}(s)\,\mathrm{d}\,s \quad \text{for } x \in [1/2, 1].$$

First, one passes from the integral of an inverse function $\Phi^{-1}(s)$ to the integral of $\Phi(s)$ using the Laisant formula.[5] Then, integration by parts gives

$$\xi(x) = \frac{1 - \exp\left(-(\Phi^{-1}(x))^2/2\right)}{\sqrt{2\,\pi}} \quad \text{for } x \in [1/2, 1].$$

We have $\Phi^{-1}(1) = \infty$, giving $\xi(1) = (2\,\pi)^{-1/2}$ and solving

$$\xi(1 - \varepsilon^*/2) = \frac{1}{2\,\sqrt{2\,\pi}}$$

gives the asymptotic breakdown point

$$\varepsilon^*(P, T) = 2\,\Phi\left(-\sqrt{\log(4)}\right) \approx 0.239.$$

This result is surprising; even though for the gross error sensitivity we have $\gamma^*(P, T) = \infty$, for the asymptotic breakdown point we found $\varepsilon^*(P, T) \approx 0.239$. That seems counter-intuitive because of formula (39) and Definition 5 of the asymptotic breakdown point. This example emphasises the inherent difference between infinitesimal robustness and robustness in neighbourhoods. △

In the next example, we show it is also possible to have a functional that is infinitesimally robust, but not robust in a neighbourhood of P.

Example 3.26. Take the efficient L-estimator for P the logistic distribution. By formula (85), its influence function is proportional to the derivative of the log-density f_0 from (120), which is

$$\frac{1 - \exp(x)}{1 + \exp(x)} \quad \text{for } x \in \mathbb{R}.$$

This function is bounded in x, and thus $\gamma^*(P, T)$ is finite. On the other hand, the support of the measure M defining T is the whole interval $[0, 1]$, and thus the asymptotic breakdown point of T is zero, because of Theorem 13. △

4 Minimax optimal estimation of location

There are two major approaches to the construction of optimal robust estimators:

[5] https://en.wikipedia.org/wiki/Integral_of_inverse_functions

- **Hampel's approach** based on the concept of infinitesimal robustness, and the influence function from Section 2.2. This approach intends to find functionals T that, for a given distribution $P_0 \in \mathcal{P}(\mathcal{X})$, minimise the asymptotic variance $A(P_0, T)$, under the condition that their gross error sensitivity (that is, the maximum absolute value of their influence function) $\gamma^*(P_0, T)$ is bounded. The bound on the gross error sensitivity corresponds to imposing robustness on T. Originally, this idea was pursued by Hampel; in detail, it is treated in [10].

- **Huber's approach** that takes a fixed distribution $P_0 \in \mathcal{P}(\mathcal{X})$ and an appropriate neighbourhood $\mathcal{P}_\varepsilon(P_0)$ of P_0 in $\mathcal{P}(\mathcal{X})$. The task is to find a functional T (and the corresponding estimators) that minimises either the maximum bias $b(\varepsilon, P_0, T)$, or the maximum variance $v(\varepsilon, P_0, T) = \sup_{P \in \mathcal{P}_\varepsilon(P_0)} A(P, T)$ introduced in Definition 4 over all (regular enough) estimators T.

We already saw that the two paradigms do share similarities, but are inherently different. The advantage of Huber's approach is that we indeed seek for robustness guaranteed in full neighbourhoods of our target measure P_0. As we will see, its disadvantages are the mathematical complexity, and the fact that we are able to obtain explicit solutions only in relatively simple models. Hampel's approach applies to much more general situations, and is somewhat easier to handle, but deals only with the infinitesimal robustness of T.

In the present section we introduce the basics of Huber's approach in the special situation of location estimation in $\mathbb{R}$. We suppose that we are given a random sample $X_1, \ldots, X_n$ from an (assumed, ideal) distribution $P_0 \in \mathcal{P}(\mathbb{R})$ with distribution function F_0, and we intend to estimate the location parameter $\theta \in \mathbb{R}$ given by the parametric model $\mathcal{F}$ of distribution functions

$$\mathcal{F} = \{F_0(\cdot - \theta) \colon \theta \in \mathbb{R}\}. \tag{121}$$

In this model, if X is a random variable with distribution F_0 we clearly have that

$$\mathsf{P}(X + \theta \leq x) = \mathsf{P}(X \leq x - \theta) = F_0(x - \theta) \quad \text{for all } x \in \mathbb{R},$$

meaning that the random variable corresponding to $F_0(\cdot - \theta)$ is, in fact, just X shifted by a constant $\theta \in \mathbb{R}$. For that reason, it is natural to impose that for any functional $T = T(X) = T(F)$ estimating θ in a location model we have

$$T(X + \theta) = T(X) + \theta \quad \text{for all } \theta \in \mathbb{R}. \tag{122}$$

This is precisely the translation equivariance of T that we observed to be true for (i) M-estimators of location in Section 3.1.3; (ii) L-estimators of location in Section 3.2.2; and (iii) R-estimators in Section 3.3.2. From now on, we therefore assume in this section that the functional T satisfies (122).

We begin by discussing the Huber's approach to the minimax bias estimation in Section 4.1. Then, in Section 4.2 we treat the more challenging situation of minimising maximum asymptotic variance of T.

4.1 Minimax bias estimation

We want to estimate a location parameter, where the model is given by $\mathcal{F} = \{F_0(\cdot - \theta) \colon \theta \in \mathbb{R}\}$. To have the location of F_0 well defined, we assume that F_0 is symmetric in the sense $F_0(-x) = 1 - F_0(x)$ for all $x \in \mathbb{R}$, with a unimodal density f_0. The latter means that $f_0(x)$ is non-increasing in $x \geq 0$. Under these conditions, the origin is certainly the most reasonable choice for the parameter θ.

We will minimise the maximum bias of a location equivariant functional T in the contamination neighbourhood

$$\mathcal{P}_\varepsilon(P_0) = \{(1-\varepsilon)\, P_0 + \varepsilon\, Q \colon Q \in \mathcal{P}(\mathbb{R})\}\,, \tag{123}$$

where $P_0 \in \mathcal{P}(\mathbb{R})$ is the measure corresponding to F_0. Just as we saw in Section 3.2.2 for general L-estimators, the maximum positive bias of the median functional T_{med} in the neighbourhood $\mathcal{P}_\varepsilon(P_0)$ must be caused by placing the whole ε-mass of Q to the point at positive infinity. By symmetry, the maximum absolute bias of T_{med} is the solution $x_0 = x_0(\varepsilon)$ to the equation

$$(1-\varepsilon)\, F_0(x_0) = 1/2,$$

that is

$$x_0(\varepsilon) = F_0^{-1}\left(\frac{1}{2(1-\varepsilon)}\right). \tag{124}$$

Note that because of our assumption of the existence of unimodal density of F_0, the distribution function F_0 is strictly increasing on its support, and applying F_0^{-1} is legitimate for any $\varepsilon \in (0, 1/2)$.

Consider now the function

$$f_+(x) = \begin{cases} (1-\varepsilon)\, f_0(x) & \text{for } x \leq x_0(\varepsilon), \\ (1-\varepsilon)\, f_0(x - 2\, x_0(\varepsilon)) & \text{for } x > x_0(\varepsilon). \end{cases}$$

First, we verify that f_+ is a density that corresponds to a distribution function $F_+ \in \mathcal{P}_\varepsilon(P_0)$.

To do that, compute

$$\begin{aligned}\int_{\mathbb{R}} f_+(x)\,\mathrm{d}\,x &= (1-\varepsilon)\left(\int_{-\infty}^{x_0(\varepsilon)} f_0(x)\,\mathrm{d}\,x + \int_{x_0(\varepsilon)}^{\infty} f_0(x-2\,x_0(\varepsilon))\,\mathrm{d}\,x\right)\\ &= (1-\varepsilon)\left(F_0(x_0(\varepsilon)) + \int_{-x_0(\varepsilon)}^{\infty} f_0(x)\,\mathrm{d}\,x\right)\\ &= (1-\varepsilon)\,(F_0(x_0(\varepsilon)) + 1 - F_0(-x_0(\varepsilon))) = 2\,(1-\varepsilon)\,F_0(x_0(\varepsilon))\\ &= 2\,(1-\varepsilon)\frac{1}{2(1-\varepsilon)} = 1.\end{aligned}$$

We used the symmetry of F_0 and (124). It remains to verify that $F_+(x) = \int_{-\infty}^{x} f_+(s)\,\mathrm{d}\,s$ is an element of $\mathcal{P}_\varepsilon(F_0)$. For that we have

$$f_+(x) - (1-\varepsilon)\,f_0(x) = \begin{cases} 0 & \text{for } x \le x_0(\varepsilon),\\ (1-\varepsilon)\;(f_0(x-2\,x_0(\varepsilon)) - f_0(x)) & \text{for } x > x_0(\varepsilon).\end{cases}$$

This difference has a total mass

$$\int_{\mathbb{R}} (f_+(x) - (1-\varepsilon)\,f_0(x))\;\mathrm{d}\,x = 1 - (1-\varepsilon) = \varepsilon.$$

We obtain that F_+ can indeed be written as $(1-\varepsilon)\,F_0 + \varepsilon\,G$ with G a distribution function supported in the interval $[x_0(\varepsilon), \infty)$. The situation is visualised in Figure 12. Because of the symmetry of f_0 around the origin, the density f_+ is symmetric around $x_0(\varepsilon)$.

Take now the distribution $F_+ \in \mathcal{P}_\varepsilon(F_0)$ and its version F_- shifted to the left

$$F_-(x) = F_+(x + 2\,x_0(\varepsilon)) \quad \text{for } x \in \mathbb{R}.$$

By obvious symmetry considerations, also $F_- \in \mathcal{P}_\varepsilon(F_0)$. Take now any translation equivariant functional T. Since F_+ is a translation of F_- by $2\,x_0(\varepsilon)$, we get that

$$T(F_+) = T(F_-) + 2\,x_0(\varepsilon).$$

Therefore, no matter what value does $T(F_0)$ take, one of the numbers $T(F_-)$ or $T(F_+)$ must be at least $x_0(\varepsilon)$-far from $T(F_0)$. In particular, the maximum bias of any translation equivariant functional in $\mathcal{P}_\varepsilon(F_0)$ cannot be lower than $x_0(\varepsilon)$, which is the maximum bias of T_{med}. We obtain the following theorem.

Theorem 16. *In a location model $\mathcal{F} = \{F_0(\cdot - \theta)\colon \theta \in \mathbb{R}\}$ with F_0 symmetric with unimodal density and the contamination neighbourhood (123), the smallest maximum bias $b(\varepsilon, F_0, T)$ among all location equivariant functionals T is equal to (124). This maximum bias is attained by the median functional T_{med}.*

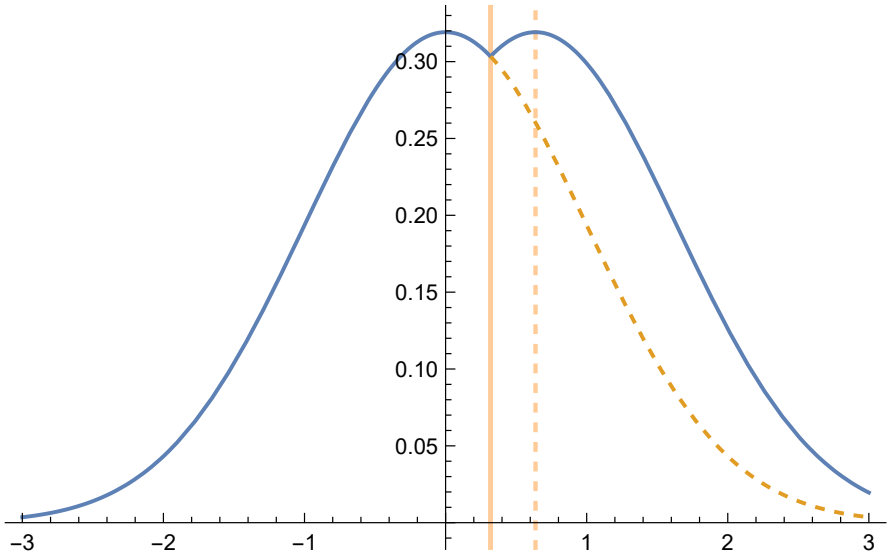

Figure 12: The density f_+ (blue) constructed in Section 4.1 for F_0 the standard normal distribution and $\varepsilon = 0.2$. The density is symmetric around $x_0(\varepsilon)$ (thick orange vertical line) and bimodal with modes at $x = 0$ and $x = 2\,x_0(\varepsilon)$ (thin vertical dashed line). The mixture distribution G corresponds to the excess mass above the dashed part of the bell curve in the interval $[x_0(\varepsilon), \infty)$.

In Theorem 16, we deal with contamination neighbourhoods. The same approach, however, works also for the Lévy neighbourhood of F_0, with the only exception that by (91) with $M = \delta_{1/2}$ we know that

$$x_0(\varepsilon) = F_0^{-1}(1/2 + \varepsilon) + \varepsilon.$$

We see that minimising the maximum bias (in location models) is simple; the median functional is always the optimal choice. This, of course, means only that the median is optimal in terms of its maximum bias. Still, it must be considered that the (asymptotic) variance of the median might be too high for the estimator to be useful in practice.

4.2 Minimax variance estimation

Our task is to minimise the asymptotic maximum variance $v(\varepsilon, G, T)$ from Definition 4 in the contamination neighbourhood (123) around a fixed distribution $G \in \mathcal{P}(\mathbb{R})$. We are in the location model (121) with $F_0 = G$, and we search for the optimal robust functional T along the M-functionals from Section 3.1.

An M-functional $T(P)$ is given as a solution to

$$\int_{\mathbb{R}} \psi(x - T(P)) \,\mathrm{d}\, P(x) = 0, \tag{125}$$

see formula (49). By Theorems 8 and 5, under appropriate conditions, its asymptotic variance is

$$A(P,T) = \frac{\int_{\mathbb{R}} (\psi(x - T(P)))^2 \,\mathrm{d}\, P(x)}{\left(\int_{\mathbb{R}} \psi'(y - T(P) \,\mathrm{d}\, P(y)\right)^2} = \frac{\mathsf{E}_P \,(\psi(X - T(P)))^2}{(\mathsf{E}_P \,\psi'(X - T(P)))^2}.$$

Theorem 15 says that for any $F \in \mathcal{P}\,(\mathbb{R})$ with density f, the asymptotic variance $A(F,T)$ is the smallest for T based on the maximum likelihood estimator given by

$$\psi(x) = \frac{\partial}{\partial x} \log(f(x)) \quad \text{for } x \in \mathbb{R},$$

and for the asymptotic variance of this M-estimator we have

$$A(F,T) = 1/J(F)$$

for J the Fisher information of F. If we now find the element $F_0 \in \mathcal{P}_\varepsilon(G)$ which minimises the Fisher information, we obtain a lower bound in our problem — no estimator can then have minimax variance lower than $1/J(F_0)$.

In the first step, let us thus find the distribution F_0 that minimises the Fisher information in an ε-contamination neighbourhood of G given by

$$\mathcal{P}(G) = \{(1-\varepsilon)\,G + \varepsilon\,H \colon H \in \mathcal{P}\,(\mathbb{R})\}\,. \tag{126}$$

In our notation we now suppress ε since it is fixed, and for now, do not assume G to be symmetric.

As we will see, it will turn out that the maximum likelihood M-estimator corresponding to F_0 is the solution to our problem of minimax variance estimation in a neighbourhood of G. We will prove that in the second step of our endeavour.

4.2.1 Step 1: Distribution minimising Fisher information

We need to find $F_0 \in \mathcal{P}(G)$ from (126) that minimises the Fisher information. Our first observation is the following simple lemma.

Lemma 5. *Let $u, v \colon [0,1] \to \mathbb{R}$ be linear functions such that $v(t) > 0$ for all $t \in [0,1]$. Then $w(t) = u(t)^2/v(t)$ is convex in $[0,1]$.*

Proof. We need to verify that the second derivative of w is non-negative. Using $u''(t) = v''(t) = 0$ and denoting $u'(t) = a$, $v'(t) = b$ for all $t \in [0,1]$, it is easy to compute

$$w''(t) = \frac{2(u(t)\,b - v(t)\,a)^2}{v(t)^3} \geq 0$$

for all $t \in [0,1]$, as we wanted to show. □

Take now the Fisher information

$$J(F) = \int_{\mathbb{R}} \left(\frac{f'(x)}{f(x)} \right)^2 f(x) \, \mathrm{d}\, x = \int_{\mathbb{R}} \frac{(f'(x))^2}{f(x)} \, \mathrm{d}\, x.$$

Here, we write f is the density of F, and similarly we will denote f_0 the density of F_0 etc. Denoting $F_t = (1-t)\, F_0 + t\, F_1$, and writing f_t for its density, we get that

$$J(F_t) = \int_{\mathbb{R}} \frac{(f_t'(x))^2}{f_t(x)} \, \mathrm{d}\, x.$$

Now, as a function of $t \in [0,1]$, both

$$u(t) = f_t'(x) \quad \text{and} \quad v(t) = f_t(x)$$

are linear functions for each $x \in \mathbb{R}$, and $v(t) > 0$ on the support of F_t. Applying Lemma 5 we get that the integrand in the Fisher information is convex, and thus also $J(F_t)$ must be a convex function in $t \in [0,1]$. For a slightly stronger result see [4].

Knowing that the Fisher information is convex, we can find its minimum over the convex set (126) by considering its directional (Gâteaux) derivatives. The directional derivative of the functional J at measure F_0 in direction $F_1 \in \mathcal{P}(G)$ is given by

$$\begin{aligned} \frac{\partial}{\partial t} J(F_t) &= \int_{\mathbb{R}} \frac{\partial}{\partial t} \frac{(f_t'(x))^2}{f_t(x)} \, \mathrm{d}\, x = \int_{\mathbb{R}} \frac{\partial}{\partial t} \frac{((1-t)\, f_0'(x) + t\, f_1'(x))^2}{(1-t)\, f_0(x) + t\, f_1(x)} \, \mathrm{d}\, x \\ &= \int_{\mathbb{R}} \frac{1}{(f_t(x))^2} \Big(2\, (f_t'(x))(f_1'(x) - f_0'(x)) f_t(x) - (f_t'(x))^2 (f_1(x) - f_0(x)) \Big) \, \mathrm{d}\, x \end{aligned}$$

which simplifies for $t = 0$ to

$$\frac{\partial}{\partial t} \left[J(F_t) \right]_{t=0} = \int_{\mathbb{R}} 2\, \frac{f_0'(x)}{f_0(x)} (f_1'(x) - f_0'(x)) - \left(\frac{f_0'(x)}{f_0(x)} \right)^2 (f_1(x) - f_0(x)) \, \mathrm{d}\, x.$$

Denoting $\psi(x) = -f_0'(x)/f_0(x)$, we can rewrite this to

$$\begin{aligned} \frac{\partial}{\partial t} \left[J(F_t) \right]_{t=0} &= \int_{\mathbb{R}} -2\, \psi(x) (f_1'(x) - f_0'(x)) - (\psi(x))^2 \, (f_1(x) - f_0(x)) \, \mathrm{d}\, x \\ &= \int_{\mathbb{R}} \Big(2\, \psi'(x) - (\psi(x))^2 \Big) \, (f_1(x) - f_0(x)) \, \mathrm{d}\, x, \end{aligned} \tag{127}$$

where in the second equality we integrated by parts. We get that a distribution $F_0 \in \mathcal{P}(G)$ that minimises the Fisher information must satisfy

$$\int_{\mathbb{R}} \Big(2\, \psi'(x) - (\psi(x))^2 \Big) \, (f_1(x) - f_0(x)) \, \mathrm{d}\, x \geq 0$$

for any density f_1 of $F_1 \in \mathcal{P}(G)$. We obtain the following result.

Theorem 17. *A distribution $F_0 \in \mathcal{P}(G)$ minimises the Fisher information over all distributions in $\mathcal{P}(G)$ if and only if*

$$\int_{\mathbb{R}} \left(2\,\psi'(x) - (\psi(x))^2\right) \mathrm{d}\,(F_1 - F_0)(x) \geq 0$$

for all $F_1 \in \mathcal{P}(G)$, where $\psi(x) = -f_0'(x)/f_0(x)$ and f_0 is the density of F_0.

To find an optimum in Theorem 17 is a task of analysis of variations. We must effectively optimise over an infinite-dimensional space of densities to get F_1. This is not easy, and there are no simple solutions. One usually has to proceed using heuristic arguments, and employ some guesswork to find F_0. We give one important example where the solution in Theorem 17 can be found explicitly. Additional explicit solutions for different neighbourhoods and setups can be found in [12].

Theorem 18. *Let $G \in \mathcal{P}(\mathbb{R})$ be a log-concave distribution, meaning that G has a twice differentiable density $g\colon \mathbb{R} \to [0,\infty)$ such that $-\log(g)$ is a convex function on the support of G. Then the minimum Fisher information in the neighbourhood*

$$\mathcal{P}(G) = \{(1-\varepsilon)\,G + \varepsilon\,H \colon H \in \mathcal{P}(\mathbb{R})\}$$

is attained at the density

$$f_0(x) = \begin{cases} (1-\varepsilon)\,g(x_0)\exp(k(x-x_0)) & \text{for } x \leq x_0, \\ (1-\varepsilon)\,g(x) & \text{for } x \in [x_0, x_1], \\ (1-\varepsilon)\,g(x_1)\exp(-k(x-x_1)) & \text{for } x \geq x_1. \end{cases} \tag{128}$$

Here, $x_0 \leq x_1$ are the (possibly infinite) endpoints of the interval where $|g'/g| \leq k$, and $k \geq 0$ is given by

$$\int_{x_0}^{x_1} g(x)\,\mathrm{d}\,x + \frac{g(x_0)+g(x_1)}{k} = \frac{1}{1-\varepsilon}.$$

Proof. First we show that f_0 is indeed a density. It is certainly non-negative. Its integral is

$$\begin{aligned} &\frac{1}{1-\varepsilon}\int_{\mathbb{R}} f_0(x)\,\mathrm{d}\,x \\ &= \int_{-\infty}^{x_0} g(x_0)\exp(k(x-x_0))\,\mathrm{d}\,x + \int_{x_0}^{x_1} g(x)\,\mathrm{d}\,x + \int_{x_1}^{\infty} g(x_1)\exp(-k(x-x_1))\,\mathrm{d}\,x \\ &= \frac{g(x_0)+g(x_1)}{k} + \int_{x_0}^{x_1} g(x)\,\mathrm{d}\,x, \end{aligned}$$

meaning that f_0 indeed integrates to 1. Next, we prove that f_0 lies in the neighbourhood $\mathcal{P}(G)$. To do that, we consider

$$h(x) = \frac{f_0(x) - (1-\varepsilon)\, g(x)}{\varepsilon}$$
$$= \begin{cases} \frac{1-\varepsilon}{\varepsilon}(g(x_0)\exp(k(x-x_0)) - g(x)) & \text{for } x \le x_0, \\ 0 & \text{for } x \in [x_0, x_1], \\ \frac{1-\varepsilon}{\varepsilon}(g(x_1)\exp(-k(x-x_1)) - g(x)) & \text{for } x \ge x_1, \end{cases}$$

and show that h is a density of some $H \in \mathcal{P}(\mathbb{R})$. The integral of h is 1 because both g and f_0 also integrate to 1; it remains to show that h is non-negative. First, note that because $\xi = -\log(g)$ is convex, its derivative $\xi' = -g'/g$ is non-decreasing [22, Section 5.4], and by the way we defined k we have that $|g'(x_0)/g(x_0)| \ge k$ and $|g'(x_1)/g(x_1)| \ge k$. We want to show

$$g(x_1)\exp(-k(x-x_1)) \ge g(x) \quad \text{for } x \ge x_1,$$

which is equivalent with

$$\xi(x_1) + k(x-x_1) \le \xi(x) \quad \text{for } x \ge x_1,$$

where $\xi = -\log(g)$ is convex. This, however, follows from the fact that $\xi'(x_1) = |g'(x_1)/g(x_1)| \ge k$ and the fact that any convex function must lie above its tangent at any point [22, Lemma 5.4.9 and Example 5.5.15], that is

$$\xi(y) + \xi'(y)(x-y) \le \xi(x) \quad \text{for all } x, y \in \mathbb{R}.$$

The analogous inequality for $x \le x_0$ follows in the same way. We have shown that f_0 is a density from $\mathcal{P}(G)$.

Denote $\psi(x) = -f_0'(x)/f_0(x)$. Then we can write

$$\psi(x) = \begin{cases} -k & \text{for } x \le x_0, \\ -g'(x)/g(x) & \text{for } x \in (x_0, x_1), \\ k & \text{for } x \ge x_1. \end{cases}$$

We now verify that f_0 satisfies the condition from Theorem 17. Because $-\log(g)$ is convex, its derivative $-g'/g$ is non-decreasing. Thus, for any $x \in (x_0, x_1)$ we get

$$2\,\psi'(x) - (\psi(x))^2 = 2\frac{\partial}{\partial x}\left(-\frac{g'(x)}{g(x)}\right) - \left(\frac{g'(x)}{g(x)}\right)^2 \ge 0 - k^2 = -k^2,$$

and hence also

$$2\,\psi'(x) - (\psi(x))^2 \ge -k^2 \quad \text{for } x \in (x_0, x_1),$$
$$2\,\psi'(x) - (\psi(x))^2 = -k^2 \quad \text{for } x \notin (x_0, x_1).$$

We can now evaluate for any distribution F_1

$$\begin{aligned}
&\int_{\mathbb{R}} \left(2\,\psi'(x) - (\psi(x))^2\right) \mathrm{d}\,(F_1 - F_0)(x) \\
&\quad = \int_{\mathbb{R}} \left(2\,\psi'(x) - (\psi(x))^2 + k^2\right) \mathrm{d}\,(F_1 - F_0)(x) - \int_{\mathbb{R}} k^2 \,\mathrm{d}\,(F_1 - F_0)(x) \\
&\quad = \int_{x_0}^{x_1} \left(2\,\psi'(x) - (\psi(x))^2 + k^2\right) \mathrm{d}\,(F_1 - F_0)(x) - \int_{\mathbb{R}} k^2 \,\mathrm{d}\,(F_1 - F_0)(x).
\end{aligned}$$

Now, surely $\int_{\mathbb{R}} k^2 \,\mathrm{d}\,(F_1 - F_0)(x) = 0$. In addition, for the density f_1 of F_1 we must have $f_1(x) = (1-\varepsilon)\,g(x) + \varepsilon\,h(x) \geq (1-\varepsilon)\,g(x) = f_0(x)$ for h the density of H for all $x \in (x_0, x_1)$. The last inequality $(1-\varepsilon)\,g(x) = f_0(x)$ holds because of (128). Thus, also the first term on the right-hand side above must be non-negative, as we wanted to prove. □

The result of Theorem 18 is surprising. The least informative density in the neighbourhood of G has tails that are quite light; they decrease only exponentially as $x \to \infty$. A very important special case is obtained for $G = \Phi$ the standard normal distribution.

Example 4.1. For $G = \Phi \in \mathcal{P}(\mathbb{R})$ and $g = \varphi = \Phi'$ we have in Theorem 18 that $g'(x)/g(x) = x$ for $x \in \mathbb{R}$ and $-x_0 = x_1 = k$. The constant $k \geq 0$ is determined by the normalising condition

$$\frac{1}{1-\varepsilon} = 2\,\Phi(k) - 1 + \frac{2\,\varphi(k)}{k}.$$

We cannot express k explicitly, but it is determined uniquely. To see that, it remains to take the derivative of the right-hand side of this formula and conclude that the function is strictly monotone in $k \in \mathbb{R}$. The least informative density in the ε-contamination neighbourhood of Φ is thus

$$f_0(x) = \begin{cases} (1-\varepsilon)\varphi(x) & \text{for } |x| \leq k, \\ \frac{1-\varepsilon}{\sqrt{2\,\pi}} \exp\left(-k^2/2 - k\,|x|\right) & \text{for } |x| > k. \end{cases}$$

The corresponding score function $\psi(x) = -f_0'(x)/f_0(x)$ takes a quite simple form

$$\psi(x) = \min\{k, \max\{-k, x\}\} = \begin{cases} -k & \text{for } x < -k, \\ x & \text{for } x \in [-k, k], \\ k & \text{for } x > k. \end{cases} \tag{129}$$

The density f_0 and its score function ψ are visualised in Figure 13.

△

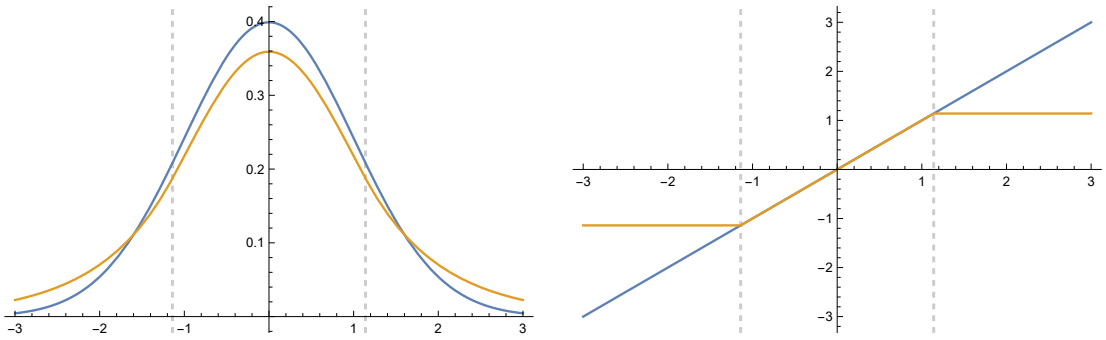

Figure 13: Example 4.1. Left panel: The density of the standard normal distribution φ (blue) and the least informative density f_0 for $\varepsilon = 1/10$ (orange). In the interval $x \in [-k, k]$ attains the density f_0 the lowest value $(1-\varepsilon)\varphi(x)$ that a density from $\mathcal{P}(\Phi)$ can get. Right panel: The score function $-\varphi'(x)/\varphi(x) = x$ for the standard normal distribution (blue) and the score function ψ for the distribution given by f_0 (orange). The vertical dashed lines in both plots are the cut-off values $-k$ and k. For $\varepsilon = 1/10$ we have $k \approx 1.140$.

4.2.2 Step 2: Optimality of the M-estimator

Let f_0 be the density of the distribution F_0 that minimises the Fisher information in $\mathcal{P}(G)$, and let

$$\psi(x) = -\frac{f_0'(x)}{f_0(x)} \quad \text{for } x \in \mathbb{R}$$

be the score function that defines the asymptotically efficient M-estimator T from (125) for F_0. For the asymptotic variance $A(F, T)$ of T at F we have by Theorem 5 and 8

$$A(F,T) = \frac{\int_{\mathbb{R}} \psi(x - T(F))^2 \,\mathrm{d}\, F(x)}{\left(\int_{\mathbb{R}} \psi'(x - T(F)) \,\mathrm{d}\, F(x)\right)^2}. \tag{130}$$

At this point, we encounter a fundamental difficulty with the expression $T(F)$ in the right-hand side of (130). This term makes optimising the supremum of $A(F, T)$ cumbersome. To get away with this term, we must restrict our problem further. We will not consider all distributions in the neighbourhood $\mathcal{P}(G)$ from (126), but only distributions in its subset

$$\mathcal{P}_0(G) = \{F \in \mathcal{P}(G) \colon T(F) = 0\}.$$

This reduced neighbourhood is still very rich; it contains all symmetric distributions $F \in \mathcal{P}(G)$ from (126).

We know that the estimator T is asymptotically efficient for F_0. Thus, we have by Theorem 15 that $A(F_0, T) = (J(F_0))^{-1}$. We will show the following result.

Theorem 19. *Let $G \in \mathcal{P}(\mathbb{R})$ be a symmetric distribution and let $F_0 \in \mathcal{P}_0(G)$ be the distribution that minimises the Fisher information in $\mathcal{P}_0(G)$. Denote by T be asymptotically efficient M-estimator of location (125) corresponding to F_0 given by*

$$\psi(x) = -\frac{f_0'(x)}{f_0(x)} \quad \text{for } x \in \mathbb{R},$$

where f_0 is the density of F_0. Then

$$A(F, T) \leq A(F_0, T) = \frac{1}{J(F_0)} \quad \text{for all } F \in \mathcal{P}_0(G). \tag{131}$$

In particular, the M-estimator T is a minimax variance optimal estimator of location in the neighbourhood $\mathcal{P}_0(G)$.

Proof. We only need to prove the inequality in (131); the optimality of T then follows from the fact that by Theorem 15, no M-estimator S of location can have $\sup_{F \in \mathcal{P}_0(G)} A(F, S)$ lower than $A(F_0, T) = 1/J(F_0)$.

To prove (131), we first observe that just like the Fisher information, also the inverse of the asymptotic variance functional (130) given by

$$\frac{1}{A(F, T)} = \frac{\left(\int_{\mathbb{R}} \psi'(x) \,\mathrm{d}\, F(x)\right)^2}{\int_{\mathbb{R}} \psi(x)^2 \,\mathrm{d}\, F(x)} \tag{132}$$

is convex in $F \in \mathcal{P}_0(G)$. This follows from Lemma 5, because

$$\int_{\mathbb{R}} h(x) \,\mathrm{d}\, F_t(x) = \int_{\mathbb{R}} h(x) \,\mathrm{d}\, F_0(x) + t \int_{\mathbb{R}} h(x) \,\mathrm{d}\, (F_1 - F_0)(x)$$

is certainly linear in $t \in [0, 1]$ for any $h \colon \mathbb{R} \to \mathbb{R}$.

We now compute the directional derivative of the functional $1/A(\cdot, T)$ at F_0 in direction $F_1 \in \mathcal{P}_0(G)$; we want to find that $1/A(F, T)$ is minimised in $F = F_0$. For $F_t = (1-t)\, F_0 + t\, F_1$ and any $h \colon \mathbb{R} \to \mathbb{R}$ we have

$$\frac{\partial}{\partial t} \left[\int_{\mathbb{R}} h(x) \,\mathrm{d}\, F_t(x)\right]_{t=0} = \int_{\mathbb{R}} h(x) \,\mathrm{d}\, (F_1 - F_0)(x).$$

We use this to obtain

$$\begin{aligned}
\frac{\partial}{\partial t}\left[\frac{1}{A(F_t,T)}\right]_{t=0} &= \frac{\partial}{\partial t}\left[\frac{\left(\int_{\mathbb{R}}\psi'(x)\,\mathrm{d}\,F_t(x)\right)^2}{\int_{\mathbb{R}}\psi(x)^2\,\mathrm{d}\,F_t(x)}\right]_{t=0} \\
&= \frac{2\left(\int_{\mathbb{R}}\psi'(x)\,\mathrm{d}\,F_0(x)\right)\left(\int_{\mathbb{R}}\psi'(x)\,\mathrm{d}\,(F_1-F_0)(x)\right)\left(\int_{\mathbb{R}}\psi(x)^2\,\mathrm{d}\,F_0(x)\right)}{\left(\int_{\mathbb{R}}\psi(x)^2\,\mathrm{d}\,F_0(x)\right)^2} \\
&\quad - \frac{\left(\int_{\mathbb{R}}\psi'(x)\,\mathrm{d}\,F_0(x)\right)^2\left(\int_{\mathbb{R}}\psi(x)^2\,\mathrm{d}\,(F_1-F_0)(x)\right)}{\left(\int_{\mathbb{R}}\psi(x)^2\,\mathrm{d}\,F_0(x)\right)^2} \\
&= 2\left(\int_{\mathbb{R}}\psi'(x)\,\mathrm{d}\,(F_1-F_0)(x)\right)\frac{\int_{\mathbb{R}}\psi'(x)\,\mathrm{d}\,F_0(x)}{\int_{\mathbb{R}}\psi(x)^2\,\mathrm{d}\,F_0(x)} \\
&\quad - \left(\int_{\mathbb{R}}\psi(x)^2\,\mathrm{d}\,(F_1-F_0)(x)\right)\left(\frac{\int_{\mathbb{R}}\psi'(x)\,\mathrm{d}\,F_0(x)}{\int_{\mathbb{R}}\psi(x)^2\,\mathrm{d}\,F_0(x)}\right)^2 \\
&= \int_{\mathbb{R}}\left(2\,\psi'(x)-\psi(x)^2\right)\,\mathrm{d}\,(F_1-F_0)(x),
\end{aligned}$$

where the last equality follows from

$$\begin{aligned}
\int_{\mathbb{R}}\psi'(x)\,\mathrm{d}\,F_0(x) &= -\int_{\mathbb{R}}\frac{\partial^2}{\partial x^2}\log(f_0(x))\,\mathrm{d}\,F_0(x) = J(F_0), \\
\int_{\mathbb{R}}\psi(x)^2\,\mathrm{d}\,F_0(x) &= \int_{\mathbb{R}}\left(-\frac{\partial}{\partial x}\log(f_0(x))\right)^2\,\mathrm{d}\,F_0(x) = J(F_0),
\end{aligned}$$

see, e.g., [19, Theorem 1]. Compare our result with (127) that we used to prove Theorem 17. We see that

$$\frac{\partial}{\partial t}\left[\frac{1}{A(F_t,T)}\right]_{t=0} = \frac{\partial}{\partial t}\left[J(F_t)\right]_{t=0},$$

where we saw that the right-hand side is for F_0 minimising the Fisher information always non-negative. We get that $1/A(F_t,T)$ is a convex function in t whose derivative in t at $t=0$ is non-negative. Necessarily, $t=0$ (or equivalently F_0) must be the measure that minimises $1/A(F_t,T)$. Since $F_1 \in \mathcal{P}_0(G)$ was chosen arbitrarily, that gives

$$A(F_0,T) \geq A(F,T) \quad \text{for all } F \in \mathcal{P}_0(G),$$

as we wanted to prove. □

Example 4.2. We have found that the M-estimator from Example 4.1 is minimax optimal in the contamination neighbourhood of the normal distribution. This important estimator is sometimes called the *Huber estimator* of location. It does not take an explicit form; for a random sample $X_1,\ldots,X_n$ it is computed as the solution in $t \in \mathbb{R}$ to

$$\sum_{i=1}^{n}\psi(X_i-t)=0$$

with ψ from (129), or equivalently as $t \in \mathbb{R}$ that minimises the *Huber loss*

$$\sum_{i=1}^{n} \rho(X_i - t),$$

where

$$\rho(x) = \begin{cases} \frac{x^2}{2} & \text{for } |x| \leq k, \\ k|x| - \frac{k^2}{2} & \text{for } x > k. \end{cases} \tag{133}$$

Observe that as $k \to \infty$ we get the squared loss function, and the estimator is just the sample average. As $k \to 0$, which corresponds to the amount of contamination $\varepsilon \to 1$, we approach the absolute loss and get the sample median. The Huber estimator can therefore be considered to be a compromise between the median and the mean, optimised in the sense of being robust and at the same time not losing much efficiency at the normal distribution. △

4.3 Minimax optimality: Additional remarks

Returning to the optimal Huber estimator of location T from Example 4.1, we see that in view of Theorem 10, this estimator is qualitatively robust and possesses asymptotic breakdown point $\varepsilon^*(P, T) = 1/2$ for any $P \in \mathcal{P}(\mathbb{R})$. Looking at its influence function in Theorem 8, we see that the effect of extremely large observations is bounded by the constant $k > 0$. This means that the estimator down-weights the effect of the outliers, but they still do contribute to the resulting estimator.

A way to approach this phenomenon might be to consider M-estimators whose influence function is restricted to be zero outside the interval $[-k, k]$. Such estimators are called *re-descending M-estimators*. They correspond to searching for score functions ψ that minimise maximum asymptotic variance, under the additional condition $\psi(x) = 0$ for $x \notin [-k, k]$. It is not difficult to apply this restriction to the theory of minimax estimation devised in the previous section. For the ε-contamination neighbourhood of the normal distribution, one gets the odd function

$$\psi(x) = -\psi(-x) = \begin{cases} x & \text{for } x \in [0, a], \\ b \tanh(b(c - x)/2) & \text{for } x \in (a, c], \\ 0 & \text{for } x > c, \end{cases}$$

for appropriate constants $a, b, c > 0$ that depend on ε. This function is displayed in the top left panel of Figure 14.

Since the shape of the optimal redescending function ψ is somewhat cumbersome, several authors have suggested simpler alternatives. Popular choices are the Hampel piecewise linear

function

$$\psi(x) = -\psi(-x) = \begin{cases} x & \text{for } x \in [0, a), \\ a & \text{for } x \in [a, b), \\ \frac{c-x}{c-b} a & \text{for } x \in [b, c), \\ 0 & \text{for } x \geq c, \end{cases}$$

the Andrews sine wave function

$$\psi(x) = \begin{cases} \sin(x) & \text{for } x \in [-\pi, \pi], \\ 0 & \text{otherwise,} \end{cases}$$

or the Tukey biweight function

$$\psi(x) = \begin{cases} x^2(1-x^2)^2 & \text{for } x \in [-1, 1], \\ 0 & \text{otherwise.} \end{cases}$$

These functions are all alike the optimal redescending score function; they are not optimal at the normal distribution, but produce similar, very robust estimators. All these four functions are drawn in Figure 14.

In this section we treated only minimax optimality for M-estimators. The situation with minimax optimal L, or R-estimators is more complicated. The main problem stems from the fact that the inverse asymptotic variance $1/A(F, T)$ from (132) is no longer convex in F, and thus the claim of Theorem 19 does not have to be true. In particular, the asymptotically efficient estimators at the least informative distributions do not have to be minimax optimal.

In particular cases, one can still derive various optimality results. As argued in [12, Section 4.7], for the ε-contaminated normal distribution, the minimax optimal L-estimator can be shown to be the trimmed mean from Example 3.13.

5 Further topics in robustness

In our treatment, we primarily considered only the one-dimensional situation, and developed the theory of minimax optimal robustness only in the case of a location parameter (121). Of course, much more can be done, and the general principles of robustness also apply in many different settings. In the present section, we outline the very basics of some of these ideas and give references.

5.1 Equivariance of robust location estimators

Recall that we say that a (location) functional $T\colon \mathcal{P}(\mathbb{R}) \to \mathbb{R}$ is translation and scale equivariant if for all $X \sim P \in \mathcal{P}(\mathbb{R})$ and $Y = aX + b \sim Q \in \mathcal{P}(\mathbb{R})$ with $a > 0$ and $b \in \mathbb{R}$, we

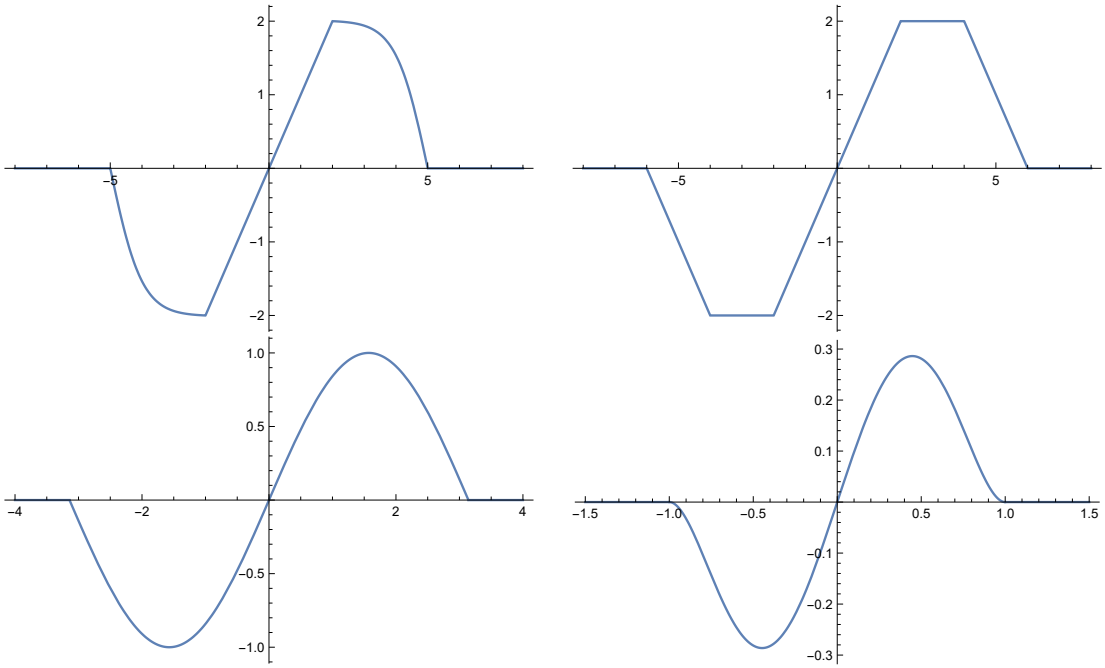

Figure 14: The four redescending score functions that generate robust location estimators: the minimax optimal one (top left), the Hampel piecewise linear function (top right), the Andrews sine wave (bottom left), and the Tukey biweight function (bottom right).

have

$$T(Q) = a\,T(P) + b. \tag{134}$$

We have shown in (89) that this property is satisfied for common L-functionals of location. Using (88) in the defining formula of R-functionals (108), it is easy to see that the same is true also for R-functionals. In contrast, for M-functionals of location, we have shown in (63) only translation equivariance, that is the special case of (134) with $a = 1$. In general, there is no reason for M-functionals to be scale equivariant, since substituting $a\,X_i + b$ instead of X_i and $a\,T_n + b$ instead of T_n in the formula

$$\sum_{i=1}^{n} \psi(X_i - T_n) = 0 \tag{135}$$

for a location Z-estimator, see (47), we get only

$$\sum_{i=1}^{n} \psi(a\,X_i + b - (a\,T_n + b)) = \sum_{i=1}^{n} \psi(a(X_i - T_n)).$$

There is no reason why the factor $a > 0$ should disappear in the last formula. This lack of equivariance might be troubling because the resulting M-estimator of location is then dependent on the unknown scale (dispersion) of the data. To solve this problem, it is possible to input the scale correction into the estimating formula (135) manually. Take $S\colon \mathcal{P}(\mathbb{R}) \to [0, \infty]$ a scale functional, meaning that S verifies

$$S(Q) = a\,S(P) \quad \text{for all } a > 0 \text{ and } b \in \mathbb{R}, \tag{136}$$

where P and Q are as in (134). Plug S into (135) to automatically account for scale differences, and define $T_n = T(P_n)$ as a solution to

$$\sum_{i=1}^{n} \psi\left(\frac{X_i - T_n}{S(P_n)}\right) = 0. \tag{137}$$

Now, substituting X_i by $a\,X_i + b$ and T_n by $a\,T_n + b$ in the previous formula, we see that not only b, but also a cancels out. We thus obtain an M-estimator T_n that is both location and scale equivariant.

It can be shown that if ψ is odd and bounded and if S verifies (136), the asymptotic breakdown point of T_n is equal to the asymptotic breakdown value of $S(P_n)$. A good choice for the scale estimator S is the *median absolute deviation* (also called simply MAD)

$$S(P_n) = \mathsf{med}\ \left(|X_1 - \mathsf{med}\,(P_n)|, \ldots, |X_n - \mathsf{med}\,(P_n)|\right). \tag{138}$$

The asymptotic breakdown point of MAD is $1/2$, and it is easy to see that this estimator is a scale functional verifying (136). When used in conjunction with the Huber estimator of

location from Example 4.2, it is even better to scale S to be a Fisher consistent estimator of the standard deviation σ in the Gaussian model $P = \mathsf{N}(\mu, \sigma^2)$. That is achieved by considering

$$S(P_n) = \frac{\mathsf{med}\ (|X_1 - \mathsf{med}\,(P_n)|, \ldots, |X_n - \mathsf{med}\,(P_n)|)}{\Phi^{-1}(3/4)}$$
$$\approx 1.4826\,\mathsf{med}\ (|X_1 - \mathsf{med}\,(P_n)|, \ldots, |X_n - \mathsf{med}\,(P_n)|)\,.$$

This improved version of the Huber location estimator is the one typically used in standard statistical software, e.g. in function huber in R package MASS, or in huberM in R package robustbase.

5.2 Computation of M-estimators of location

Solving equations defining M, or R-estimators does not have to be straightforward. The case of location M-estimators and equations (135) or (137) is, however, easy to solve iteratively. We consider (137) and the scale equivariant M-estimator of location. Defining

$$W(x) = \begin{cases} \psi(x)/x & \text{for } x \neq 0, \\ \psi'(0) & \text{for } x = 0, \end{cases}$$

we can write T_n from (137) as a solution to

$$\sum_{i=1}^{n} \left(\frac{X_i - T_n}{S(P_n)} \right) \cdot W \left(\frac{X_i - T_n}{S(P_n)} \right) = 0.$$

Considering that $S(P_n)$ is fixed and positive and interpreting

$$w_i = W \left(\frac{X_i - T_n}{S(P_n)} \right)$$

as weights, this gives that T_n can be written as a weighted mean

$$T_n = \frac{\sum_{i=1}^{n} w_i\, X_i}{\sum_{j=1}^{n} w_j}.$$

This is, of course, not precise because the weights w_i still depend on the unknown T_n. But, it suggests the following simple iterative procedure:

1. Compute $s = S(P_n)$, set $k = 0$ and $t_0 = \mathsf{med}\,(P_n)$,
2. Increase k by one and compute the weights
$$w_{k,i} = W \left(\frac{X_i - t_{k-1}}{s} \right).$$
3. Set
$$t_k = \frac{\sum_{i=1}^{n} w_{k,i}\, X_i}{\sum_{j=1}^{n} w_{k,j}}. \tag{139}$$

4. If $|t_k - t_{k-1}| < 10^{-6}\, s$, then stop and return $T_n = t_k$; otherwise return to step (2).

The constant 10^{-6} in step (4) was taken as an arbitrary small number; we use it to signal whether the weighted mean t_k still changes. Because the weighted mean in (139) is a special case of weighted least squares, the procedure above is called *iteratively reweighted least squares* (IRLS). If the function W is bounded, symmetric, and non-increasing for $x > 0$, the sequence $\{t_k\}_{k=0}^{\infty}$ is bound to converge to a solution to (137).

5.3 Estimation of location and scale

Consider the two-parameter location-scale model given by the system of distributions

$$\mathcal{F} = \left\{ F_0\left(\frac{\cdot - \theta}{\sigma}\right) : \theta \in \mathbb{R} \text{ and } \sigma > 0 \right\},$$

where F_0 is a distribution function with density f_0. One starts with the maximum likelihood estimators, which maximise the log-likelihood function

$$\sum_{i=1}^{n} \log\left(\frac{1}{\sigma} f_0\left(\frac{x_i - \theta}{\sigma}\right)\right) = -n\log(\sigma) + \sum_{i=1}^{n} \log\left(f_0\left(\frac{x_i - \theta}{\sigma}\right)\right)$$

in θ and σ. This is solved by considering the likelihood equations

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \psi\left(\frac{x_i - \theta}{\sigma}\right), \\ 0 &= \sum_{i=1}^{n} \left(\psi\left(\frac{x_i - \theta}{\sigma}\right)\frac{x_i - \theta}{\sigma} - 1\right), \end{aligned} \tag{140}$$

where we denoted by $\psi(x) = -f_0'(x)/f_0(x)$ the negative score function of F_0. Generalising this system of equations, one can define M-estimators of location and scale as a pair of estimators (T_n, S_n) that satisfy the system

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \psi\left(\frac{x_i - T_n}{S_n}\right), \\ 0 &= \sum_{i=1}^{n} \chi\left(\frac{x_i - T_n}{S_n}\right). \end{aligned}$$

Here, ψ and χ are appropriate functions. In most cases, in analogy with (140) for f_0 symmetric around the origin, ψ is taken to be an odd function and χ is even. For $F_0 = \Phi$ the standard normal distribution, we get $\psi(x) = x$ and $\chi(x) = x^2 - 1$. The M-functionals corresponding to T_n and S_n are naturally T and S given by

$$\begin{aligned} 0 &= \int_{\mathbb{R}} \psi\left(\frac{x - T}{S}\right) \mathrm{d}\, P(x), \\ 0 &= \int_{\mathbb{R}} \chi\left(\frac{x - T}{S}\right) \mathrm{d}\, P(x). \end{aligned} \tag{141}$$

A theory of M-estimators that is more complicated, but similar to what we proved in Sections 3.1 and 4, can also be developed for the pair (T, S), and for more general multidimensional M-estimators [12, Chapter 6].

An estimator of scale S analogous to the median $T(F) = \mathsf{med}\,(F)$ is the defined by solving (141) for

$$\psi(x) = \operatorname{sign}(x),$$
$$\chi(x) = \operatorname{sign}(|x| - 1).$$

One obtains a pair of estimators (T, S) with $T(F_n) = \mathsf{med}\,(X_1, \ldots, X_n)$ and $S(F_n)$ the MAD from (138).

It is not hard to see that the asymptotic breakdown point of S is $1/2$, the maximum possible (sensible) value.

The scale equivalent to the Huber estimator from Examples 4.1 and 4.2 is the *Huber M-estimator of scale* S obtained by taking in (141) the functions

$$\psi(x) = \min\{k, \max\{-k, x\}\},$$
$$\chi(x) = \psi(x)^2 - \beta(k).$$

Here ψ is the same as in (129), and $\beta(k)$ is chosen appropriately so that for F the standard normal distribution, the functional S is Fisher consistent.

5.4 Robustness in multidimensional spaces

Principles of robustness nicely expand also to multidimensional data from $\mathbb{R}^d$. The location parameter of $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ is now a vector $\boldsymbol{\theta} \in \mathbb{R}^d$, and the role of a scale parameter is played by a positive definite matrix $\Sigma \in \mathbb{R}^{d\times d}$. In multi-dimensional spaces, there is no natural ordering of the observations, so L-estimators and R-estimators cannot be considered directly. As for the M-estimators, one can begin from the maximum likelihood approach for systems of elliptically symmetric densities. Such densities $f\colon \mathbb{R}^d \to [0, \infty)$ are characterised by assuming the general form

$$f(\boldsymbol{x}, \boldsymbol{\theta}, \Sigma) = \left|\det\left(\Sigma^{-1/2}\right)\right| g\left((\boldsymbol{x} - \boldsymbol{\theta})^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\theta})\right) \quad \text{for } \boldsymbol{x} \in \mathbb{R}^d,$$

where $g\colon [0, \infty) \to [0, \infty)$ is a univariate function, and det is the determinant of a matrix. We can again approach the problem from the angle of maximum likelihood estimation. One can express the likelihood equations corresponding to f as in (140), and generalise M-estimators to the multivariate situation. In taking the derivatives with respect to the vector (or matrix) parameters $\boldsymbol{\theta}$ and Σ, one however has to use some matrix differential calculus. This theory is expounded in, e.g., [12, Chapter 8].

5.5 Robustness in regression

A major topic which we did not cover is the problem of robust estimation of regression parameters. Here, we are given independent observations $(\boldsymbol{X}_i, Y_i)^\mathsf{T} \in \mathbb{R}^d \times \mathbb{R}$ for $i = 1, \dots, n$ that take the form

$$Y_i = \boldsymbol{X}_i^\mathsf{T} \boldsymbol{\beta} + \varepsilon_i,$$

where ε_i is a sequence of independent (unobserved) errors with distribution symmetric around zero, and $\boldsymbol{\beta} \in \mathbb{R}^d$ is the unknown parameter of interest. The standard least squares estimator [14, Section 2] is the minimiser of the sum of squares

$$\sum_{i=1}^n \left(Y_i - \boldsymbol{X}_i^\mathsf{T} \boldsymbol{b}\right)^2 \tag{142}$$

in $\boldsymbol{b} \in \mathbb{R}^d$. It corresponds to the maximum likelihood estimator of $\boldsymbol{\beta}$ if the conditional distribution of Y_i given $\boldsymbol{X}_i$ is normal. Just as the sample mean (the maximum likelihood estimator in a normal location model, see Example 1.2), the least squares model is also very non-robust. It is easy to see that the least squares estimator collapses completely by taking a single observation Y_i far away from the rest of the data. In terms of robustness, its asymptotic breakdown point is thus zero.

There are many approaches towards robustifying the least squares estimator. A straightforward idea is to replace the (non-robust) square function in (142) with a function that does not grow so fast at infinity. That would be analogous to our approach from M-estimation of location. We thus want to minimise the function

$$\boldsymbol{b} \mapsto \sum_{i=1}^n \rho\left(Y_i - \boldsymbol{X}_i^\mathsf{T} \boldsymbol{b}\right), \tag{143}$$

for $\rho \colon \mathbb{R} \to \mathbb{R}$ given, and take its argument of minima as the estimator of $\boldsymbol{\beta}$. This is precisely in line with our treatment of M-estimators from Section 3.1. We can thus call such an estimator a regression M-estimator. Its functional counterpart is the minimiser of the integral

$$\boldsymbol{b} \mapsto \int_{\mathbb{R}^{d+1}} \rho(y - \boldsymbol{x}^\mathsf{T} \boldsymbol{b}) \, \mathrm{d}\, P_{\boldsymbol{X},Y}(\boldsymbol{x}, y)$$

where $P_{\boldsymbol{X},Y}$ is the joint distribution of the random vector $(\boldsymbol{X}_1, Y_1)^\mathsf{T} \in \mathbb{R}^{d+1}$. In the special case of $\rho(x) = |x|$ being the absolute value function, we obtain the famous least absolute deviations estimator treated in [21, Section 3]. Another sensible choice could be to employ the Huber loss function (133) in (143).

One particular property of regression M-estimators from (143) is that they are typically robust only with respect to changes in the response Y_i. If, however, the regressors $\boldsymbol{X}_i$ are taken to the extreme, the estimators still break down easily. This is known as the problem of

leverage points [14, Chapter 11]. It can be resolved by modifying (143), introducing another weight function w into the formula

$$\boldsymbol{b} \mapsto \sum_{i=1}^{n} w(\boldsymbol{X}_i) \cdot \rho\left(Y_i - \boldsymbol{X}_i^{\mathsf{T}} \boldsymbol{b}\right)$$

with w that does not grow fast as its argument $\boldsymbol{x}$ drifts away.

A detailed treatment of all these topics can be found in [12, Chapter 7] or [24]. There are also notions of L and R-like estimators for regression problems. An L-estimator is, for example, given by the concept of regression quantiles [21, Section 5].

5.6 Final comments

The problem of robustness of estimation has now already been thoroughly studied in many setups. For robust estimation in mixed and generalised linear models, longitudinal data analysis, and survival data, one can see [11]. Robust methods in time series analysis are treated in [17]. Glimpses of robust testing procedures and robust Bayesian analysis can be found in [12, Chapters 13 and 15] and [10].

References

[1] Brenton R. Clarke. Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations. *Ann. Statist.*, 11(4):1196–1205, 1983.

[2] Brenton R. Clarke. Nonsmooth analysis and Fréchet differentiability of M-functionals. *Probab. Theory Relat. Fields*, 73(2):197–209, 1986.

[3] Brenton R. Clarke. *Robustness theory and application*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, 2018.

[4] M. Cohen. The Fisher information and convexity (corresp.). *IEEE Transactions on Information Theory*, 14(4):591–592, 1968.

[5] P. L. Davies and U. Gather. The breakdown point—examples and counterexamples. *REVSTAT*, 5(1):1–17, 2007.

[6] P. Laurie Davies and Ursula Gather. Breakdown and groups. *Ann. Statist.*, 33(3):977–1035, 2005.

[7] R. M. Dudley. *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.

[8] Luisa Turrin Fernholz. *von Mises calculus for statistical functionals*, volume 19 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1983.

[9] Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *Int. Stat. Rev.*, 70(3):419–435, 2002.

[10] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust statistics: The approach based on influence functions*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986.

[11] Stephane Heritier, Eva Cantoni, Samuel Copt, and Maria-Pia Victoria-Feser. *Robust methods in biostatistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2009.

[12] Peter J. Huber and Elvezio M. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2009.

[13] Jana Jurečková. *Robustní statistické metody*. Karolinum, Praha, 2001.

[14] A. Komárek. NMSA407 Linear Regression: Course notes. https://www2.karlin.mff.cuni.cz/~kulich/vyuka/linreg/doc/2021-NMSA407-notes.pdf, 2021. Accessed: 2023-04-12.

[15] Michal Kulich and Marek Omelka. NMSA331 Matematická statistika 1. Poznámky k přednášce. Univerzita Karlova, 2022. https://www2.karlin.mff.cuni.cz/~omelka/Soubory/nmsa331/ms1.pdf. Accessed: 2022-01-30.

[16] Petr Lachout. *Teorie pravděpodobnosti*. Karolinum, 1998.

[17] Ricardo A. Maronna, R. Douglas Martin, Victor J. Yohai, and Matías Salibián-Barrera. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, 2019.

[18] Ivan Mizera. Qualitative robustness and weak continuity: the extreme unction? In *Nonparametrics and robustness in modern statistical inference and time series analysis: a Festschrift in honor of Professor Jana Jurečková*, volume 7 of *Inst. Math. Stat. Collect.*, pages 169–181. Inst. Math. Statist., Beachwood, OH, 2010.

[19] Stanislav Nagy. NMSA332: Mathematical Statistics 2. https://www2.karlin.mff.cuni.cz/~nagy/NMSA332/NMSA332.pdf, 2023.

[20] M. Omelka. NMSA434: Modern Statistical Methods. https://www2.karlin.mff.cuni.cz/~omelka/Soubory/nmst434/nmst434_course-notes.pdf, 2022. Accessed: 2023-02-14.

[21] M. Omelka. NMST424: Mathematical Statistics 3. https://www2.karlin.mff.cuni.cz/~omelka/Soubory/nmst424/nmst424_course-notes.pdf, 2023. Accessed: 2023-04-12.

[22] L. Pick, S. Hencl, J. Spurný, and M. Zelený. Matematická analýza 1. https://www2.karlin.mff.cuni.cz/~pick/analyza.pdf, 2022. Accessed: 2023-02-12.

[23] J. Rataj. Teorie míry a integrálu. https://www2.karlin.mff.cuni.cz/~rataj/TMI/TMI-text_2017.pdf, 2018. Accessed: 2023-02-19.

[24] Peter J. Rousseeuw and Annick M. Leroy. *Robust regression and outlier detection*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1987.

[25] Robert Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons Inc., New York, 1980. Wiley Series in Probability and Mathematical Statistics.

[26] J. Spurný. Funkcionální analýza. https://www2.karlin.mff.cuni.cz/~spurny/doc/faprednaska.pdf, 2017. Accessed: 2023-02-12.

[27] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[28] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.