Robust tests and location estimators over uncertainty classes

Bachelor Thesis

by
Genc Kqiku

supervised by
Prof. Sara van de Geer

Department of Mathematics
Seminar for Statistics
ETH Zürich

Spring semester 2020

Abstract. This thesis discusses the existence and the construction of robust tests as minimax tests between composite hypotheses formed of specified uncertainty classes of distributions. It also presents the construction of location estimators that minimize the maximum asymptotic variance or bias when the underlying distribution lies in some uncertainty set. Robust confidence intervals are also discussed in this setting. Several different types of uncertainty classes are considered.

Contents

1 Introduction

When taking a statistical decision, a few bad observations in the sample can ruin the output, however good our model is. Consider for example a likelihood ratio test $\Phi(X_1, ..., X_n) = \mathbb{1}\{\prod p_1(X_i)/p_0(X_i) \geq c\}$. A single deviation (due to bad measurement, model misspecification, adversarial attack, etc.) can send the likelihood ratio p_1/p_0 to $+\infty$ or to 0 and hence completely change the decision. Deviations also make the probability of errors rise: the significance level and power are not correct anymore. The same issue shows up with estimators. Take as an example data following a $N(\mu, \sigma)$ distribution. Then $\hat{\sigma} = \left[\frac{1}{n}\sum(x_i - \bar{x})^2\right]^{1/2}$ and MAD $= \frac{1}{n}\sum|x_i - \bar{x}|$ are asymptotically normal estimators of σ and $\sqrt{2/\pi}\,\sigma$ respectively. Tukey 1960 showed that only 2% of observations following $N(\mu, 3\sigma)$ instead of $N(\mu, \sigma)$ suffice to reverse the advantage of $\hat{\sigma}$ against MAD with respect to the asymptotic variance.

Peter J. Huber wrote that "a small minority of the observations should never be able to override the evidence of the majority" (Huber 1965). This is the main goal of robust statistics. Huber has been part of the foundation of this area, and this thesis is based mostly on his work. We focus on the case of outliers coming from model misspecification, i.e. when a minority of the observations do not come from the believed distribution, called the *nominal distribution*. The idea is to let ourselves some space to make mistakes, by assuming that the underlying distribution of the data lies in a certain set of distributions. The latter is a neighborhood of the nominal distribution and we call these type of sets *uncertainty classes*. The aim of the thesis is to show how to build decisions that minimize the maximum degradation of performance possible over an appropriately specified uncertainty class. The decisions we look for are in that sense *minimax*. Although robustness is also about how much efficiency one is ready to sacrifice in order to gain stability, or how many bad observations suffice to cause a catastrophe in the decision (breakdown point), we only treat stability issues.

This thesis covers both finite sample size and asymptotic minimax decisions. The separation is very important. Indeed, assume our data contains 1% of bad observations. Then, if the sample size is 1000, we expect 10 bad observations per sample. But if the sample size is 5, then 19 out of 20 samples will not be contaminated at all. Thus, it can seem inappropriate to use asymptotically optimal robust decisions on a small sample size. The gain we obtain on an occasional erroneous value may be negligible in front of the loss of efficiency on the dominating good observations. For this reason we also need results which do not depend on the sample size. We will see that sometimes both methods lead to the same decision, which is a very pleasant property.

We cover three types of decisions. In the first two chapters we build robust hypothesis tests. Instead of testing between two distributions P_0, P_1, we test

between two neighborhoods of P_0 and P_1. In Chapter 2 we cover (according to Huber and Strassen 1973) the proof of existence of a robust version of the Neyman-Pearson test, for which we can control the significance level and maximize the power for testing distributions lying in very general uncertainty classes. These classes cover some well known special cases like Kolmogorov neighborhoods, or the contamination case, on which we focus in Section 2.3. These very general tests being hard to compute, we treat in Chapter 3 a more simple case where the uncertainty classes consist of the distributions for which the associated densities are close to the nominal densities, according to the Kullback-Leibler divergence. For this case we are able to analyze the performance of the robust test (see Section 3.2). Then we direct our attention on robust estimators, so we assume that the distribution of the data lies in some set. Chapter 4 focuses on estimators with minimax properties for fixed and finite sample size. Those estimators are derived from the tests developed in Chapter 2 and are helpful for the construction of minimax confidence intervals. Next, we build estimators with minimax asymptotic properties. First is the asymptotic bias the value to minimize over the uncertainty class, and then the asymptotic variance. For the latter we look at a general way to find minimax estimators, and finally we see how functions of the order statistics can minimize the worst-case asymptotic variance.

Acknowledgments. I am very grateful to Prof. van de Geer for her involvement in this thesis. I would like to thank her for her time and precious comments as well as for inspiring me through her excellent lectures. They provide valuable knowledge, and the frequent side comments give a good picture of the possible applications, difficulties or extensions of the studied tools. It is with one of those remarks that I discovered Huber's work in robust statistics and wanted to study it in more depth.

2 Minimax tests using 2-alternating capacities

We first direct our interest on building robust tests. Let $(\Omega, \mathcal{F})$ be a measurable space and $\mathcal{M}$ be the set of all probability measures on $(\Omega, \mathcal{F})$. Let X a random variable following an unknown distribution $P \in \mathcal{M}$. We consider the case where one wants to test between two simple hypotheses

$$H_0 : P = P_0, \quad H_1 : P = P_1,$$

where $P_0, P_1 \in \mathcal{M}$. Since we are aware of possible deviations from the nominal distributions P_0, P_1, we capture that uncertainty by considering classes of distributions $\mathcal{P}_0, \mathcal{P}_1$, which are neighborhoods of P_0 and P_1 respectively. Then, we rewrite our hypotheses and the robust test decides between

$$H_0 : P \in \mathcal{P}_0, \quad H_1 : P \in \mathcal{P}_1.$$

Heuristically, we want to find the pair of distribution $(Q_0, Q_1) \in \mathcal{P}_0 \times \mathcal{P}_1$ which is the "most difficult" to test and build a test Φ of fixed level α for the according pair of simple hypotheses. Then, we expect that for every other pair of distributions in $\mathcal{P}_0 \times \mathcal{P}_1$, the test will have at most a level α and at least the same power as Φ. We will consider different types of uncertainty classes and show existence of this pair (Q_0, Q_1) that is called the *least favorable pair (LFP)*. Then we compare the structure of the obtained robust test.

In this chapter we have a look at robust tests for particular uncertainty sets, which were introduced in Huber and Strassen 1973. All the proofs of this chapter follow from the latter paper. First, we need to define *2-alternating capacities*. To get a better intuition of these measures, observe that for each subset $\mathcal{H} \subset \mathcal{M}$, one can define an *upper probability* v and a *lower probability* u as

$$v(A) = sup\{P(A) : P \in \mathcal{H}\}, \quad u(A) = inf\{P(A) : P \in \mathcal{H}\}, \quad \text{for } A \in \mathcal{F}.$$

Then, v and u are conjugate to each other, i.e. we have

$$u(A) + v(A^c) = 1. \tag{2.1}$$

It is easy to see that v satisfies the following properties:

1. $v(\varnothing) = 0, \quad v(\Omega) = 1$
2. for $A \subset B$, $v(A) \leq v(B)$
3. for any increasing sequence of sets (A_n) with $\lim_{n\to\infty} A_n = A$, one has $\lim_{n\to\infty} v(A_n) = v(A)$
4. if any sequence in $\mathcal{H}$ has a weakly convergent subsequence ($\mathcal{H}$ is weakly compact), then for any decreasing sequence of closed sets (A_n) with $\lim_{n\to\infty} A_n = A$, one has $\lim_{n\to\infty} v(A_n) = v(A)$.

Definition 2.1. We call $v : \mathcal{F} \to \mathbb{R}$ a *2-alternating capacity* if it satisfies $1. - 4.$ and if we additionally have that

$$5.\ \ v(A \cup B) + v(A \cap B) \leq v(A) + v(B) \tag{2.2}$$

We can now take the reverse path, and define for a 2-alternating capacity v its conjugate function $u : \mathcal{F} \to \mathbb{R}$ as $v(A) = 1 - u(A^c)$ for $A \in \mathcal{F}$. We call such a function u a *2-monotone capacity*. Such a u clearly also satisfies $1. - 2.$ It satisfies 3. if the sets are open and 4. for any sequence of decreasing sets. More importantly, we have

$$5'.\ \ u(A \cup B) + u(A \cap B) \geq u(A) + u(B) \tag{2.3}$$

Note that this last inequality implies that for every $A \in \mathcal{F}$, one has $u(A) \leq v(A)$ (take $B = A^c$).

Now that we have defined 2-alternating capacities, we can define the uncertainty classes for which we want to build a test. We consider tests for the composite hypotheses

$$\begin{aligned} H_0 &: P \in \mathcal{P}_0 = \{P \in \mathcal{M} : P(A) \leq v_0(A),\ \forall A \in \mathcal{F}\} \\ H_1 &: P \in \mathcal{P}_1 = \{P \in \mathcal{M} : P(A) \leq v_1(A),\ \forall A \in \mathcal{F}\}, \end{aligned} \tag{2.4}$$

where v_0 and v_1 are 2-alternating capacities. These are very general classes for which we will prove the results. Then these general results can be used for specific cases. Also, one can show (see Huber and Strassen 1973) that if the LFP exists, $\mathcal{P}_0$ and $\mathcal{P}_1$ can be written as in (2.4). However in practice, we only use particular cases of those classes, for instance, for nominal distributions $\mathcal{P}_0, \mathcal{P}_1$:

- ϵ-contamination:

$$\mathcal{P}_j = \{Q \in \mathcal{M} : Q \leq (1 - \epsilon_j)P_j + \epsilon_j\}, \quad 0 \leq \epsilon_j \leq 1, j = 0, 1$$

  We will focus on this model in Section 2.3.

- Kolmogorov:

$$\mathcal{P}_j = \{Q \in \mathcal{M} : |Q(x) - P_j(x)| \leq \delta_j, \forall\, x\}, \quad \delta_j > 0,\ j = 0, 1$$

- Total variation:

$$\mathcal{P}_j = \{Q \in \mathcal{M} : |Q(A) - P_j(A)| \leq \delta_j, \forall\, A \in \mathcal{F}\}, \quad \delta_j > 0,\ j = 0, 1$$

2.1 Existence of the Bayes test

For the hypotheses (2.4), we consider for $A \in \mathcal{F}$ the test

$$\Phi_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$$

We work with tests in this very theoretical form, defined on Ω, to simplify the notations in the proofs. In practice we would of course not use a test this way, but rather with a sample $\mathbf{X}$, a statistic T and a critical region $C = \{T(\mathbf{X}(\omega)) : \omega \in A\}$. The upper probabilities of error of type I and II are respectively given by

$$\sup_{P_0 \in \mathcal{P}_0} \mathbb{E}_{P_0} \Phi_A = v_0(A) \tag{2.5}$$

$$\sup_{P_1 \in \mathcal{P}_1} \mathbb{E}_{P_1}[1 - \Phi_A] = v_1(A^c) = 1 - u_1(A). \tag{2.6}$$

Our goal is to build a minimax test for the Bayes risk. Take as a prior distribution (on $\Theta = \{\mathcal{P}_0, \mathcal{P}_1\}$)

$$\Pi_t(\mathcal{P}_0) = \frac{t}{1+t}, \quad \Pi_t(\mathcal{P}_1) = \frac{1}{1+t},$$

with $0 \leq t \leq \infty$. Note that we cover all prior distributions as t goes from 0 to ∞, with the special cases $t = \infty$, where all the mass is put on $\mathcal{P}_0$, and $t = 0$, where all the mass is put on $\mathcal{P}_1$. Now, the upper Bayes risk is given by

$$\bar{r}(\Pi_t, \Phi_A) \coloneqq \sup_{(P_0, P_1) \in \mathcal{P}_0 \times \mathcal{P}_1} r(\Pi_t, \Phi_A) = \frac{t}{1+t} v_0(A) + \frac{1}{1+t}(1 - u_1(A)).$$

The aim is to find a test Φ_A, i.e. a critical set A which minimizes $\bar{r}(\Pi_t, \Phi_A)$. This is equivalent to minimizing

$$w(t) \coloneqq t\, v_0(A) - u_1(A).$$

Note and keep in mind that w is a 2-alternating capacity. The first result we are showing gives us some hope that we can find a minimax test.

Lemma 2.1. *For all $t \in [0, \infty]$, there exists $A_t \in \mathcal{F}$ such that*

$$w_t(A_t) = \inf_{A \in \mathcal{A}} w_t(A).$$

Hence, the infimum is reached. This lemma ensures the existence of a minimax critical set and hence of a minimax Bayes test.

Proof (from Huber and Strassen 1973).
For readability, we write w for w_t and $a \coloneqq inf_{A \in \mathcal{A}} w(A)$. By characterization of the infimum, one can choose $(\epsilon_n)_{n \geq 1} \subset (0, \infty)$ with $\sum_{n=1}^{\infty} \epsilon_n < \infty$ and $(A_n)_{n \geq 1} \subset \mathcal{F}$ such that

$$w(A_n) \leq a + \epsilon_n \,.$$

Using that w is by definition a 2-alternating capacity, we have for $n \leq m$

$$w(A_n \cup A_m) + w(A_n \cap A_m) \leq w(A_n) + w(A_m) \leq 2a + \epsilon_n + \epsilon_m \,.$$

Combining this with

$$w(A_n \cup A_m) \geq a, \quad w(A_n \cap A_m) \geq a$$

yields

$$w(A_n \cup A_m) \leq a + \epsilon_n + \epsilon_m, \quad w(A_n \cap A_m) \leq a + \epsilon_n + \epsilon_m \,.$$

By iterating this operation, we get that for fixed $n < N$

$$w\left(\bigcup_{n\leq m\leq N} A_m\right) \leq a + \sum_{n\leq m\leq N} \epsilon_m, \quad w\left(\bigcap_{n\leq m\leq N} A_m\right) \leq a + \sum_{n\leq m\leq N} \epsilon_m \,.$$

Letting $N \to \infty$ yields

$$w\left(\bigcup_{m\geq n} A_m\right) \leq a + \sum_{m\geq n} \epsilon_m, \quad w\left(\bigcap_{m\geq n} A_m\right) \leq a + \sum_{m\geq n} \epsilon_m, \tag{2.7}$$

where we used that $w(\lim G_n) \leq \lim w(G_n)$ for any monotone sequence of sets $(G)_n$, which follows from (2.2) and (2.3). We use that once again on $(\cup_{m\geq n} A_m)_n$, which is decreasing in n, and the convergence of the series $\sum \epsilon_n$ implies

$$\begin{aligned} w\left(\bigcap_{n\geq 1} \cup_{m\geq n} A_m\right) &= w\left(\lim_{n\to\infty} \bigcup_{m\geq n} A_m\right) \\ &\leq \lim_{n\to\infty} w\left(\bigcup_{m\geq n} A_m\right) \overset{(2.7)}{\leq} \lim_{n\to\infty} \left(a + \sum_{m\geq n} \epsilon_m\right) = a \end{aligned}$$

Since the left-hand side quantity is also obviously at least a, we have found

$$w\left(\bigcap_{n\geq 1} \bigcup_{m\geq n} A_m\right) = a \,.$$

□

Remark. *We can similarly obtain* $w(\bigcup_{n\geq 1} \bigcap_{m\geq n} A_m) = a$*, by using the second relation of* (2.7).

So we have proved that for any prior distribution on $\{\mathcal{P}_0, \mathcal{P}_1\}$, there is a critical set which minimizes the upper Bayes risk. Let us now see how these different sets are linked.

Lemma 2.2. *The sequence of minimizing sets* $(A_t)_{0\leq t\leq\infty}$ *can be chosen to be decreasing, i.e. such that* $A_t \subset A_s$, $\forall s < t$. *We then have for all* $0 < t < \infty$ *either*

$$A_t = \bigcup_{s>t} A_s \text{ or } A_t = \bigcap_{s<t} A_s.$$

Moreover,

$$A_0 = \bigcup_{t>0} A_t, \quad A_\infty = \bigcap_{t<\infty} A_t.$$

The reader can find the proof in the Appendix.

Remark. *The later implies that the more prior belief we put on H_0, the smaller will the probability of rejecting H_0, which matches with the intuition. It also tells us that the critical sets for different priors are contained in each other, which will be useful later.*

Now, say that we choose the minimizing sets as $A_t = \bigcup_{s>t} A_s$. With Lemma 2.2, it makes sense to define $\lambda : \Omega \to \mathbb{R}$ as

$$\lambda(\omega) \coloneqq \inf \{t : \omega \notin A_t\} \tag{2.8}$$

so that

$$\omega \in A_t \Leftrightarrow \lambda(\omega) > t. \tag{2.9}$$

Remark. *The function λ is the Radon-Nikodym derivative of v_1 with respect to v_0. Indeed, observe that for $t < s$, $A_s \subset B \subset A_t$*

$$\begin{aligned} & w_t(B) \geq w_t(A_t), \quad w_s(B) \geq w_s(A_s) \\ \Leftrightarrow \quad & t(v_0(B) - v_0(A_t)) \geq u_1(B) - u_1(A_t), \\ & s(v_0(B) - v_0(A_s)) \geq u_1(B) - u_1(A_s) \\ \Rightarrow \quad & \frac{u_1(A_t) - u_1(B)}{v_0(A_t) - v_0(B)} \geq t \qquad (2.10) \\ & \frac{u_1(B) - u_1(A_s)}{v_0(B) - v_0(A_s)} \leq s \qquad (2.11) \end{aligned}$$

where we assume that the fractions are well defined. The last two relations will be of importance, but we will not show that λ is the Radon-Nikodym derivative of $\frac{dv_1}{dv_0}$. The interested reader can use the development above and compare with the usual proof of the Radon-Nikodym theorem to prove it.

2.2 The main result

The main result of this chapter is the existence of a pair (Q_0, Q_1) in $\mathcal{P}_0 \times \mathcal{P}_1$ which represents the upper probabilities (v_0, v_1).

Theorem 2.1. *Let λ be defined as in* (2.8)*. There exist $Q_0 \leq v_0$, $Q_1 \leq v_1$ such that $\forall c \in \mathbb{R}$*

$$Q_0\{\lambda > c\} = v_0\{\lambda > c\} \tag{2.12}$$
$$Q_1\{\lambda > c\} = u_1\{\lambda > c\}, \tag{2.13}$$

Moreover, λ is also the Radon-Nikodym derivative $\frac{dQ_1}{dQ_0}$, i.e. the likelihood ratio when testing between Q_0 and Q_1.

A direct consequence is that we know exactly the form of the minimax test between $\mathcal{P}_0$ and $\mathcal{P}_1$.

Corollary 2.1. *The Neyman-Pearson test (NP test) of level α between Q_0 and Q_1*

$$\phi(X_1, ..., X_n) = \begin{cases} 1 & \text{if } \prod_{i=1}^n \lambda(X_i) > c \\ q & \text{if } \prod_{i=1}^n \lambda(X_i) = c \\ 0 & \text{if } \prod_{i=1}^n \lambda(X_i) < c \end{cases}$$

where q and c are such that $\mathbb{E}_{Q_0}\phi = \alpha$, is also a minimax test for $H_0 : P \in \mathcal{P}_0$, $H_1 : P \in \mathcal{P}_1$, with same level α and same upper probability of type II error (same minimum power).

Remark. *A very suitable feature of this result is that it is valid for any sample size n. It is thus interesting to use it even for a small amount of bad observations.*

Proof of Corollary 2.1 (based on Huber and Strassen 1973).

Recall that we want the minimax test ϕ^* between $\mathcal{P}_0$ and $\mathcal{P}_1$ to be such that

$$\inf_{\phi' \text{ test}} \sup_{(P_0,P_1)\in\mathcal{P}_0\times\mathcal{P}_1} r(\Pi_t, \phi') = r(\Pi_t, \phi^*), \quad \forall t \in [0, \infty].$$

In particular, if we fix a pair $(P_0, P_1) \in \mathcal{P}_0 \times \mathcal{P}_1$ and take $t = 0$, the test between P_0 and P_1 has to minimize the probability or type II error, or equivalently, must be of maximal power. Hence, we know from the Neyman-Pearson lemma that the test has to be a Neyman-Pearson test, i.e. the critical set A is of the form $\{\lambda \geq c\}$. Then, the upper probability of type I error is

$$v_0\{\lambda \geq c\} = Q_0\{\lambda \geq c\} = \alpha$$

and the minimum power is

$$1 - v_1\{\lambda < c\} = 1 - u_1\{\lambda \geq c\} = 1 - Q_1\{\lambda \geq c\}$$

which is the minimum power of ϕ as wanted. Thus, $\phi^* = \phi$. Note than we used the result of the theorem with a non-strict inequality. We will see in Lemma 2.4 that we can indeed do that. □

2.2.1 Preliminary results

To prove Theorem 2.1, we need the following tool. For $\mu : \mathcal{F} \to \mathbb{R}$ monotone and bounded, we define a function on the set of bounded and continuous functions on Ω. Let $\tilde{\mu} : C^b(\Omega) \to \mathbb{R}$ be defined as

$$\tilde{\mu}(f) \coloneqq \int_0^{+\infty} \mu\{\omega : f(\omega) > t\}\, dt. \tag{2.14}$$

Then, $\tilde{\mu}$ is positive, monotone and continuous. If μ is a probability measure, then $\tilde{\mu}$ simply is $\mathbb{E}_\mu f$. This function will be the key of our proof. Let us now look at a nice feature of $\tilde{\mu}$.

Lemma 2.3. *Let P be a probability measure and v a 2-alternating capacity with $P \leq v$. If $\tilde{P}$ satisfies $\tilde{P}(1) = 1$ and $\tilde{P}(f) \leq \tilde{v}(f)$ for all $f \in C^b(\Omega)$, then there is a one to one correspondence between P and $\tilde{P}$.*

The reader can find the proof in the Appendix. The next lemma is at the foundation of the proof of Theorem 2.1. It can seem to be sufficient to prove the theorem, but we will see in the proof that it is certainly not the case.

Lemma 2.4. *Let v be a 2-alternating capacity and h an upper semi-continuous function on Ω. Then there is a $Q \in \{P \in \mathcal{M} : P \leq v\}$ such that for all $t \in \mathbb{R}$, $Q\{h > t\} = v\{h > t\}$ and $Q\{h \geq t\} = v\{h \geq t\}$.*

The reader can find the proof in the Appendix.

Lemma 2.5. *Let v be a 2-alternating capacity. Then for each $A \in \mathcal{F}$ there is a $Q \in \{P \in \mathcal{M} : P \leq v\}$ such that $Q(A) = v(A)$.*

The reader can find the proof in the Appendix. A consequence of this last Lemma is that v is the upper probability for the set $\{P \in \mathcal{M} : P \leq v\}$. Also note that with the help of (2.1), it also gives the existence of $Q' \geq u$ such that $Q' = u$, where u is the conjugate of v.

2.2.2 Proof of Theorem 2.1

(From Huber and Strassen 1973).
As said above, the existence of Q_0 comes directly from Lemma 2.4. However, this Q_0 is not necessarily adequate, in the sense that the corresponding Q_1 given by $dQ_1 = \lambda\, dQ_0$ may not satisfy $Q \leq v_1$. Thus, our aim is to construct $Q_0 \leq v_0$ such that the corresponding Q_1 satisfies $Q \leq v_1$. Note that $Q \leq v_1$ is equivalent to $Q \geq u_1$, indeed:

$$Q(A) \leq v_1(A) \Leftrightarrow 1 - Q(A) \geq 1 - v_1(A) \Leftrightarrow Q(A^c) \geq u_1(A^c),$$

and the A^c cover $\mathcal{F}$ when A does. So, heuristically, we will choose Q_0 such that "$\pi^{-1}\, u_1 \leq Q_0 \leq v_0$".

Construction of Q_0:

Let $t < s$. For $B \in \mathcal{F}$, we define

$$F_{t,s}(B) \coloneqq u_1((B \cap A_t) \cup A_s) - u_1(A_s)$$
$$G_{t,s}(B) \coloneqq v_0((B \cap A_t) \cup A_s) - v_0(A_s).$$

Clearly, $F_{t,s}$ is a 2-monotone capacity and $G_{t,s}$ a 2-alternating capacity. Let us now derive some useful properties of these functions. Keep in mind the result of Lemma 2.2 which will be useful throughout the proof. First, we can in (2.10) and (2.11) take $(B \cap A_t) \cup A_s$ for B, since it is contained in A_s and contains A_t. This yields

$$F_{t,s}(B) \leq s\, G_{t,s}(B) \tag{2.15}$$
$$F_{t,s}(B) \geq t\, G_{t,s}(B) \tag{2.16}$$

Also, one has,

$$\begin{aligned} G_{t,s}(B) = v_0((B \cap A_t) \cup A_s) - v_0(A_s) &\overset{t\leq s}{=} v_0(B \cap A_t) - v_0(A_s) \\ &\leq v_0(B \cap A_t) - v_0(B \cap A_s) \qquad (2.17) \\ &\leq v_0(B). \qquad (2.18) \end{aligned}$$

Now, let $t_1 < t_2 < t_3$, so that $A_{t_3} \subset A_{t_2} \subset A_{t_1}$ and let v a 2-alternating capacity. Property (2.2) of capacities implies $v(A_{t_1} \cup A_{t_2}) \leq v(A_{t_1}) + v(A_{t_2}) - v(A_{t_1} \cap A_{t_2}) = v(A_{t_1})$. We have also $v(A_{t_2}) \leq v(A_{t_1})$ and $v(A_{t_2} \cup A_{t_3}) \leq v(A_{t_1} \cup A_{t_3})$. Hence, we get

$$v(A_{t_1} \cup A_{t_2}) - v(A_{t_2}) + v(A_{t_2} \cup A_{t_3}) - v(A_{t_3}) \leq v(A_{t_1} \cup A_{t_3}) - v(A_{t_3})$$

which implies

$$G_{t_1,t_3} \geq G_{t_1,t_2} + G_{t_2,t_3}. \qquad (2.19)$$

Proceeding similarly with 2-monotone capacities yields

$$F_{t_1,t_3} \leq F_{t_1,t_2} + F_{t_2,t_3}. \qquad (2.20)$$

Now, for $T = (t_0, t_1, ..., t_n) \subset (0, \infty)$ a finite increasing sequence, we define

$$u_1^T(A) \coloneqq \sum_{i=1}^{n} \frac{1}{t_i} F_{t_{i-1},t_i}(A), \quad A \in \mathcal{F}$$

Then, $\forall A \in \mathcal{F}$, using consecutively (2.15), (2.19) and (2.18):

$$u_1^T(A) \leq \sum_{i=1}^{n} G_{t_{i-1},t_i}(A) \leq G_{t_{i_0},t_n}(A) \leq v_0(A) \qquad (2.21)$$

so if we define

$$\bar{u}_1(A) \coloneqq \sup_{\text{T finite seq.}} u_1^T(A), \quad A \in \mathcal{F},$$

then $\bar{u}_1$ is a 2-monotone capacity with $\bar{u}_1 \leq v_0$.

We now try to understand deeper $\bar{u}_1$. Let $\alpha < 1$ and an increasing sequence $(t_i)_{i\geq 0} \subset [0, \infty)$ with $t_i > 0$ $\forall i > 0$, and with $\frac{t_{i-1}}{t_i} \geq \alpha$, $\forall i \geq 1$. We write $T_n = (t_0, t_1, ..., t_n)$, $n \geq 0$. Then, if $t = t_j$ for some j and $n > j$, we have

$$\begin{aligned} \bar{u}_1(A_t) \geq u_1^{T_n}(A_t) &\overset{(2.16)}{\geq} \sum_{i=1}^{n} \frac{t_{i-1}}{t_i} G_{t_{i-1},t_i}(A_t) \\ &\geq \alpha \sum_{i=1}^{n} G_{t_{i-1},t_i}(A_t) = \alpha \left(v_0(A_t) - v_0(A_{t_n})\right) \end{aligned}$$

since

$$G_{t_{i-1},t_i}(A_t) = \begin{cases} 0 & \text{if } t_i < t \\ v_0(A_t) - v_0(A_{t_i}) & \text{if } t_{i-1} \leq t \leq t_i \\ v_0(A_{t_{i-1}}) - v_0(A_{t_i}) & \text{if } t < t_{i-1} \end{cases}$$

because

$$\left(A_t \cap A_{t_{i-1}}\right) \cup A_{t_i} = \begin{cases} A_{t_i} & \text{if } t_i < t \\ A_t & \text{if } t_{i-1} \le t \le t_i \\ A_{t_{i-1}} & \text{if } t < t_{i-1}. \end{cases}$$

Thus, since $\bar{u}_1(A_t) \le v_0(A_t)$ by (2.21), and using (6.3), we obtain letting $\alpha \to 1$ and $n \to \infty$

$$\bar{u}_1(A_t) = v_0(A_t) \tag{2.22}$$

which is satisfied for all $t \in [0, \infty)$ by the arbitrary nature of the sequence $(t_i)_{i \ge 0}$. In particular, $\bar{u}_1(\Omega) = \bar{u}_1(A_0) = v_0(A_0)$. We can then assume without loss of generality that $\bar{u}_1(\Omega) = 1$, because if $v_0(A_0) < 1$, we just replace $\bar{u}_1$ by $\bar{u}_1 + Q|_{A_0^c}$, where $Q_0 \le v_0$ is a probability measure with $Q(A_0) = v_0(A_0)$ so that $\bar{u}_1(\Omega) = 1$.

We now construct the adequate Q_0. Define

$$U = \left\{ f \in C_+^b(\Omega) : \tilde{\bar{u}}_1(f) \ge 1 \right\},$$
$$V = \left\{ g \in C^b(\Omega) : \tilde{v}_0(|g|) \ge 1 \right\}$$

where $\tilde{\bar{u}}_1$ and $\tilde{v}_0$ are defined as in (2.14). U and V are convex and V is open, so by the separation theorem (Hahn-Banach), there is a function $\tilde{Q}_0$ on $C^b(\Omega)$ which separates U and V, such that

$$\tilde{Q}_0(g) < \tilde{Q}_0(f), \quad \forall f \in U, g \in V.$$

By definition, $\tilde{Q}_0$ is positive and without loss of generality, $\tilde{Q}_0(1) = 1$. Moreover, the last inequality implies

$$\tilde{\bar{u}} \le \tilde{Q}_0 \le \tilde{v}_0. \tag{2.23}$$

The assumptions of Lemma 2.3 being satisfied, $\tilde{Q}_0$ defines a probability measure Q_0 for which we have thanks to (2.22) and (2.23)

$$Q_0(A_t) = v_0(A_t),$$

which shows the first part because we know that A_t is of the form $\{\lambda > t\}$ (see proof of Corollary 2.1).

Validity of Q_1:

On A_∞: Lemma 2.5 gives us the existence of $Q \ge u_1$ with $Q(A_\infty) = u_1(A_\infty)$, so we can simply choose

$$Q_1|_{A_\infty} = Q|_{A_\infty} \tag{2.24}$$

On A_∞^c: we put

$$dQ_1 = \lambda \, dQ_0 \tag{2.25}$$

and we first check that $Q_1 \geq u_1$. Let $\alpha < 1$, $T = (t_0, ..., t_n)$ with $\frac{t_{i-1}}{i} \geq \alpha$, $i \geq 1$. Let $B \in \mathcal{F}$. We develop:

$$\begin{aligned} Q_1(B \cap A_{t_0}) - Q_1(B \cap A_{t_1}) &\geq Q_1\left((B \cap A_{t_0}) \cup A_{t_1}\right) - Q_1(A_{t_1}) = \\ Q\left(B \cap (A_{t_0} \setminus A_{t_1})\right) &\overset{(2.25)}{=} \int\limits_{B \cap (A_{t_0} \setminus A_{t_1})} \lambda \, dQ_0 \overset{(2.9)}{\geq} \int\limits_{B \cap (A_{t_0} \setminus A_{t_1})} t_0 \, dQ_0 \\ &= t_0 \, Q_0\left(B \cap (A_{t_0} \setminus A_{t_1})\right) \geq t_0 \, u_1^T\left(B \cap (A_{t_0} \setminus A_{t_1})\right) \\ &= t_0 \, u_1^T(B) = \frac{t_0}{t_1} \, F_{t_0,t_1}(B) \end{aligned}$$

So, iterating this for i and $i+1$ instead of 0 and 1, for $i = 0, ..., n-1$, we get

$$\begin{aligned} &Q_1(B \cap A_{t_0}) - Q_1(B \cap A_{t_n}) \geq Q_1\left((B \cap A_{t_0}) \cup A_{t_n}\right) - Q_1(A_{t_n}) \\ &\geq \alpha \sum_{i=1}^{n} F_{t_{i-1},t_i}(B) \overset{(2.20)}{\geq} \alpha \, F_{t_0,t_n}(B) = \alpha \left[u_1((B \cap A_{t_0}) \cup A_{t_n}) - u_1(A_{t_n})\right] \\ &\geq \alpha \left[u_1(B \cap A_{t_0}) - u_1(B \cap A_{t_n})\right] \end{aligned}$$

where last step comes from (2.17) in which we reverse the inequality sign because u is 2-monotone. Now, we let $\alpha \to 1$, $t_0 \to 0$, $t_n \to \infty$, and using the monotone convergence theorem, we have

$$Q_1(B \cap A_0) - Q_1(B \cap A_\infty) \geq u_1(B \cap A_0) - u_1(B \cap A_\infty) = u_1(B \cap A_0). \quad (2.26)$$

Yet,

$$\begin{aligned} u_1(B \cup A_0) + u_1(B \cap A_0) &\overset{(2.3)}{\geq} u_1(B) + u_1(A_0), \text{ and} \\ u_1(B \cup A_0) &= u_1(A_0) = 1 \\ \Rightarrow \quad u_1(B \cap A_0) &\geq u_1(B) \end{aligned}$$

so (2.26) becomes

$$Q_1(B \cap A_0) \geq u_1(B) + Q_1(B \cap A_\infty) \geq u_1(B)$$

and since $Q_1(B) \geq Q_1(B \cap A_0)$, we have obtained $Q_1 \geq u_1$ as wanted.

It remains to show that $Q_1(A_t) = u_1(A_t)$ $\forall \, t \in [0, \infty)$. We use a similar approach. Let $t \in [0, \infty)$. For $t_1 \leq t \leq t_2$

$$\begin{aligned} Q_1(A_{t_1}) - Q_1(A_{t_2}) &= \int\limits_{A_1 \setminus A_2} \lambda \, dQ_2 \leq t_2 \int\limits_{A_1 \setminus A_2} dQ_2 = \\ t_2 \, \left[Q_0(A_{t_1}) - Q_0(A_{t_2})\right] &= t_2 \, \left[v_0(A_{t_1}) - v_0(A_{t_2})\right] = \\ t_2 \, G t_1, t_2(A_t) &\overset{(2.16)}{\leq} \frac{t_2}{t_1} \, F t_1, t_2(A_t) = \frac{t_2}{t_1} \, \left[u_1(A_{t_1}) - u_1(A_{t_2})\right]. \end{aligned}$$

Again, for $\alpha < 1$, and for $t = t_0 < t_1 < ... < t_n$, with $\frac{t_i}{t_{i-1}} \leq \alpha$, $i \geq 1$, we sequentially get

$$Q_1(A_t) - Q_1(A_{t_n}) \leq \alpha \left[u_1(A_t) - u_1(A_{t_n}) \right]$$

so, letting $\alpha \to 1$ and $t_n \to \infty$ yields

$$Q_1(A_t) - Q_1(A_\infty) \leq u_1(A_t) - u_1(A_\infty)$$

and since $Q_1(A_\infty) = u_1(A_\infty)$, we indeed have

$$Q_1(A_t) = u_1(A_t).$$

□

2.3 Example: contamination

The developments and proofs of this section are based on Section 10.3 of Huber and Ronchetti 2009.

Uncertainty sets defined by 2-alternative capacities are very interesting. Indeed, we will see later that one can build location estimators with the help of the minimax tests we saw in this chapter. Moreover, a lot of the uncertainty classes that are usually considered can be written as in (2.4). In this section we focus on neighborhoods of the following form:

$$\begin{aligned} \mathcal{P}_0 &= \{(1-\epsilon_0)P_0 + \epsilon_0 H : H \in \mathcal{M}\}, \\ \mathcal{P}_1 &= \{(1-\epsilon_1)P_1 + \epsilon_1 H : H \in \mathcal{M}\} \end{aligned}$$

for some $0 \leq \epsilon_0, \epsilon_1 \leq 1$. That is, we know that some small proportion of the data (given by ϵ_j) does not come from the nominal distribution P_j. We say that our sample has been contaminated. The probability of an observation being a "bad" one is ϵ_j. This case will come often in the sequel, we call it *contamination*. $\mathcal{P}_0$ and $\mathcal{P}_1$ can be written as

$$\begin{aligned} \mathcal{P}_0 &= \{Q \in \mathcal{M} : Q \leq (1-\epsilon_0)P_0 + \epsilon_0\}, \\ \mathcal{P}_1 &= \{Q \in \mathcal{M} : Q \leq (1-\epsilon_1)P_1 + \epsilon_1\}. \end{aligned}$$

One can easily check that $v_j = (1-\epsilon_j)P_j + \epsilon_j$, $j = 0, 1$, satisfy the conditions of 2-alternating capacities. Hence, we know from Theorem 2.1 that there is a LFP (Q_0, Q_1). We will now explicitly construct the minimax test. We write the uncertainty sets in another equivalent form with which it is easier to work. Let X be a random variable and set

$$\begin{aligned} \mathcal{P}_0 &= \{Q \in \mathcal{M} : Q\{X < x\} \geq (1-\epsilon_0)P_0\{X < x\}\}, \\ \mathcal{P}_1 &= \{Q \in \mathcal{M} : Q\{X > x\} \geq (1-\epsilon_1)P_1\{X > x\}\}. \end{aligned} \tag{2.27}$$

We write q_j the density function of Q_j with respect to some dominating measure μ, for $j = 0, 1$. We will not derive q_0 and q_1 but we will first define them,

and then verify that their likelihood ratio test is indeed the minimax test. If $\epsilon_j = 1$, the solution is trivially $Q_j(A) = sup\{P(A) : P \in \mathcal{M}\}$ and if $\epsilon_j = 0$, then $q_j = p_j$ obviously does the job. Thus, we suppose $0 < \epsilon_0, \epsilon_1 < 1$. The densities will be such that there exist $x_0 < x_1$ such that they are equal to the boundaries of the uncertainty sets between x_0 and x_1, i.e.

$$q_j(x) = (1 - \epsilon_j)p_j(x), \quad \text{for } x_0 \leq x \leq x_1,\ j = 0, 1.$$

On $(-\infty, x_0)$ and (x_1, ∞) we want the likelihood ratio $\frac{q_1(x)}{q_0(x)}$ to be constant. More precisely, set

$$a_0 = \frac{\epsilon_0}{1 - \epsilon_0}, \quad a_1 = \frac{\epsilon_1}{1 - \epsilon_1}.$$

We write

$$c(x) = \frac{p_1(x)}{p_0(x)}$$

the nominal likelihood ratio and we assume it is monotone. Instead of distinguishing the cases $x < x_0$, $x_0 \leq x \leq x_1$, $x > x_1$, we will distinguish the cases $c(x) < r$, $r \leq c(x) \leq \frac{1}{s}$, $c(x) > \frac{1}{s}$ for some s, r such that $0 < r < \frac{1}{s}$, $s \neq 0$. Note that these two approaches are not equivalent, since $c(x)$ is not necessarily strictly monotone and continuous. However, $c(x)$ being monotone guarantees that the latter three regions are disjoint. Let us write $I_1 = \{x \in \mathbb{R} : c(x) < r\}$, $I_2 = \{x \in \mathbb{R} : r \leq c(x) \leq \frac{1}{s}\}$, $I_3 = \{x \in \mathbb{R} : c(x) > \frac{1}{s}\}$. We choose

$$q_0(x) = \begin{cases} (1 - \epsilon_0)\, p_0(x) & \text{on } I_1 \\ (1 - \epsilon_0)\, p_0(x) & \text{on } I_2 \\ (1 - \epsilon_0)\, s\, p_1(x) & \text{on } I_3, \end{cases}$$

$$q_1(x) = \begin{cases} (1 - \epsilon_1)\, r\, p_0(x) & \text{on } I_1 \\ (1 - \epsilon_1)\, p_1(x) & \text{on } I_2 \\ (1 - \epsilon_1)\, p_1(x) & \text{on } I_3. \end{cases} \tag{2.28}$$

Let us first check that there exist adequate r, s such that q_0, q_1 are well defined densities and that $Q_0 \in \mathcal{P}_0$ and $Q_1 \in \mathcal{P}_1$, where Q_0, Q_1 are the corresponding distribution functions. We discuss q_0, one can deal with q_1 in in a similar fashion.

Obviously, $q_0 \geq 0$. Now, let $0 < \epsilon_0 < 1$. We want

$$\int_{I_3} q_0\, d\mu = \int_{I_3} (1 - \epsilon_0) p_0\, d\mu + \epsilon_0 \tag{2.29}$$

so that

$$\int q_0\, d\mu = \int_{I_1 \cup I_2} (1 - \epsilon_0) p_0\, d\mu + \int_{I_3} (1 - \epsilon_0) p_0\, d\mu + \epsilon_0 = 1 - \epsilon_0 + \epsilon_0 = 1.$$

We can rewrite (2.29) as

$$\int_{I_3} [q_0 - (1-\epsilon_0)p_0]\, d\mu = \epsilon_0$$
$$\Leftrightarrow (1-\epsilon_0) \int_{I_3} [s\, p_1 - p_0]\, d\mu = \epsilon_0$$
$$\Leftrightarrow \int_{I_3} [s\, p_1 - p_0]\, d\mu = \frac{\epsilon_0}{1-\epsilon_0} \tag{2.30}$$

So if we let

$$f(z) = \int_{\{c(x)>1/z\}} [z\, p_1(x) - p_0(x)]\, d\mu(x),$$

we have to show that there exists z with $f(z) = \frac{\epsilon_0}{1-\epsilon_0}$. Let $h \in \mathbb{R}$ and compute

$$\begin{aligned} f(z+h) - f(z) = & \int_{\{1/(z+h)<c<=1/z\}} [(z+h)\, p_1 - p_0]\, d\mu + \\ & \int_{\{c>1/z\}} [(z+h)\, p_1 - p_0]\, d\mu - \int_{\{c>1/z\}} [z\, p_1 - p_0]\, d\mu \\ \overset{c=p_1/p_0}{=} & \int_{\{1/(z+h)<c<=1/z\}} p_0\, [(z+h)\, c - 1]\, d\mu + h \int_{\{c>1/z\}} c\, p_0\, d\mu \\ & \leq \left[(z+h)\, \frac{1}{z} - 1\right] \int p_0\, d\mu + h \int p_1\, d\mu = \frac{h(1+z)}{z}. \end{aligned}$$

From this we get

$$0 \leq \frac{f(z+h) - f(z)}{h} \leq \frac{1+z}{z}$$

Thus, $f(z)$ is monotone increasing, and continuous (because differentiable). Since $\lim_{z\to\infty} f(z) = \infty$ and $\lim_{z\to 0} f(z) = 0$, we get from the Intermediate value theorem that $\frac{\epsilon_0}{1-\epsilon_0} \in Imf = (0, \infty)$, as wanted.

Now that we know that q_0 is a well defined density, we check that $Q_0 \in \mathcal{P}_0$. Suppose that $q_0 < (1-\epsilon_0)p_0$ on I_3. Then, we would have $\int q_0\, d\mu < \int (1-\epsilon_0)p_0\, d\mu < 1$, which is a contradiction. Hence, $q_0 \geq (1-\epsilon_0)\, p_0$ so that $\forall x$

$$Q_0\{X<x\} = \int_{\{X<x\}} q_0\, d\mu \geq \int_{\{X<x\}} (1-\epsilon_0)p_0\, d\mu = (1-\epsilon_0)P_0\{X<x\}$$

which shows that $Q_0 \in \mathcal{P}_0$.

With the densities we found, the least favorable likelihood ratio is

$$\frac{q_1(x)}{q_0(x)} = \frac{1-\epsilon_1}{1-\epsilon_0}\,\tilde{\lambda}(x)$$

where

$$\tilde{\lambda}(x) = \begin{cases} r & \text{if } c(x) < r \\ c(x) & \text{if } r \leq c(x) \leq 1/s \\ 1/s & \text{if } 1/s < c(x) \end{cases} \tag{2.31}$$

As said before, this likelihood ratio is equal to the nominal one on a certain interval $[x_0, x_1]$ and constant above and below this interval. We can see it as a truncation of $c(x)$. It is interesting to see that the minimax test, which is a Neyman-Pearson test between Q_0 and Q_1 as we will see below, takes the same decisions than the NP test between P_0 and P_1 for some values of $c(x)$. However, when $c(x)$ takes very big or very small values, which implies an "easier" decision, we force the values of $\tilde{\lambda}$ to stay constant, which leads to "more difficult" decisions, with bigger probability of errors. This is of course not surprising for a least favorable pair of hypotheses.

Proposition 2.1. *Let Q_0, Q_1 the distribution functions of the densities in* (2.28)*. One has*

$$Q_0'\left\{\tilde{\lambda} < t\right\} \geq Q_0\left\{\tilde{\lambda} < t\right\} \quad \forall\, Q_0' \in \mathcal{P}_0,$$
$$Q_1'\left\{\tilde{\lambda} < t\right\} \leq Q_1\left\{\tilde{\lambda} < t\right\} \quad \forall\, Q_1' \in \mathcal{P}_1$$

Proof (from Huber and Ronchetti 2009 Sect. 10.3).
If $t \leq r$, then $\{\tilde{\lambda} < t\} = \varnothing$ and if $t > 1/s$, then $\{\tilde{\lambda} < t\} = \mathbb{R}$. Thus, both cases are trivial. On $\{r \leq c(x) \leq 1/s\}$, we chose Q_j to be the boundaries, so the inequalities are immediately fulfilled. One can look at how we described $\mathcal{P}_j$ in (2.27). □

Corollary 2.2. *The Neyman-Pearson test of level α between Q_0 and Q_1*

$$\phi(X_1, ..., X_n) = \begin{cases} 1 & \text{if } \prod_{i=1}^n \tilde{\lambda}(X_i) > C \\ \gamma & \text{if } \prod_{i=1}^n \tilde{\lambda}(X_i) = C \\ 0 & \text{if } \prod_{i=1}^n \tilde{\lambda}(X_i) < C \end{cases}$$

where q and γ are such that $\mathbb{E}_{Q_0}\phi = \alpha$, is also a minimax test for $H_0 : P \in \mathcal{P}_0$, $H_1 : P \in \mathcal{P}_1$, with same level α and same upper probability of type II error (same minimum power).

Proof Directly from Proposition 2.1. □

3 Minimax tests using Kullback-Leibler divergence

As we have seen, uncertainty sets defined by 2-alternating capacities are very powerful, since they include plenty of class types and provide very general results of existence, for any sample size. However, the tests provided in Chapter 2 are computationally difficult to implement. Let us now restrict our attention on other particular uncertainty classes, for which we will be able to numerically construct the minimax test. We consider neighborhoods of the nominal densities p_0, p_1, where the "distance" used is the Kullback-Leibler divergence. Precisely, for a random variable X we are interested of deciding what is its density p, between

$$H_0 : p \in \mathcal{G}_0 = \{\text{density function } g : \mathrm{KL}(g \,|\, p_0) \leq \epsilon_0\} \tag{3.1}$$
$$H_1 : p \in \mathcal{G}_1 = \{\text{density function } g : \mathrm{KL}(g \,|\, p_1) \leq \epsilon_1\}, \tag{3.2}$$

where $\epsilon_0, \epsilon_1 > 0$ and $\mathrm{KL}(g \,|\, f)$ denotes the *Kullback-Leibler divergence* or *relative entropy* of density functions f, g, defined as

$$\mathrm{KL}(g \,|\, f) = \int_{-\infty}^{\infty} \log\left(\frac{g(x)}{f(x)}\right) g(x)\, dx.$$

Note that it is not truly a distance because it is not symmetric and does not satisfy the triangle inequality. However, $\mathrm{KL}(g \,|\, f) \geq 0$ and is zero iff $f = g$. Also, $x \mapsto \log x$ is a convex function, so that $\mathrm{KL}(g \,|\, f)$ is convex in g, which in turn implies that $\mathcal{G}_0, \mathcal{G}_1$ are convex sets. Moreover, they are compact with respect to the KL distance. We have to be careful here because KL is not a metric, but one can show that $\mathrm{KL}(f_n \,|\, f) \to 0$ implies that $f_n \to f$ uniformly. Recall the Bayesian setting we introduced in Section 2.1. We now consider for simplicity the case $t = 1/2$, i.e. put equal mass on both hypotheses. Hence, our minimax problem is

$$\min_{\delta \in D} \max_{(g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1} r(\delta, g_0, g_1), \quad \text{where}$$
$$r(\delta, g_0, g_1) = \frac{1}{2}\left[\mathbf{E}_{g_0}\delta(X) + (1 - \mathbf{E}_{g_1}\delta(X))\right] \tag{3.3}$$

and where D is the set of randomized tests between H_0 and H_1. Observe that D is convex, since $\gamma\,\delta_1 + (1-\gamma)\,\delta_2$ clearly belongs to D for any $\delta_1, \delta_2 \in D$ and for all $0 \leq \gamma \leq 1$. Also, D is compact with respect to the norm $||.||_\infty$ since $\sup \delta \leq 1 \;\forall\, \delta \in D$.

The following developments are based on Levy 2009.

This time, the existence of a solution to the problem (3.3) comes more quickly

than in Chapter 2. Indeed, $r(\delta, g_0, g_1) = \frac{1}{2}\int \delta\, g_0 + \frac{1}{2}\int (1-\delta)\, g_1$ is linear in δ, g_0, g_1. In particular, $r(\cdot, g_0, g_1)$ is convex for fixed g_0, g_1, $r(\delta, \cdot, g_1)$ is concave for fixed δ, g_1, and $r(\delta, g_0, \cdot)$ is concave for fixed δ, g_0. Moreover, D and $\mathcal{G}_0 \times \mathcal{G}_1$ are convex and compact. Thus, the existence of a minimax solution (δ^*, q_0, q_1) comes directly from the Minimax theorem (von Neumann). This solution is such that for all $\delta' \in D, g_0 \in \mathcal{G}_0, g_1 \in \mathcal{G}_1$, one has

$$r(\delta^*, g_0, g_1) \leq r(\delta^*, q_0, q_1) \leq r(\delta', q_0, q_1). \tag{3.4}$$

Like in the previous chapter, when we know q_0, q_1, the form of the test follows. Indeed, for fixed q_0, q_1, the second inequality in (3.4) implies that δ^* is a Bayes test for the chosen prior, i.e. it minimizes

$$r(\delta, q_0, q_1) = \frac{1}{2}\int \delta\, q_0 + \frac{1}{2}\left(1 - \int \delta\, q_1\right) = \frac{1}{2}\int \delta\,(q_0 - q_1) + \frac{1}{2}$$

which is clearly minimized by

$$\begin{aligned}\delta^*(x) &= \begin{cases} 1 & \text{if } q_0(x) - q_1(x) < 0 \\ k & \text{if } q_0(x) - q_1(x) = 0 \\ 0 & \text{if } q_0(x) - q_1(x) > 0 \end{cases} \\ &= \begin{cases} 1 & \text{if } \lambda(x) > 1 \\ k & \text{if } \lambda(x) = 1 \\ 0 & \text{if } \lambda(x) < 1 \end{cases} \end{aligned} \tag{3.5}$$

with $\lambda(x) = q_1(x)/q_0(x)$, and k arbitrary.

Moreover, expanding the first inequality in (3.4) yields $\mathbb{E}_{g_0}\delta^* + (1 - \mathbb{E}_{g_1}\delta^*) \leq \mathbb{E}_{q_0}\delta^* + (1 - \mathbb{E}_{q_1}\delta^*)$, $\forall g_0 \in \mathcal{G}_0, g_1 \in \mathcal{G}_1$. This implies

$$\mathbb{E}_{g_0}\delta^* \leq \mathbb{E}_{q_0}\delta^* \tag{3.6}$$

$$1 - \mathbb{E}_{g_1}\delta^* \leq 1 - \mathbb{E}_{q_1}\delta^*. \tag{3.7}$$

This will help us characterize g_0, g_1, for a fixed δ^*, what will be used later to guess a minimax solution. We cover the derivation of q_0. Given δ^*, (3.6) tells that q_0 is found by maximizing $\mathbb{E}_{q_0}\delta^*$ over all $g_0 \in \mathcal{G}_0$. Thus, q_0 is the solution of a maximization under constraints problem, with constraints

$$\mathrm{KL}(g_0 \,|\, p_0) \leq \epsilon_0 \tag{3.8}$$

$$\int g_0 = 1 \tag{3.9}$$

We also must have $g_0 \geq 0$ but this will automatically be fulfilled with the maximization. We will proceed with the method of the Lagrange multipliers, or more precisely the Karush-Kuhn-Tucker theorem, since constraint (3.8) is an

inequality constraint. The Langragian is

$$\mathcal{L}(g_0, \lambda_1, \lambda_2) = \mathbb{E}_{g_0}\delta^* + \lambda_1 \, (\epsilon_0 - \mathrm{KL}(g_0 \,|\, p_0)) + \lambda_2 \left(1 - \int g_0\right) = \int_{-\infty}^{\infty} \left[\delta^*(x) - \lambda_1 \, \log\left(\frac{g_0(x)}{p_0(x)}\right) - \lambda_2\right] g_0(x)\, dx + \lambda_1 \, \epsilon_0 + \lambda_2.$$

We take the derivative with respect to g_0, in the direction of a function z (Gateaux derivative). Expanding the expression $\mathcal{L}(g_0 + h \cdot z, \lambda_1, \lambda_2)$ for some $h \in \mathbb{R}$, we find

$$\bigtriangledown_{g_0,z} \; \mathcal{L}(g_0, \lambda_1, \lambda_2) = \lim_{h \to 0} \frac{1}{h} \left[\mathcal{L}(g_0 + h \cdot z, \lambda_1, \lambda_2) - \mathcal{L}(g_0, \lambda_1, \lambda_2)\right] = \int_{-\infty}^{\infty} \left[\delta^*(x) - (\lambda_1 + \lambda_2) - \lambda_1 \, \log\left(\frac{g_0(x)}{p_0(x)}\right)\right] z(x)\, dx$$

Hence since z is arbitrary, $\bigtriangledown_{g_0,z} \, \mathcal{L}(g_0, \lambda_1, \lambda_2) = 0$ implies

$$\delta^* - (\lambda_1 + \lambda_2) - \lambda_1 \, \log\left(\frac{g_0}{p_0}\right) \equiv 0$$

which easily gives

$$q_0(x) = p_0(x) \, \exp\left[\frac{\delta^*(x) - (\lambda_1 + \lambda_2)}{\lambda_1}\right] = \frac{p_0(x)}{\eta_0} \, \exp(\mu_0 \, \delta^*(x)), \tag{3.10}$$

with $\eta_0 = \exp\left(1 + \frac{\lambda_2}{\lambda_1}\right)$ and $\mu_0 = \frac{1}{\lambda_1}$. Moreover, we have

$$\frac{\partial \mathcal{L}(g_0, \lambda_1, \lambda_2)}{\partial \lambda_1} = 0 \; \Rightarrow \; \mathrm{KL}(g_0 \,|\, p_0) = \epsilon_0$$

so that q_0 belongs to the boundary of $\mathcal{G}_0$ as expected. In a similar way, we get

$$q_1(x) = \frac{p_1(x)}{\eta_1} \, \exp\left(\mu_1 \, (1 - \delta^*(x))\right) \tag{3.11}$$

for some constants η_1, μ_1.

3.1 The minimax test

We will now describe the structure of the solution (δ^*, q_0, q_1) and show that it is indeed a solution to our minimax problem. We will assume the following:

(i) The nominal likelihood ratio

$$c(x) = \frac{p_1(x)}{p_0(x)}$$

is strictly increasing.

(ii) p_0 and p_1 are symmetric in the sense that $\forall x \in \mathbb{R}$

$$p_1(x) = p_0(-x).$$

Remember that we asked for monotonicity of c in Section 2.3. We are here more restrictive. Assumption (ii) is very strong, and will rarely be satisfied in practice. Song et al. 2018 constructed a minimax test without this assumption, but the proof is lengthy and not part of this thesis. Also, the test we will see suffices to get a flavor of the differences with the one we saw in the case of ϵ-contamination. We will need a density function that is at the same KL distance from p_0 and p_1. Define for $0 \leq u \leq 1$

$$p_u(x) = \rho_u \, p_0^{1-u}(x) \, p_1^u(x),$$

with $\rho_u = \left(\int p_0^{1-u} \, p_1^u\right)^{-1}$ the normalizing constant. Now, $p_{1/2}$ achieves the desired property. Indeed, one has

$$\mathrm{KL}(p_{1/2} \,|\, p_0) = \int \log\left(\frac{p_1^{1/2} \, p_0^{1/2} \, \rho_{1/2}}{p_0}\right) p_{1/2} = \frac{1}{2} \int \log\left(\frac{p_1}{p_0}\right) p_{1/2} + \log \rho_{1/2}$$

and since

$$\int_{-\infty}^{\infty} \log\left(\frac{p_1(x)}{p_0(x)}\right) dx = \int_{-\infty}^{\infty} \log\left(\frac{p_1(-x)}{p_0(-x)}\right) dx = \int_{-\infty}^{\infty} \log\left(\frac{p_0(x)}{p_1(x)}\right) dx,$$

we have

$$\mathrm{KL}(p_{1/2} \,|\, p_0) = \mathrm{KL}(p_{1/2} \,|\, p_1).$$

Now, as in Levy 2009, let $s > 0$ and the test

$$\delta^*(x) = \begin{cases} 0 & \text{if } x < -s \\ \frac{1}{2}\left(1 + \frac{\log c(x)}{\log c(s)}\right) & \text{if } -s \leq x \leq s \\ 1 & \text{if } x > s. \end{cases} \tag{3.12}$$

We also define the pair of density functions

$$q_0(x) = \frac{1}{k(s)} \cdot \begin{cases} p_0(x) & \text{if } x < -s \\ c^{1/2}(s) \, p_1^{1/2}(x) \, p_0^{1/2}(x) & \text{if } -s \leq x \leq s \\ c(s) \, p_0(x) & \text{if } x > s \end{cases} \tag{3.13}$$

$$q_1(x) = \frac{1}{k(s)} \cdot \begin{cases} c(s) \, p_1(x) & \text{if } x < -s \\ c^{1/2}(s) \, p_1^{1/2}(x) \, p_0^{1/2}(x) & \text{if } -s \leq x \leq s \\ p_1(x) & \text{if } x > s, \end{cases} \tag{3.14}$$

where for s fixed, the normalizing constant $k(s)$ is chosen so that $\int q_0 = \int q_1 = 1$.

Theorem 3.1. *Assume that $(i)-(ii)$ are satisfied, and that $\epsilon_0 = \epsilon_1 = \epsilon$, with*

$$0 < \epsilon < KL(p_{1/2} \,|\, p_0),$$

which ensures $\mathcal{G}_0 \cap \mathcal{G}_1 = \varnothing$. Then, there exists $s > 0$ such that for δ^*, q_0, q_1 defined as above, we have

$$KL(q_0 \,|\, p_0) = KL(q_1 \,|\, p_1) = \epsilon$$

and (δ^*, q_0, q_1) is a solution of the minimax problem (3.3).

Remark. *As we did for the ϵ-contamination model, we have a look at the least favorable likelihood ratio*

$$\lambda(x) = \begin{cases} c(s)\, c(x) & \text{if } x < -s \\ 1 & \text{if } -s \le x \le s \\ \frac{1}{c(s)}\, c(x) & \text{if } x > s \end{cases} \tag{3.15}$$

Like for ϵ-contamination, the least favorable likelihood ratio is a non-linear function of the nominal likelihood ratio. However, here we do not truncate the nominal LR at high and low values but rather force it to be close to 1. *It is not surprising. Indeed, when the LR is* 1, *the data does not provide any information and the decision is the hardest.*

Proof (from Levy 2009).
Take δ^*, q_0, q_1 as defined in (3.12), (3.13), (3.14) respectively. First, we see from (3.15) that $\lambda(x) < 1$ when $x < -s$ and $\lambda(x) > 1$ when $x > s$. Thus, δ^* has as wanted the form described in (3.5). Also, $0 \le \delta^* \le 1$ is ensured by the fact that $-1 \le \frac{\log c(x)}{\log c(s)} \le 1$.

Now, we verify that q_0 has for our δ^* the structure we found in (3.10). Observe that

$$\frac{p_0(x)}{\eta_0}\, \exp(\mu_0\, \delta^*(x)) = \frac{1}{\eta_0} \cdot \begin{cases} p_0(x) & \text{if } x < -s \\ \exp\left[\mu_0\, \frac{1}{2}\left(1 + \frac{\log c(x)}{\log c(s)}\right)\right] & \text{if } -s \le x \le s \\ e^{\mu_0}\, p_0(x) & \text{if } x > s \end{cases}$$

so we indeed have

$$q_0(x) = \frac{p_0(x)}{\eta_0}\, \exp(\mu_0\, \delta^*(x))$$

if we choose

$$\eta_0 = k(s), \quad \mu_0 = \log c(s).$$

Similarly we have that q_1 can be written as in (3.11).

Finally, we show the existence of an s such that q_0 belongs to the boundary of $\mathcal{G}_0$, i.e. such that $\mathrm{KL}(q_0 \,|\, p_0) = \epsilon$. One can then reproduce the same development to get that q_1 belongs to the boundary of $\mathcal{G}_1$. We write $q_0 = q_0^s$

to emphasise the dependence of q_0 on s, and we look at the distance between q_0^s and p_0 as a function of s. Using the definition of the KL divergence and simplifying the expression, we find

$$d(s) \coloneqq \mathrm{KL}(q_0^s \,|\, p_0) = -\log k(s) + \\ \frac{1}{k(s)} \left[c(s)\, \log c(s) \int_s^\infty p_0(x)\, dx + c^{1/2}(s)\, \log c(s) \int_0^s p_1^{1/2}(x)\, p_0^{1/2}(x)\, dx \right].$$

Note that our symmetry assumption (ii) was useful above to write the integral between $-s$ and s as 2 times the integral between 0 and s. Let us now differentiate this expression:

$$\frac{d}{ds} d(s) = -\frac{1}{k(s)} \frac{dk(s)}{ds} \\ - \frac{1}{k^2(s)} \frac{dk(s)}{ds} \left[c(s)\, \log c(s) \int_s^\infty p_0(x)\, dx + c^{1/2}(s)\, \log c(s) \int_0^s p_1^{1/2}(x)\, p_0^{1/2}(x)\, dx \right] \\ + \frac{1}{k(s)} \left[\frac{d}{ds} \left(c(s)\, \log c(s) \right) \int_s^\infty p_0(x)\, dx + \frac{d}{ds} \left(c^{1/2}(s)\, \log c(s) \right) \int_0^s p_1^{1/2}(x)\, p_0^{1/2}(x)\, dx \right]$$

We used Leibniz's rule to get that the derivatives w.r.t. s of the integrals are zero. We can rewrite this as

$$\frac{d}{ds} d(s) = \frac{R(s)}{k^2(s)} \frac{dc(s)}{ds},$$

with

$$R(s) = \log c(s) \int_s^\infty p_0(x)\, dx \int_s^\infty p_1(x)\, dx + \\ \frac{1}{2} \log c(s) \left[c^{1/2}(s) \int_0^s p_0^{1/2}(x)\, p_1^{1/2}(x)\, dx + \int_0^s \left(p_0(x) + \frac{1}{c(s)}\, p_1(x) \right) dx \right] > 0$$

so that, using also assumption (i):

$$\frac{d}{ds} d(s) > 0$$

Hence, $d(s)$ is strictly increasing, and continuous (because differentiable). Now, note that $q_0^0 = p_0$. Thus,

$$d(0) = \mathrm{KL}(q_0^0 \,|\, p_0) = 0.$$

In the same way, $q_0^{+\infty} = p_{1/2}$, which implies

$$\lim_{s\to\infty} d(s) = \mathrm{KL}(q_0^{+\infty} \,|\, p_0) = \mathrm{KL}(p_{1/2} \,|\, p_0).$$

Recall that we assumed $\mathrm{KL}(q_0^{+\infty} \,|\, p_0) > \epsilon$. Hence, by the Intermediate value theorem, there exists an s such that

$$d(s) = \epsilon.$$

□

3.2 Performance of the test on simulated data

The main advantage of the approach we discuss in this chapter is that the minimax tests can be easily implemented on a computer. Let us now view a practical example and see how the test works. Let $X_1, X_2, ..., X_n$ be n independent and identically distributed (i.i.d.) copies of a random variable X. We consider the location problem, and want to test between

$$H_0 : X \sim N(-1, \sigma^2), \quad H_1 : X \sim N(1, \sigma^2)$$

for σ^2 known. First, let us look at the least favorable pair of density functions in this case. We have

$$p_0(x) = (2\pi\sigma)^{-1/2} \exp\left\{-\frac{(x+1)^2}{2\sigma^2}\right\}, \;\; p_1(x) = (2\pi\sigma)^{-1/2} \exp\left\{-\frac{(x-1)^2}{2\sigma^2}\right\}$$

We then have

$$p_{1/2}(x) = (2\pi\sigma)^{-1/2} \exp\left(\frac{-x^2}{2\sigma^2}\right)$$

and it is no surprise that the mid-way distribution is $N(0, \sigma^2)$. Moreover, the nominal likelihood ratio

$$c(x) = \frac{p_1(x)}{p_0(x)} = \exp\left(\frac{2x}{\sigma^2}\right)$$

is strictly increasing, so assumption (i) is satisfied. Assumption (ii) is also clearly fulfilled. Thus, we know exactly the form of the minimax solution (δ^*, q_0, q_1). For instance, we can use (3.13) to compute

$$q_0(x) = \frac{1}{k(s)} \cdot \begin{cases} p_0(x) & \text{if } x < -s \\ e^{\frac{2s-1}{2\sigma^2}} \, (2\pi\sigma)^{-1/2} \, p_{1/2}(x) & \text{if } -s \le x \le s \\ e^{\frac{2s}{\sigma^2}} \, p_0(x) & \text{if } x > s \end{cases}$$

So we see that in order to get the least favorable distribution under the null, we attenuate the nominal density p_0 on $(-\infty, -s)$, we scale the mid-way density $p_{1/2}$ on $[-s, s]$ and we amplify the nominal density p_0 on (s, ∞). Attenuating

the density on the left and amplifying it on the right will shift mass on $[-s, s]$, where the least favorable LR is 1 and the data uninformative, and shift mass to the right, where we reject H_0, hence increase the probability of type I error, under the null.

Let us now analyze the robustness of the minimax test δ^* defined in (3.12) versus the classical Neyman-Pearson test. For that purpose, we estimate by simulation the probability of type I error of each test, under multiple null distributions, starting with the nominal, and contaminating it more and more. For the robust test, we describe the uncertainty we have in the distributions with the Kullback-Leibler divergence model (see (3.1)). We fix $\sigma^2 = 1$ and choose $\epsilon_0 = \epsilon_1 = \epsilon = 0.125$, i.e $\mathcal{G}_j = \{\text{density function } g : \mathrm{KL}(g \,|\, p_j) \leq 0.125\}$, $j = 0, 1$. We simulate data under four different distributions:

- $P_{0,1} = P_0 = N(-1, 1)$, the nominal null distribution
- $P_{0,2} = N(-0.99, 1)$
- $P_{0,3} = N(-0.9, 1)$
- $P_{0,4} = N(-0.75, 1)$
- $P_{0,5} = N(-0.6, 1)$
- $P_{0,6} = N(-0.5, 1)$

We call $p_{0,i}$ the density function of the distribution $P_{0,i}$, $i = 1, 2, ..., 6$. One can easily compute

$$\mathrm{KL}(p_{0,6} \,|\, p_0) = \epsilon$$

Hence, $p_{0,6}$ lies at the boundary of $\mathcal{G}_0$. Also, $p_{0,i} \in \mathcal{G}_0$, $i = 1, 2, ..., 6$, so that all $p_{0,j}$ are valid null densities for the robust test of H_0' : true density is in $\mathcal{G}_0$, H_1' : true density is in $\mathcal{G}_1$.

We compare the performance of two tests. We set $\alpha = 0.05$. Let δ^{NP} be the Neyman-Pearson test of level α for the hypotheses H_0 : true density is p_0, H_1 : true density is p_1. For our nominal distributions, it is given by

$$\delta^{NP}(X_1, ..., X_n) = \begin{cases} 1 & \text{if } \sum_{i=1}^n X_i \geq \gamma \\ 0 & \text{if } \sum_{i=1}^n X_i < \gamma \end{cases}$$

where $\gamma = -n + \sqrt{n}\, \Phi^{-1}(1-\alpha)$ ensures a significance level of α, with $\Phi^{-1}(1-\alpha)$ being the $(1-\alpha)$-quantile of the standard normal distribution. On the other hand, all assumptions of Theorem 3.1 are satisfied, so we know that there exists s such that the test

$$\delta^*(x) = \begin{cases} 0 & \text{if } x < -s \\ \frac{1}{2}\left(1 + \frac{\log c(x)}{\log c(s)}\right), & \text{if } -s \leq x \leq s \\ 1 & \text{if } x > s \end{cases}$$

is minimax. However, we still need to find that s. We know that it is such that

$$\mathrm{KL}(q_0 \,|\, p_0) = 0.125, \tag{3.16}$$

for q_0 as in (3.13). Hence, we solve the latter equation numerically and find $s = 0.4586459$.

Now that δ^* is completely defined, we sample $n = 100$ observations of the distribution $P_{0,i}$, for $i = 1, 2, ..., 6$. Then, we run both our tests $n_{sim} = 10000$ times and look at the frequency of rejection of H_0. Since the true distribution is $P_{0,i}$, this ratio is an estimate of the probability of type I error. Table 1 summarizes the results we get. We first look at the results for δ^{NP}. Observe

	$P_{0,1}$	$P_{0,2}$	$P_{0,3}$	$P_{0,4}$	$P_{0,5}$	$P_{0,6}$
δ^{NP}	0.0494	0.0586	0.2592	0.8072	0.9926	0.9996
δ^*	0	0	0	0.0004	0.0038	0.0212

Table 1: Frequencies of rejection of the null

that, as expected, the ratio of false rejections of H_0 is approximately α under the nominal distribution. It is important to highlight that a tiny deviation from the nominal distribution suffices to exceed α (see the result for $P_{0,2}$). Also, if the deviation is a bit bigger, the number of times the tests falsely rejects H_0 explodes. The test almost reaches 100% of errors at the boundary of $\mathcal{G}_0$. The advantage of δ^* is glaring. Working with a minimax test for $\mathcal{G}_0$ versus $\mathcal{G}_1$ ensures to control the significance level over the whole uncertainty classe! Of course, this robustness has a price: δ^{NP} has a bigger power than δ^* if the true distribution is the nominal distribution. Indeed, the test δ_{NP} has an estimated power of 1 under each distribution, and δ^* showed an estimated power of 0.9986, 0.9232 and 0.6545 under $P_{0,4}, P_{0,5}$ and $P_{0,6}$ respectively.

4 Minimax location estimators derived from tests

4.1 From tests to equivariant estimators

It is well know that there exists a duality between tests and confidence intervals. We can thus legitimately ask ourselves if the minimax property of the robust tests we constructed in the previous chapters also translates in the confidence intervals context. In that purpose, we construct estimators from the tests we have, inspiring ourselves with the duality mentioned above. The developments and proofs follow from Huber 1968 and Huber and Ronchetti 2009 (10.6-10.7).

We consider the location estimation problem. Let $X_1, ..., X_n$, with $X_i \sim F_i(x-\theta)$, where only the location parameter θ is unknown. Write $\mathbf{X} = (X_1, ..., X_n)$ and consider the following test between $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$:

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{if } h(\mathbf{X}) > C \\ \gamma & \text{if } h(\mathbf{X}) = C \\ 0 & \text{if } h(\mathbf{X}) < C \end{cases}$$

where $h(\mathbf{x}+\theta) = h(x_1+\theta, ..., x_n+\theta)$ is assumed to be increasing in θ and C is chosen such that the level of the test is $\mathbb{E}_{\theta_0}\varphi = \alpha$. Now we define two estimators

$$\begin{aligned} T_1(\mathbf{X}) &= \sup\{\theta : h(\mathbf{X}-\theta) > C\} \\ T_2(\mathbf{X}) &= \inf\{\theta : h(\mathbf{X}-\theta) < C\}, \end{aligned} \tag{4.1}$$

where $h(\mathbf{X}-\theta)$ will mean $h(X_1-\theta, ..., X_n-\theta)$ in the whole chapter. We now recall an important concept.

Definition 4.1. A statistic T is called *location equivariant* if $\forall\, c \in \mathbb{R}$ and $\forall\, \mathbf{x} = (x_1, ..., x_n)$ one has

$$T(\mathbf{x}+c) = T(\mathbf{x}) + c$$

We will in the sequel simply say *equivariant*, because we only talk about the location case. Equivariant estimators are very pleasant, because if we reparametrize our data, we simply need to reparametrize the estimate and do not need to recompute it. Now, note that T_1 and T_2 are equivariant since h is increasing. We define a randomized estimator

$$T^0 = \begin{cases} T_1 & \text{with probability } 1-\gamma \\ T_2 & \text{with probability } \gamma \end{cases} \tag{4.2}$$

which is thus also equivariant. We have $T_1 \leq T_2$ and it is easy to see that

$$\begin{aligned} \{\mathbf{x} : T_1(\mathbf{x}) > \theta\} &\subset \{\mathbf{x} : h(\mathbf{x}-\theta) > C\} \\ \{\mathbf{x} : T_2(\mathbf{x}) > \theta\} &\subset \{\mathbf{x} : h(\mathbf{x}-\theta) \geq C\}. \end{aligned} \tag{4.3}$$

and

$$\{\mathbf{x} : T_1(\mathbf{x}) < \theta\} \subset \{\mathbf{x} : h(\mathbf{x} - \theta) \leq C\}$$
$$\{\mathbf{x} : T_2(\mathbf{x}) < \theta\} \subset \{\mathbf{x} : h(\mathbf{x} - \theta) < C\}. \tag{4.4}$$

Then, for any distribution of $\mathbf{X} = (X_1, ..., X_n)$ and $\forall\, \theta$:

$$\begin{aligned} P(T^0 > \theta) &= (1-\gamma)\, P(T_1 > \theta) + \gamma\, P(T_2 > \theta) \\ &\overset{(4.3)}{\leq} (1-\gamma)\, P(h(\mathbf{X} - \theta) > C) + \gamma\, P(h(\mathbf{X} - \theta) \geq C) = \mathbb{E}\, \varphi(\mathbf{X} - \theta) \end{aligned} \tag{4.5}$$

and

$$\begin{aligned} P(T^0 < \theta) &= (1-\gamma)\, P(T_1 < \theta) + \gamma\, P(T_2 < \theta) \\ &\overset{(4.4)}{\leq} (1-\gamma)\, P(h(\mathbf{X} - \theta) \leq C) + \gamma\, P(h(\mathbf{X} - \theta) < C) \\ &= 1 - \mathbb{E}\, \varphi(\mathbf{X} - \theta) \end{aligned} \tag{4.6}$$

4.2 Minimaxity

We now have all the tools we need to seek for minimaxity. We consider the following framework. Our sample $X_1, ..., X_n$ is believed to be i.i.d with $X_i \sim G(x-\theta)$, $\theta \in \Theta$ unknown. We suppose G is a continuous distribution with $-\log g$ strictly convex on its convex support, where g is the density of G. However, as for our tests, there is uncertainty about the latter statement, so we say that X_i has distribution $F_i(x - \theta)$ with possibly different F_i, $i = 1, ..., n$. We still assume that $X_1, ..., X_n$ are independent. The F_i are supposed to belong to the following neighborhood of G:

$$\mathcal{P} = \{F : (1 - \epsilon_0)\, G(x) - \delta_0 \leq F(x) \leq (1 - \epsilon_1)\, G(x) + \epsilon_1 - \delta_1,\, \forall\, x\} \tag{4.7}$$

Observe that this class covers contamination and Kolmogorov classes as special cases.

Our goal is now to find an estimator T that minimizes the probability of being far from θ by more than a, for $a > 0$ fixed. I.e., we want to minimize

$$\sup_{P \in \mathcal{P}, \theta \in \Theta} \max \left\{ P(T < \theta - a),\, P(T > \theta + a) \right\}. \tag{4.8}$$

To build our test, we consider shifted distributions G_{-a} and G_{+a}, that we define through their densities

$$g_{-a}(x) = g(x + a), \quad g_{+a}(x) = g(x - a)$$

Thus, G_{-a} and G_{+a} are obtained by shifting G by amount a to the left and to the right respectively. They will be our nominal distributions, i.e. $P_0 = G_{-a}$, $P_1 = G_{+a}$. For the robust test, we consider the neighborhoods

$$\mathcal{P}_0 = \{Q : Q(\mathbf{X} < \mathbf{x}) \geq (1 - \epsilon_0) P_0(\mathbf{X} < \mathbf{x}) - \delta_0, \forall x\}$$
$$\mathcal{P}_1 = \{Q : Q(\mathbf{X} < \mathbf{x}) \geq (1 - \epsilon_1) P_1(\mathbf{X} < \mathbf{x}) - \delta_1, \forall x\},$$

with $0 < \epsilon_0, \epsilon_1, \delta_0, \delta_1 < 1$. When $\delta_0 = \delta_1 = 0$, it is the same case as in Section 2.3. For general δ_0, δ_1, the LFP is only very slightly different that the one we found, but Corollary 2.2 is still valid (see Huber and Ronchetti 2009 (10.3)), because

$$c(x) = \frac{g(x-a)}{g(x+a)}$$

is strictly increasing due to our condition on $-\log g$, so the condition needed is fulfilled. Thus, we know that the test

$$\varphi(X_1, ..., X_n) = \begin{cases} 1 & \text{if } \prod_{i=1}^n \tilde{\lambda}(X_i) > C \\ \gamma & \text{if } \prod_{i=1}^n \tilde{\lambda}(X_i) = C \\ 0 & \text{if } \prod_{i=1}^n \tilde{\lambda}(X_i) < C \end{cases}$$

is minimax for testing between $\mathcal{P}_0$ and $\mathcal{P}_1$. Recall that $\tilde{\lambda}$ is proportional to the likelihood ratio of the least favorable pair (Q_0, Q_1). We choose C and γ such that

$$\mathbb{E}_{Q_0} \varphi = \mathbb{E}_{Q_1} (1 - \varphi) = \alpha, \tag{4.9}$$

i.e. we choose the test with same probability of error of each type.

Now, like we did in the beginning of the chapter, we can derive an equivariant estimator from the test φ. With $h(\mathbf{x}) = \prod \tilde{\lambda}(x_i)$, we take T^0 as it is defined by (4.1) and (4.2), and we have the duality result:

Theorem 4.1. *If the distributions of the $X_i - \theta$ are contained in $\mathcal{P}$, then for any θ one has*

(i)

$$P(T^0 < \theta - a) \leq \alpha,$$
$$P(T^0 > \theta + a) \leq \alpha$$

(ii) The bound α is the best possible for equivariant estimators.

Hence, the probability of overshooting or undershooting θ by an amount a is at most α, for any underlying distribution in $\mathcal{P}$. In particular, the interval $(T^0 - a, T^0 + a)$ is a confidence interval of level (at most) $1 - 2\alpha$, robustly over all possible distributions in $\mathcal{P}$.

Proof (from Huber 1968 and Huber and Ronchetti 2009 Sect. 10.7).
By construction, T^0 satisfies (4.5) and (4.6) for any distribution of $X_1, ..., X_n$ and for any $\theta \in \Theta$. In particular, switching to distributions of $X_1 - \theta, ..., X_n - \theta$, we have for any $Q_0' \in \mathcal{P}_0$ and any $Q_1' \in \mathcal{P}_1$

$$Q_0'(T^0 > 0) \leq \mathbb{E}_{Q_0'} \varphi(\mathbf{X}) \leq \alpha \tag{4.10}$$

$$Q_1'(T^0 < 0) \leq \mathbb{E}_{Q_1'} (1 - \varphi(\mathbf{X})) \leq \alpha, \tag{4.11}$$

where the last two inequalities are valid because φ is minimax between $\mathcal{P}_0$ and $\mathcal{P}_1$. Also, it is clear that

$$Q_0 \in \mathcal{P}_{-a} \subset \mathcal{P}_0 \tag{4.12}$$

$$Q_1 \in \mathcal{P}_{+a} \subset \mathcal{P}_1, \tag{4.13}$$

where $\mathcal{P}_{-a}$ and $\mathcal{P}_{+a}$ are the set of distributions in $\mathcal{P}$ shifted by an amount a to the left and right respectively. Now, if the distributions of the $X_i - \theta$ are contained in $\mathcal{P}$, there exists $Q_0'' \in \mathcal{P}_0$ such that

$$P(T^0 > \theta + a) = Q_0''(T^0 > a) = Q_0'(T^0 > 0)$$

where $Q_0' \in \mathcal{P}_{-a}$ corresponds to Q_0'' shifted to the left by a. We also used the equivariance of T^0 to get the last equality. Now, $Q_0' \in \mathcal{P}_0$ so that $Q_0'(T^0 > 0) \leq \alpha$ by (4.12) and we get by (4.10)

$$P(T^0 > \theta + a) \leq \alpha$$

as wanted. In the same way, there exists $Q_1'' \in \mathcal{P}_1$ and its shifted to the right version $Q_1' \in \mathcal{P}_{+a} \subset \mathcal{P}_1$ such that

$$P(T^0 < \theta - a) = Q_1''(T^0 < -a) = Q_1'(T^0 < 0) \leq \alpha$$

and (i) is proved.

Let us now prove (ii). First, recall that the distributions of $X_1, ..., X_n$ are assumed to be continuous. Hence, since T^0 is equivariant and function of $X_1, ..., X_n$, its distribution is also continuous. Thus,

$$Q_0(T^0 = 0) = Q_1(T^0 = 0) = 0. \tag{4.14}$$

But for any statistic T, we must have

$$\max\{Q_0(T \geq 0), Q_1(T \leq 0)\} \geq \alpha$$

because we can see T as a test statistic for testing between Q_0 and Q_1, and from (4.9) we know that the minimax risk is α. So if T also satisfies (4.14), we have

$$\max\{Q_0(T > 0), Q_1(T < 0)\} \geq \alpha. \tag{4.15}$$

In particular, T^0 satisfies (4.15) and (ii) is proved. □

4.3 The Gaussian case

First, let us simply assume that G is symmetric, $\epsilon_0 = \epsilon_1$ and $\delta_0 = \delta_1$. Then, because of the assumption (4.9), we must have $C = 1$, $\gamma = 1/2$ (see how we found the test (3.5) and use symmetry assumption). Therefore, our test can be written as $\varphi(x) = \mathbb{1}\{q_1(x)/q_0(x) > 1\}$ or equivalently $\varphi(x) = \mathbb{1}\{\log \frac{q_1(x)}{q_0(x)} > 0\}$. Thus, we can take $h(x) = \log \frac{q_1(x)}{q_0(x)}$ and $C = 0$ instead of 1. Then, by definition (see (4.1)), T_1 and T_2 are the smallest and largest solutions of

$$\sum_{i=1}^{n} h(x_i - T) = 0,$$

i.e. they can be seen as M-estimators with derivative of loss function $\psi(x) = h(x)$. And since T^0 randomizes between T_1 and T_2, it can also be seen as an M-estimator. Also observe that in this case, the least favorable likelihood ratio has the same form as in (2.31), so we have

$$\psi(x) = \begin{cases} -k & \text{if } c(x) < -k \\ \log c(x) & \text{if } -k \le c(x) \le k \\ k & \text{if } c(x) > k \end{cases} = \max\{-k, \min\{k, \log c(x)\}\},$$

where $c(x) = \frac{g(x-a)}{g(x+a)}$.

In the special case $G = \Phi$ is the standard normal distribution, $\log \frac{g(x-a)}{g(x+a)}$ simplifies to $2ax$ and we get as a derivative of the loss function:

$$\psi(x) = \max\{-k, \min\{k, 2ax\}\} = \max\{-k', \min\{k', x\}\} \tag{4.16}$$

We will meet this function ψ in the next chapter!

Remark. *In fact one can observe that T_1 and T_2 are the same with high probability, so the non-randomized estimator $\frac{1}{2}(T_1 + T_2)$ can appear to have better properties than T^0. However, it has been shown by Huber 1968 that $\frac{1}{2}(T_1 + T_2)$ is not minimax!*

5 Asymptotically minimax location estimators

We have seen that finite sample results exist for two cases: robust version of the Neyman-Pearson lemma (see (2.1)) and robust confidence intervals (see Chapter 4). One may want to develop estimators that have nice asymptotic properties, robustly over a set of possible underlying distributions. That is the topic of this chapter. We treat only the case of location estimation. Suppose we have an infinite number of observations $X_1, X_2, ...$ i.i.d. with distribution $F_0(x-\theta)$, $\theta \in \Theta$ unknown, and F_0 lying in a certain set of distributions $\mathcal{P}$.

We recall the notion of *empirical distribution*. For each $n \geq 1$, it is the function defined as

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

The law of large numbers implies that $F_n(x) \to F(x)$ $\mathbb{P}_\theta - a.s$ for all $x \in \mathbb{R}$. We write T_n a statistic $T_n(X_1, ..., X_n)$ depending on the first n observations and we will assume that all our estimators are function of the empirical distributions, i.e. for any estimator T_n there is a function T such that

$$T_n = T(F_n).$$

Almost any estimator is of that form, so this is a very mild condition.

5.1 Minimax asymptotic bias

Let us first try to minimize the maximal asymptotic bias over $\mathcal{P}$. We consider the neighbourhoods

$$\mathcal{P} = \mathcal{P}_\epsilon = \{F = (1-\epsilon)\,\Phi + \epsilon\, H,\, H \in \mathcal{M}\}, \quad 0 \leq \epsilon \leq 1,$$

with Φ the standard normal distribution, i.e. we suppose $F_0 = \Phi$. This means that each observation is standard normal with probability $1-\epsilon$, and has another distribution with probability ϵ. For a sequence of estimators $T_n = T(F_n)$, its *maximum asymptotic bias* (over $\mathcal{P}_\epsilon$) is

$$b(\epsilon) = \sup_{F \in, \mathcal{P}_\epsilon} |T(F) - T(\Phi)|.$$

Intuitively, $b(\epsilon)$ measures the limit distance between the estimate and the true parameter, when the distribution of the observations covers the neighborhood $\mathcal{P}_\epsilon$, and takes the worst case. This quantity depends on ϵ, but we say ϵ is fixed and write $b = b(\epsilon)$. The minimax problem is then simply to minimize the maximum asymptotic bias, i.e. to find T that minimizes $b(\epsilon)$, among all equivariant estimators.

Proposition 5.1. *The solution of the minimax problem stated above is to take $T(F)$ the median of F, i.e. the estimator T_n to be the sample median of $X_1, ...X_n$, $n \geq 1$.*

Proof (from Huber 1964 (section 7) and Huber and Ronchetti 2009 Sect. 4.2). Let T be the median. Since $T(\Phi) = 0$, for a fixed $F = (1-\epsilon)\,\Phi + \epsilon\,H \in \mathcal{P}_\epsilon$, the asymptotic bias is the value x_0 such that $(1-\epsilon)\,\Phi(x_0) + \epsilon\,H(x_0) = 1/2$. Then, to attain the maximal value of b, all the contaminating mass (given by H) has to be on one side of zero, say on the right, and therefore the asymptotic maximal bias of the median is the solution x_0 of

$$(1-\epsilon)\,\Phi(x_0) = 1/2,$$

or equivalently, it is

$$x_0 = \Phi^{-1}\left(\frac{1}{2(1-\epsilon)}\right). \tag{5.1}$$

Now, define

$$f_+ = \begin{cases} (1-\epsilon)\,\varphi(x) & \text{if } x \le x_0 \\ (1-\epsilon)\,\varphi(x-2x_0) & \text{if } x > x_0 \end{cases}$$

and observe that for any $c \in \mathbb{R}$

$$f_+(x_0+c) = (1-\epsilon)\,\varphi(c-x_0) = (1-\epsilon)\,\varphi(x_0-c) = f_+(x_0-c).$$

Hence, the corresponding distribution F_+ is symmetric around x_0. We also define a translated version of F_+

$$F_-(x) = F_+(x+2x_0), \tag{5.2}$$

which is symmetric around $-x_0$. It is easy to verify that $F_+, F_- \in \mathcal{P}_\epsilon$.

Let T'_n an equivariant estimator. Then, the associated function T' is also location equivariant, in the sense that

$$T'(F(x+c)) = T'(F(x)) + c, \quad \text{for any } F \in \mathcal{P}_\epsilon,\ c \in \mathbb{R}.$$

We use this property and the relation (5.2) to compute

$$(T'(F_-) - T'(\Phi)) - (T'(F_+) - T'(\Phi)) = T'(F_-) - T'(F_+) = 2\,x_0.$$

Therefore, it is impossible to have simultaneously

$$|T'(F_+) - T'(\Phi)| < x_0 \quad \text{and} \quad |T'(F_-) - T'(\Phi)| < x_0.$$

Hence, we have either

$$\begin{gathered} |T'(F_+) - T'(\Phi)| = |T'(F_-) - T'(\Phi)| = x_0 \\ \Rightarrow T' \text{ not better than the median} \end{gathered}$$

or

$$\begin{gathered} |T'(F_+) - T'(\Phi)| > x_0 \text{ or } |T'(F_+) - T'(\Phi)| > x_0 \\ \Rightarrow T' \text{ worst than median in worst case} \end{gathered}$$

and the result is proved. □

5.2 Minimax asymptotic variance: simple case

In the previous section we have showed that the median minimizes the asymptotic bias over a neighborhood of the normal distribution. Now we direct our interest on minimizing the maximal asymptotic variance. When one has two asymptotically normal estimators T_1, T_2, the only way to chose the best performing one asymptotically is to compare their asymptotic variances. Thus, it might be useful to ensure that this measure is robust to bad observations. We start with a neighborhood similar to the one for the bias (a bit more restrictive) and another class of estimators. We follow Huber 1964 and Huber and Ronchetti 2009 (Ch.4). Assume $X_1, X_2, ...$ are i.i.d copies of X with distribution F in

$$\mathcal{P}_\epsilon = \{F = (1-\epsilon)\,\Phi + \epsilon\, H,\, H \text{ symmetric}\}, \quad 0 \leq \epsilon \leq 1 \tag{5.3}$$

i.e. our data is normal and contaminated by a symmetric distribution. The symmetry assumption on the contamination distribution can be judged to be very strong. This symmetry, as well as the equivariance assumption in the previous section, are crucial in the construction of asymptotically minimax estimators. However, as discussed in Section 4.9 of Huber and Ronchetti 2009, if one uses a symmetric model to develop a minimax estimator, the latter will be almost minimax even if H is in fact non symmetric. Also, the following asymptotic methods are able to deal with nuisance parameters, for instance when the scale is also unknown (not covered in this thesis), which is not the case of the finite sample decisions from Chapter 2-4. Again, we assume our estimators are functions of the empirical distribution. Moreover, we consider only *M-estimators*. Recall that T_n is an M-estimator if it minimizes the empirical risk, i.e. if there exists a function ρ (the *loss function*) such that

$$\frac{1}{n}\sum_{i=1}^{n} \rho(X_i - c) \text{ is minimized by } c = T_n(X_1, ..., X_n), \text{ and}$$
$$\mathbb{E}_c\, \rho(X - c) \text{ is minimized by } c = \theta, \text{ the true parameter}$$

or equivalently, if

$$\frac{1}{n}\sum_{i=1}^{n} \psi(X_i - T_n) = 0, \text{ and}$$
$$\mathbb{E}_\theta\, \psi(X - \theta) = 0$$

where $\psi = \rho'$. So, assume $T_n = T(F_n)$ is an M-estimator, with loss function ρ and $\psi = \rho'$. We know from mathematical statistics lecture that T_n is then asymptotically normal, i.e.

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, V_F(T)), \tag{5.4}$$

with asymptotic variance

$$V_F(T) = \frac{\mathbb{E}_F(\psi^2)}{(\mathbb{E}_F \psi')^2}. \tag{5.5}$$

The minimax problem consists in finding T which minimizes the maximum asymptotic variance

$$\sup_{F \in \mathcal{P}_\epsilon} V_F(T).$$

Let us set the foundations of the solution. Consider the following loss function

$$\rho(x) = \begin{cases} \frac{1}{2}\, x^2 & \text{if } |x| \leq k \\ k\,|x| - \frac{1}{2}\, k^2 & \text{if } |x| > k, \end{cases} \tag{5.6}$$

where k is a tuning parameter. It has been introduced by Huber 1964 and is often called *Huber's loss function*. Recall that the M-estimator for $\rho_1(x) = x^2$ is the minimizer of $\sum (X_i - T_n)^2$, i.e. the sample mean. And for $\rho_2(x) = |x|$, it is the sample median, which minimizes $\sum |X_i - T_n|$, and which is much more robust to outliers than the sample mean. The Huber loss function is simply the least squares (LS) for values smaller than k in absolute, and the least absolute values (LAV) for values above k in absolute. The smaller k is, the more robust will the M-estimator be. The associated ψ is

$$\psi(x) = \begin{cases} -k & \text{if } x < -k \\ x & \text{if } |x| \leq k \\ k & \text{if } x > k \end{cases} \quad = \quad \max\{-k, \min\{k, x\}\}. \tag{5.7}$$

Proposition 5.2. *Let ψ as above and T^* be such that $\sum_{i=1}^n \psi(X_i - T^*(F_n)) = 0$ for all $n \geq 1$. Then T^* solves the minimax problem, i.e. there exists $F^* \in \mathcal{P}_\epsilon$ such that*

$$V_{F^*}(T^*) = \min_T \sup_{F \in \mathcal{P}_\epsilon} V_F(T),$$

where we take the minimum over asymptotically normal estimators.

Remark. *We will characterize F^* in the proof and see later that it is in a sense least favorable, like the least favorable pair was for minimax tests.*

Remark. *Note that the assumption of H being symmetric implies that $\mathbb{E}_F \psi = 0$ so T^* is the $M-$estimator associated with Huber's loss function. Observe that ψ is exactly the same as found in* (4.16)*. Hence, this M-estimator not only minimizes the maximal asymptotic variance for ϵ-contamination by a symmetric distribution, but also yields minimax confidence intervals for finite sample and for more general indeterminacy of the form* (4.7)*. This gives support and legitimacy to the use of minimax asymptotic variance as a measure of robust optimality.*

Proof (from Huber 1964 and Huber and Ronchetti 2009 Sect. 4.3).
For $F \in \mathcal{P}_\epsilon$, we compute using (5.5)

$$V_F(T^*) = \frac{(1-\epsilon)\,\mathbb{E}_\Phi(\psi^2) + \epsilon\,\mathbb{E}_H(\psi^2)}{\left[(1-\epsilon)\,\mathbb{E}_\Phi \psi' + \epsilon\,\mathbb{E}_H \psi'\right]^2} \leq \frac{(1-\epsilon)\,\mathbb{E}_\Phi(\psi^2) + \epsilon\, k^2}{\left[(1-\epsilon)\,\mathbb{E}_\Phi \psi'\right]^2}$$

and this upper bound is attained whenever H has all its mass outside $[-k, k]$ (because then $\mathbb{E}_H(\psi^2) = k^2$ and $\mathbb{E}_H \psi' = 0$).

We want to consider T^* as a maximum likelihood estimator (MLE), and we know there is a correspondence between MLE and M-estimators. More precisely, if f is the density function of F, T is the MLE for θ if it maximizes $\prod f_\theta(X_i)$. It is the same as maximizing $\prod f_\theta(X_i)\, c'$, which is clearly equivalent to maximizing $\frac{1}{n}\sum \log(f_\theta(X_i)\, c')$, which is in turn equivalent to minimizing $\frac{1}{n}\sum -\log(f_\theta(X_i)\, c')$. This implies that T is an M-estimator, with loss function $\rho_\theta(x) = -\log(f_\theta(x)\, c')$. Equivalently, our M-estimator T^* is also a maximum likelihood estimator, for a density of the form

$$f^*(x) = c\, \exp^{-\rho(x)},$$

with ρ the Huber loss function. One can choose the tuning parameter k in (5.6) so that $c = (1-\epsilon)/\sqrt{2\pi}$. The reader can verify that statement, it suffices to choose k such that $2\,\varphi(k)/k - 2\,\Phi(-k) = \epsilon/(1-\epsilon)$, where φ is the density of the standard normal distribution. With such a k, we have

$$f^*(x) = \frac{1-\epsilon}{\sqrt{2\pi}}\, e^{-\rho(x)} = \begin{cases} \frac{1-\epsilon}{\sqrt{2\pi}} e^{-x^2/2} & \text{if } |x| \le k \\ \frac{1-\epsilon}{\sqrt{2\pi}} e^{\frac{k^2}{2} - k\,|x|} & \text{if } |x| > k. \end{cases} \tag{5.8}$$

That way, the associated distribution function F^* is of the form

$$(1-\epsilon)\,\Phi + \epsilon\, H^* \ \in \mathcal{P}_\epsilon,$$

with H^* putting all the mass outside $[-k, k]$. Hence, by the argument at the beginning of the proof:

$$V_{F^*}(T^*) = \sup_{F \in \mathcal{P}_\epsilon} V_F(T^*).$$

Moreover, we know that the MLE is asymptotically most efficient under mild conditions (see (5.10)), i.e.

$$V_{F^*}(T^*) = \min_T V_{F^*}(T)$$

and the result is proved. □

5.3 Least favorable distributions

It is interesting to highlight similarities between Proposition 5.2 and results of existence of minimax tests. In the latter, the minimax solution (ϕ, Q_0, Q_1) is composed of the least favorable pair $(Q_0, Q_1) \in \mathcal{P}_0 \times \mathcal{P}_1$, in the sense that it is the most difficult pair to test, and of ϕ which is the most efficient test (NP test) for this pair. In the minimax solution (T^*, F^*) for the asymptotic variance, $F^* \in \mathcal{P}_\epsilon$ plays the role of (Q_0, Q_1), and T^* is the most efficient estimator for this distribution, for instance the MLE. So we can legitimately hope that F^*

is in a sense a "least favorable" distribution, for which the estimators have the worst asymptotic variance. Indeed, it turns out that F^* has the smallest Fisher information among all distributions in $\mathcal{P}_\epsilon$. This is not surprising. Recall that the Fisher information $I(F)$ of a distribution F is the expectation of the squared score function, the latter being the derivative of the log-likelihood. Under some mild conditions, $I(F)$ can also be expressed as the expectation of the negative second derivative of the log-likelihood, so that it gives a feeling of how sharp the log-likelihood function is. Hence, the smaller $I(F)$, the flatter the log-likelihood function. A flat log-likelihood yields to larger variance for the MLE, because it is "more difficult" to find the maximum: there is more uncertainty.

We will now see that even for a very general uncertainty class, the MLE of the distribution minimizing the Fisher information is a candidate to solve our minimax problem, under some regularity conditions. We consider the location model, with $X_1, X_2, ...$ independent and with common distribution $F(x-\theta)$ in $\mathcal{P}$, a given class of distributions satisfying $I(F) < \infty \; \forall F \in \mathcal{P}$. The Fisher information of F can be written as

$$I(F) = \int \left(\frac{f'(x)}{f(x)} \right)^2 f(x)\, dx.$$

For an estimator $T_n = T(F_n)$, we write $V_F(T)$ its asymptotic variance under F. Let F_{LF} be the distribution minimizing the Fisher information over $\mathcal{P}$. It is unique under some mild conditions (see section 8 of Huber 1964). We call such a distribution a *least favorable distribution (LFD)*. Recall the minimax problem of finding

$$\arg\min_T \sup_{F \in \mathcal{P}} V_F(T). \tag{5.9}$$

If Θ is an open set, the support of the likelihood f_θ does not depend on θ, and under smoothness assumptions on f_θ, the best (i.e. smallest) asymptotic variance for a "regular" estimator T is

$$\min_T V_F(T) = \frac{1}{I(F)}. \tag{5.10}$$

Let $T_n = T(F_n)$ be a regular estimator attaining this lower bound for all θ, we call it *asymptotically efficient*. The regularity assumption insures that we avoid estimators that have zero variance for one fixed θ_0. For details about this assumption, the reader can look at Bickel et al. 1998. Under the conditions above, the MLE is asymptotically efficient. We have by assumption

$$\sup_{F \in \mathcal{P}} \frac{1}{I(F)} = \frac{1}{I(F_{LF})}. \tag{5.11}$$

So, if T also satisfies

$$V_F(T) \leq \frac{1}{I(F_{LF})} \quad \forall F \in \mathcal{P}, \tag{5.12}$$

then it follows from (5.10) and (5.11) that T solves the problem (5.9).

We now wish to derive a necessary and sufficient condition for a distribution to be least favorable, so that it will be quite easy to check it for particular distributions.

Proposition 5.3. *$F \in \mathcal{P}$ minimizes the Fisher information iff*

$$-4 \int \frac{(\sqrt{f})''}{\sqrt{f}}(g-f) \geq 0 \quad \forall\, G \in \mathcal{P} \text{ with } I(G) < \infty, \tag{5.13}$$

where f, g are the density functions of F and G respectively.

Proof (from Huber and Ronchetti 2009 Sect. 4.5).
We know that $F \mapsto I(F)$ is a convex function, so F minimizes $I(\cdot)$ iff for $F_t = (1-t)\,F + t\,G$, one has

$$\frac{d}{dt} I(F_t)\bigg|_{t=0} \geq 0 \quad \forall\, G \in \mathcal{P} \text{ with } I(G) < \infty. \tag{5.14}$$

We have

$$I(F_t) = \int \left(\frac{f_t'}{f_t}\right)^2 f_t = \int \frac{(f_t')^2}{f_t}$$

and we compute, for any $G \in \mathcal{P}$

$$\begin{aligned}\frac{d}{dt} I(F_t)\bigg|_{t=0} &= \frac{d}{dt} \int \left[\frac{\left((1-t)\,f' + t\,g'\right)^2}{(1-t)\,f + t\,g}\right]\bigg|_{t=0} \\ &= \int \left[2\frac{f'}{f}\,(g'-f') - \left(\frac{f'}{f}\right)^2 (g-f)\right],\end{aligned}$$

where we used the monotone convergence theorem to switch derivative and integral. We recognize f'/f the score function of f. Recall that the MLE of f is an M-estimator, with loss function $\rho = -\log f$. Taking the derivatives we have

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

so we can write

$$\begin{aligned}&\frac{d}{dt} I(F_t)\bigg|_{t=0} = -\int \left[2\psi\,(g'-f') - \psi^2(g-f)\right] \\ &= \left[-2(g-f)\psi\right]_{-\infty}^{+\infty} + \int 2\psi'(g-f) - \int \psi^2(g-f) = \int (2\psi' - \psi^2)(g-f).\end{aligned}$$

Thus,

$$\frac{d}{dt} I(F_t)\bigg|_{t=0} \geq 0 \;\Leftrightarrow$$
$$\int (2\psi' - \psi^2)(g-f) \geq 0$$

and it is easy to verify that the latter is equivalent to

$$-4 \int \frac{(\sqrt{f})''}{\sqrt{f}}(g-f) \geq 0$$

□

One can use condition (5.13) to show Proposition 5.2, with the least favorable density in (5.8). We can also use this condition to check that the density

$$f(x) = \begin{cases} (1-\epsilon)\, g(x_0)\, e^{k(x-x_0)} & \text{if } x \leq x_0 \\ (1-\epsilon)\, g(x) & \text{if } x_0 < x < x_1 \\ (1-\epsilon)\, g(x_1)\, e^{-k(x-x_1)} & \text{if } x \geq x_1, \end{cases}$$

is least favorable for the uncertainty class $\mathcal{P}_\epsilon = \{F = (1-\epsilon)\, G + \epsilon\, H\}$, where $[x_0, x_1]$ is the interval where $|g'/g|$ is below the tuning parameter k, for g the density of G. So we know the LFD for the contamination of any distribution G. Let us now look at another type of uncertainty class.

5.4 Kolmogorov neighborhoods of Gaussian distribution

We assume that the distribution of the observations is at distance at most ϵ of the normal distribution, where the distance considered is the Kolmogorov distance (distance generated by $||\cdot||_\infty$ for distribution functions). Precisely, fix $0 \leq \epsilon \leq 1$ and consider

$$\mathcal{P} = \{F : \sup |F(x) - \Phi(x)| \leq \epsilon\}. \tag{5.15}$$

We follow Huber 1964 and Huber and Ronchetti 2009 (Chap. 4) and we try for the least favorable density a symmetric function of the form

$$f^*(x) = f^*(-x) = \begin{cases} \frac{\varphi(x_0)}{\cos^2(\omega\, x_0/2)} \cos^2\left(\frac{\omega\, x}{2}\right) & \text{if } 0 \leq x < x_0 \\ \varphi(x) & \text{if } x_0 \leq x \leq x_1 \\ \varphi(x_1)\, e^{-\lambda(x-x_1)} & \text{if } x > x_1, \end{cases} \tag{5.16}$$

with $\omega, \lambda \in \mathbb{R}$, $0 < x_0 < x_1$ and φ the standard normal density.

Proposition 5.4. *For $\epsilon < \epsilon_0 \approx 0.0303$, f^* is least favorable for the asymptotic variance over $\mathcal{P}$ defined in* (5.15).

The above implies that the MLE for (5.16) solves the minimax problem (5.9) if it satisfies (5.12).

Proof (from Huber 1964 and Huber and Ronchetti 2009 Sect. 4.5).
Let us derive four equations in order to characterize the variable $x_0, x_1, \omega, \lambda$. First we require

$$\psi(x) = -(\log f^*(x))' = \begin{cases} \omega \tan\left(\frac{\omega\, x}{2}\right) & \text{if } 0 \leq x < x_0 \\ x & \text{if } x_0 \leq x \leq x_1 \\ \lambda & \text{if } x > x_1, \end{cases}$$

to be continuous. This requires

$$\omega \tan\left(\frac{\omega\, x_0}{2}\right) = x_0, \tag{5.17}$$

$$\lambda = x_1. \tag{5.18}$$

We want the LFD to be at the boundary of $\mathcal{P}$ between x_0 and x_1, i.e.

$$F^*(x) = \Phi(x) - \epsilon \quad \text{for } x_0 \le x \le x_1. \tag{5.19}$$

To ensure this as well as f^* integrating to 1, we need

$$\int_0^{x_0} f^* = \int_0^{x_0} \varphi \ - \epsilon \tag{5.20}$$

$$\int_{x_1}^{\infty} f^* = \int_{x_1}^{\infty} \varphi \ + \epsilon. \tag{5.21}$$

Indeed, we then have for $x \in [x_0, x_1]$

$$\begin{aligned} F^*(x) = \int_{-\infty}^{x} f^* = \frac{1}{2} + \int_0^{x_0} f^* + \int_{x_0}^{x} f^* = \frac{1}{2} + \int_0^{x_0} f^* + \frac{1}{2} - \int_0^{x_0} f^* - \int_x^{\infty} f^* \\ = 1 - \int_x^{x_1} f^* - \int_{x_1}^{\infty} \varphi - \epsilon = 1 - \int_x^{\infty} \varphi - \epsilon = \Phi(x) - \epsilon \end{aligned}$$

and

$$\int_0^{\infty} f^* = \int_0^{x_0} \varphi - \epsilon + \int_{x_0}^{x_1} \varphi + \int_{x_1}^{\infty} \varphi + \epsilon = \frac{1}{2}$$

So (5.17)-(5.21) determine $x_0, x_1, \omega, \lambda$ when ϵ is fixed. For simplicity, we set $u \coloneqq \omega\, x_0$ and express everything in terms of u instead of ϵ, which yields the four following equations

$$\begin{aligned} x_0 &= \sqrt{u \, \tan \frac{u}{2}} \\ \omega &= \frac{u}{x_0} \\ \epsilon &= \Phi(x_0) - \frac{1}{2} - x_0\, \varphi(x_0) \frac{1 + \sin u/u}{1 + \cos u} \\ \epsilon &= \frac{\varphi(x_1)}{x_1} - \Phi(-x_1) \end{aligned}$$

Here we have to be careful because to ensure $x_0 < x_1$ we can compute numerically that we need $\epsilon < \epsilon_0 \approx 0.0303$. That is the reason of this assumption in

the statement.

Before proving that F^* is a LFD, we have to check that it belongs to $\mathcal{P}$. The fact that $f^* = \phi$ on $[x_0, x_1]$ combined with (5.20) and (5.21) implies

$$f^* \leq \varphi \quad \text{on } [0, x_0] \tag{5.22}$$
$$f^* \geq \varphi \quad \text{on } [x_1, +\infty). \tag{5.23}$$

And now, (5.19) combined with (5.22) implies that $|F - \Phi| \leq \epsilon$ on $(-\infty, x_0)$. Combining (5.23), (5.21) and (5.19) yields $|F - \Phi| \leq \epsilon$ on $(x_1, +\infty)$ so indeed $F^* \in \mathcal{P}$.

Finally, we check for condition (5.13). Note that it suffices to have (5.14) for symmetric distributions, because for $G \in \mathcal{P}$, the symmetrized distribution $\tilde{G}(x) = \frac{1}{2}[G(x) + 1 - G(-x)]$ has smaller information since $I(\cdot)$ is convex. Hence it suffices to verify (5.13) for symmetric distributions. So let G symmetric with density g. We compute

$$-4\frac{(\sqrt{f^*})''}{\sqrt{f^*}} = \begin{cases} \omega^2 & \text{if } 0 \leq x < x_0 \\ 2 - x^2 & \text{if } x_0 \leq x \leq x_1 \\ -x_1^2 & \text{if } x > x_1. \end{cases} \tag{5.24}$$

It is interesting to observe that it is constant outside $[x_0, x_1]$. Set $F_d = G - F^*$ and compute

$$\begin{aligned} -4\int \frac{(\sqrt{f^*})''}{\sqrt{f^*}}(g - f^*) &= \int_0^{x_0} \omega^2 \, dF_d + \int_{x_0}^{x_1} (2 - x^2) \, dF_d - \int_{x_1}^{+\infty} x_1^2 \, dF_d \\ &= \omega^2 F_d(x_0) + 2F_d(x_1) - 2F_d(x_0) - x_1^2 F_d(+\infty) + x_1^2 F_d(x_1) - \int_{x_0}^{x_1} x^2 \, dF_d \\ &= \omega^2 F_d(x_0) + 2F_d(x_1) - 2F_d(x_0) + x_1^2 F_d(x_1) - \left[x_1^2 F_d(x_1) - x_0^2 F_d(x_0)\right] \\ &\quad + 2\int_{x_0}^{x_1} x F_d(x) \, dx = (\omega^2 + x_0^2 - 2) F_d(x_0) + 2F_d(x_1) + 2\int_{x_0}^{x_1} x F_d(x) \, dx, \end{aligned}$$

where we integrated by part and used $F_d(+\infty) = 0$ to get the second equality. Yet,

$$\omega^2 + x_0^2 - 2 = \frac{u}{\tan\frac{u}{2}} + u\tan\left(\frac{u}{2}\right) - 2 = 2\left(\frac{u}{\sin u} - 1\right) \geq 0$$

because $\sin u \leq u$, and

$$F_d(x) \geq 0 \text{ for any } x_0 \leq x \leq x_1$$

so that

$$-4\int \frac{(\sqrt{f^*})''}{\sqrt{f^*}}(g - f^*) \geq 0$$

and condition (5.13) is satisfied. □

For the case $\epsilon > \epsilon_0$, Sacks and Ylvisaker 1972 showed that the least favorable density is of the form

$$f(x) = \begin{cases} \frac{\beta}{A\,(1+\beta)} \cos^2(\alpha\, x) & \text{if } |x| \leq 1 \\ \frac{\beta}{A\,(1+\beta)} \cos^2(\alpha)\, e^{2\beta}\, e^{-2\beta\,|x|} & \text{if } |x| > 1, \end{cases}$$

with A, α constants and $\beta = \alpha \tan \alpha$.

5.5 Minimax L-estimators

In this section we build new estimators that are minimax for the asymptotic variance. We look for most efficient estimators for the LFD $F^* \in \mathcal{P}$, i.e. estimators T_n with $\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, V_F(T_n))$ under $F \in \mathcal{P}$, for which

$$V_{F^*}(T) \leq \frac{1}{I(F^*)}.$$

We know that the MLE of F^* does the job under some conditions, but we want other types of estimators. We consider estimators that are a linear combination of the order statistics. Precisely, for the sample $X_1, X_2, ..., X_n$, we write the order statistics $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ and consider estimators of the form

$$T_n(X_1, ..., X_n) = \sum_{i=1}^{n} b_i\, X_{(i)}, \quad b_i \in \mathbb{R},\, i = 1, ..., n.$$

These estimators are often called *L-estimators*. We rewrite an L-estimator T_n in the equivalent form

$$T_n(X_1, ..., X_n) = \frac{1}{n} \sum_{i=1}^{n} w\left(\frac{i}{n+1}\right) X_{(i)},$$

with $w : [0, 1] \to \mathbb{R}$ a weighting function.

Let us first suppose that there is no uncertainty on the distribution $F(x-\theta)$ of the sample. For the aim of our following developments, we consider $U_1, ...U_n$ i.i.d. random variables following a uniform distribution $\mathcal{U}[0, 1]$. Then, it is known that

$$X_i = G(U_i),\ i = 1, ...n$$

if $G = F^{-1}$ is the generalized inverse of F, that can be defined as

$$G(u) = \inf\{x : F(x) \geq u\},\ u \in \mathbb{R}.$$

So T_n becomes

$$T_n(X_1, ..., X_n) = \frac{1}{n} \sum_{i=1}^{n} w\left(\frac{i}{n+1}\right) G(U_{(i)}). \tag{5.25}$$

It has been shown by Huber 1969 (Chap.3) that T_n is then asymptotically normal ($\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, V_F(T_n))$), with asymptotic variance

$$V_F(T_n) = \int A^2(t)\, dt - \left[\int A(t)\, dt\right]^2, \tag{5.26}$$

where

$$dA(t) = w(t)\, dG(t), \tag{5.27}$$

i.e. $A(E) = \int_E w\, dG + C$ for any Borel set E. Now we want to find w (which determines T_n) such that $V_F(T_n)$ is as small as possible. We look for w such that

(i) $\int_0^1 w(t)\, dt = 1$ (total weight is 1)

(ii) $\int A(t)\, dt = 0$

(iii) $\int A^2(t)\, dt$ is minimal.

We follow Huber 1969 (Chap.3). Note that $(ii) - (iii)$ are equivalent to minimizing $V_F(T_n)$, when we choose the integrating constant C in the definition of $A(t)$ such that $\int A(t)\, dt = 0$.

First, we rewrite w:

$$w(t)\left(= \frac{\text{“}dA(t)\text{”}}{dG(t)}\right) = \frac{A'(t)}{G'(t)}, \tag{5.28}$$

with

$$G'(t) = (F^{-1})'(t) = \frac{1}{F'(F^{-1}(t))} = \frac{1}{f(F^{-1}(t))}.$$

Then,

$$\begin{aligned}\int_0^1 w(t)\, dt = \int_0^1 A'(t)\, f(F^{-1}(t))\, dt &= \left[A(t)\, f(F^{-1}(t))\right]_0^1 - \int_0^1 A(t)\frac{f'(F^{-1}(t))}{f(F^{-1}(t))}\, dt \\ &= -\int_0^1 A(t)\frac{f'(F^{-1}(t))}{f(F^{-1}(t))}\, dt = \int_0^1 A(t)\, \psi(F^{-1}(t))\, dt,\end{aligned}$$

where we used that f is zero at $F^{-1}(1) = \infty$ and $F^{-1}(0) = -\infty$ and we set $\psi = -\frac{f'}{f}$ as usual. Hence, the first condition is equivalent to

$$(i)' \quad \int_0^1 A(t)\, \psi(F^{-1}(t))\, dt = 1$$

Now, we use Cauchy-Schwarz inequality to get

$$\int A^2(t)\,dt \int [\psi(F^{-1}(t))]^2\,dt \geq \left[\int_0^1 A(t)\,\psi(F^{-1}(t))\,dt\right]^2 \overset{(i)'}{=} 1$$

$$\Rightarrow \int A^2(t)\,dt \geq \frac{1}{\int_0^1 [\psi(F^{-1}(t))]^2\,dt} = \frac{1}{\int_{-\infty}^{+\infty} \psi^2(u)\,f(u)\,du} = \frac{1}{\mathbb{E}\,\psi^2} = \frac{1}{I(F)},$$

where we use the change of variables $u = F^{-1}(t)$. We already knew this lower bound, but Cauchy-Schwarz also tells that we have equality (and hence attain the minimal asymptotic variance) if and only if

$$A(t) = a\,\psi(F^{-1}(t)),\ a \in \mathbb{R}. \tag{5.29}$$

Then, for $A(t)$ as in (5.29):

$$\int_0^1 A(t)\,dt = a\int_0^1 \psi(F^{-1}(t))\,dt = a\int_{-\infty}^{+\infty} \psi(u)\,f(u)\,du = -a\int_{-\infty}^{+\infty} f'(u)\,du = 0$$

so condition (ii) is satisfied. Also,

$$\int_0^1 A(t)\,\psi(F^{-1}(t))\,dt = a\int_0^1 [\psi(F^{-1}(t))]^2\,dt = \frac{1}{a}\int A^2(t)\,dt = \frac{1}{a}\,\frac{1}{I(F)}$$

so in order to satisfy condition (i) we must have

$$a = \frac{1}{I(F)}.$$

Thus,

$$A(t) = \frac{1}{I(F)}\,\psi(F^{-1}(t))$$

minimizes the asymptotic variance, and we get the w we were looking for using (5.28):

$$w(t) = \frac{A'(t)}{G'(t)} = A'(t)\,f(F^{-1}(t)) = \frac{1}{I(F)}\,\frac{\psi'(F^{-1}(t))}{f(F^{-1}(t))}\,f(F^{-1}(t)).$$

So, using the weight function

$$w(t) = \frac{1}{I(F)}\,\psi'(F^{-1}(t)) \tag{5.30}$$

yields T_n minimizing $V_F(\cdot)$.

This is nice, but if we only know that the distribution is in a certain set $\mathcal{P}$, we might want our L-estimator to have minimal asymptotic variance over $\mathcal{P}$, i.e. find

$$\arg\min_{T} \sup_{F \in \mathcal{P}_\epsilon} V_F(T)$$

as in Section 5.2. The idea is to use equation (5.30) for the least favorable distribution of $\mathcal{P}$. Consider the normal symmetric contamination case, in which we know that, for fixed $0 \leq \epsilon \leq 1$, F lies in

$$\mathcal{P} = \{F = (1-\epsilon)\,\Phi + \epsilon\, H,\ H \text{ symmetric}\}\,.$$

It is exactly the same class we treated in Section 5.2. Hence we already know the LFD density f_* (see (5.8)) and

$$\psi(x) = -\frac{f_*'(x)}{f_*(x)} = \max\{-k, \min\{k, x\}\}$$

as in (5.7). Thus,

$$\psi'(x) = \begin{cases} 1 & \text{if } |x| < k \\ 0 & \text{if } |x| > k, \end{cases}$$

so

$$\psi'(F_*^{-1}(t)) = \begin{cases} 1 & \text{if } |F_*^{-1}(t)| < k \\ 0 & \text{if } |F_*^{-1}(t)| > k, \end{cases} = \begin{cases} 1 & \text{if } \alpha < t < 1-\alpha \\ 0 & \text{else,} \end{cases} \tag{5.31}$$

where $\alpha = F_*(-k)$. We also compute

$$\frac{1}{I(F_*)} = \frac{1}{-\mathbb{E}\,\psi'} = \frac{1}{-P(-k < x < k)} = \frac{1}{1-2\alpha}. \tag{5.32}$$

Combining (5.31) and (5.32) in (5.30) yields

$$w(t) = \frac{1}{1-2\alpha}\,\mathbb{1}\{\alpha < t < 1-\alpha\} \tag{5.33}$$

so the estimator minimizing $V_{F_*}(\cdot)$ is

$$T_n(X_1, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1-2\alpha}\,\mathbb{1}\{\alpha < t < 1-\alpha\}\ X_{(i)}$$

which simplifies to

$$T_n(X_1, ..., X_n) = \frac{1}{n-2n\alpha}\left(X_{([n\alpha+1])} + ... + X_{([n-n\alpha])}\right). \tag{5.34}$$

T_n is a trimmed mean: we do not take into account the first $[n\alpha]$ and last $[n\alpha]$ terms. For T_n to be minimax, we still need

Proposition 5.5.

$$V_F(T_n) \leq \frac{1}{I(F_*)} \ \forall\, F \in \mathcal{P}, \tag{5.35}$$

for T_n the trimmed mean as in (5.34)*.*

Remark. *Once the result is proved, we know that T_n is minimax (see* (5.12)*).*

Remark. *It might appear that the result is trivial since F_* is LFD, but this only implies*

$$V_{F_*}(T_n) \leq \frac{1}{I(F_*)} \leq \frac{1}{I(F)} \ \forall F \in \mathcal{P},$$

which is not sufficient because we want to bound the asymptotic variance under each $F \in \mathcal{P}$.

Proof (from Huber 1969 Chap. 3).
Let $F \in \mathcal{P}$, $F \neq F_*$. We know

$$V_F(T_n) = \int A_F^2(t)\, dt - \left[\int A_F(t)\, dt\right]^2,$$

where A_F is defined like in (5.27), with $G = F^{-1}$ the generalized inverse of F. We can choose the integrating constant C to be such that $A_F(1/2) = 0$, so we have for each $t \in [0,1]$

$$A_F(t) = \int_{1/2}^{1} w(s)\, dG(s).$$

F is by assumption a symmetric distribution, so the graph of F is symmetric with respect to the point $(0, 1/2)$. Hence, the graph of G is symmetric w.r.t. the point $(1/2, 0)$. In particular, $G(1/2) = 0$ and hence, recalling the expression for w (see (5.33)), we have

$$A_F(t) = \begin{cases} \frac{G(\alpha)}{1-2\alpha} & \text{if } t \leq \alpha \\ \frac{G(t)}{1-2\alpha} & \text{if } \alpha < t < 1-\alpha \\ \frac{G(1-\alpha)}{1-2\alpha} & \text{if } t \geq 1-\alpha \end{cases}. \tag{5.36}$$

Also, since G is symmetric w.r.t. $(1/2, 0)$, $\int_0^1 A_F(t)\, dt = 0$, so the asymptotic variance reduces to

$$V_F(T_n) = \int A_F^2(t)\, dt.$$

Recall that for $x \in [-k, k]$, $F_*(x) = (1-\epsilon)\,\Phi(x)$ (see proof of Prop. 5.2). Hence, $F_*(x) \leq F(x)$ for all $F \in \mathcal{P}$ on $[-k, k]$, or equivalently

$$G(t) \leq G_*(t) \ \ \forall\, t \in [\alpha, 1-\alpha]$$

which implies by (5.36) that

$$|A_F(t)| \leq |A_{F_*}(t)| \quad \forall\, t \in [0,1].$$

Thus, finally

$$V_F(T_n) \leq V_{F_*}(T_n).$$

□

A legit question is now if one can find a minimax L-estimator for any uncertainty class $\mathcal{P}$. Sacks and Ylvisaker 1972 showed that the answer is no. Indeed, they showed that $\mathcal{P} = \{F : \sup |F(x) - \Phi(x)| \leq \epsilon\}$ with $\epsilon > \epsilon_0 \approx 0.0303$ is a counterexample. For $T_n = 1/n \sum w(i/(n+1))\, X_{(i)}$ with w like in (5.30) and F^* the LFD of $\mathcal{P}$, there exists a distribution $F \in \mathcal{P}$ with

$$V_F(T_n) > \frac{1}{I(F^*)}.$$

This proves that there is no minimax L-estimator for $\mathcal{P}$. Indeed, T_n was the unique candidate because it is the unique L-estimator with $V_{F^*}(T_n) \leq 1/I(F^*)$ and any minimax estimator must verify the latter inequality, otherwise it would be beaten by the MLE of F^*.

References

Bickel, Peter J. et al. *Efficient and Adaptive Estimation for Semiparametric Models*. 1st ed. Springer-Verlag New York, 1998. ISBN: 978-0-387-98473-5.

Huber, Peter J. "A Robust Version of the Probability Ratio Test". In: *Annals of Mathematical Statistics* 36.6 (1965), pp. 1753–1758.

— "Robust confidence limits". In: *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 10 (1968), pp. 269–278.

— "Robust Estimation of a Location Parameter". In: *Annals of Mathematical Statistics* 35.1 (1964), pp. 73–101.

— *Théorie de l'inference statistique robuste*. Vol. 31. Séminaire de mathématiques supérieures. Les Presses del'Université de Montreal, 1969.

Huber, Peter J. and Elvezio M. Ronchetti. *Robust statistics*. 2nd ed. John Wiley & Sons, Inc., 2009. ISBN: 978-0-470-12990-6.

Huber, Peter J. and Volker Strassen. "Minimax Tests and the Neyman-Pearson Lemma for Capacities". In: *The Annals of Statistics* 1.2 (1973), pp. 251–263.

Levy, Bernard C. "Robust Hypothesis Testing With a Relative Entropy Tolerance". In: *IEEE Transactions on Information Theory* 55.1 (2009), pp. 413–421.

Sacks, Jerome and Donald Ylvisaker. "A Note on Huber's Robust Estimation of a Location Parameter". In: *Annals of Mathematical Statistics* 43.4 (1972), pp. 1068–1075.

Song, Enbin et al. "Robust hypothesis testing for asymmetric nominal densities under a relative entropy tolerance". In: *SCIENCE CHINA Mathematics* 61.10 (2018), pp. 1851–1880.

Tukey, John W. "A survey of sampling from contaminated distributions". In: *Contributions to Probability and Statistics Essays in Honor of Harold Hotelling* (1960), pp. 448–485.

6 Appendix

Proof of Lemma 2.2 (from Huber and Strassen 1973).
We show that for any $t \in [0, \infty]$, one can choose $A_t = \bigcup_{s>t} A_s$ as minimizing w_t. Let $t < s$. Property (2.2) of v_0 implies that $w_t - w_s = (t-s)v_0$ is a decreasing function. In particular, for any $A, B \in \mathcal{F}$

$$w_t(A \cup B) - w_s(A \cup B) \leq w_t(A) - w_s(A)\,.$$

Since w_s is a 2-alternating capacity, we also have

$$w_s(A \cup B) + w_s(A \cap B) \leq w_s(A) + w_s(B)\,.$$

Adding the last two inequalities yields

$$w_t(A \cup B) + w_s(A \cap B) \leq w_t(A) + w_s(B)\,.$$

In particular, we can instead of A and B choose sets A_t^* and A_s^*, minimizing w_t and w_s respectively:

$$w_t(A_t^* \cup A_s^*) + w_s(A_t^* \cap A_s^*) \leq w_t(A_t^*) + w_s(A_s^*)$$

and obtain, since A_t^* and A_s^*, minimize w_t and w_s

$$w_t(A_t^* \cup A_s^*) = w_t(A_t^*) \tag{6.1}$$
$$w_s(A_t^* \cap A_s^*) = w_s(A_s^*) \tag{6.2}$$

We have a closer look at (6.1), which says that $A_t^* \cup A_s^*$ minimizes w_t. Thus, we can replace A_t^* by $A_t^* \cup A_s^*$. Also, since, s was arbitrary, with $s > t$, we can take any $q > t$ and replace A_s^* by A_q^*. Repeating the previous steps, we get that $A_t^* \cup A_s^* \cup A_q^*$ minimizes w_t. Iterating the procedure, we get that for $t < t_1 < t_2 < ... < t_K$, $K \geq 1$, w_t is minimized by $A_t^* \bigcup \left(\bigcup_{i=1}^K A_{t_i}^* \right)$. We use the same property that yielded (2.7) to extend this result for $K \to \infty$. Let $(t_n)_{n \geq 1}$ be a dense sequence in $(0, \infty)$, then

$$\tilde{A}_{t_n} \coloneqq \bigcup_{t_m \geq t_n} A_{t_m}^*$$

minimizes w_{t_n}. Note that, proceeding the same way with (6.2), we find that $\tilde{A}_{t_n} \coloneqq \bigcap_{t_m \leq t_n} A_{t_m}^*$ also minimizes w_t, but we let this case to the reader. Now we still have to extend the last statement for all $t \in [0, \infty]$.

Case $t < \infty$: we use that $|w_{s_1}(A) - w_{s_2}(A)| \leq |s_1 - s_2|$. Since for any $t \in [0, \infty)$ we can find n such that t_n and t are arbitrarily close, we can make w_{t_n} and w_t to be arbitrarily close. Hence, we can pick $\bigcup_{t_n > t} \tilde{A}_{t_n}$, or equivalently $A_t = \bigcup_{s>t} A_s$ to minimize w_t.

Case $t = \infty$: for any $t < \infty$ one has

$$\begin{aligned} & w_t(A_t) \le w_t(\varnothing) = 0 \\ \Leftrightarrow\ & t\, v_0(A_t) - u_1(A_t) \le 0 \\ \Leftrightarrow\ & t\, v_0(A_t) \le u_1(A_t) \le 1 \\ \Leftrightarrow\ & v_0(A_t) \le \frac{1}{t} \end{aligned}$$

Hence, since we found $A_t = \bigcap_{s<t} A_s$. Letting $t \to \infty$ yields

$$v_0\left(\bigcap_{t<\infty} A_t\right) = 0, \tag{6.3}$$

so we can choose $A_\infty = \bigcap_{t<\infty} A_t$ since $w_\infty = v_0$. □

Proof of Lemma 2.3 (from Huber and Strassen 1973).
It suffices to show that for a decreasing sequence $(f_n) \subset C^b(\Omega)$ with $f_n \to 0$ we have $\tilde{P}(f_n) \to 0$ for each $\tilde{P} \le \tilde{v}$. So let such a sequence (f_n). We have $v\{f_n \ge t\} \to 0$ for every $t > 0$ by assumption 4. of capacities. Thus, $\tilde{P}(f_n) \le \tilde{v}(f_n) = \int_0^\infty v\{f_n \ge t\}\, dt \to 0$ where we used the monotone convergence theorem. □

Proof of Lemma 2.4 (from Huber and Strassen 1973).
Without loss of generality, we can assume $h > 0$ (otherwise replace h by $2 + \arctan h$). There is a decreasing sequence $(g_n) \subset C^b(\Omega)$ with $g_n \to h$. Then $\tilde{v}(g_n) \to \tilde{v}(h)$ by assumption 4. of capacities. $U = \{f \in C(\Omega) : \tilde{v}(|f|) < \tilde{v}(h)\}$ and $V = \{g \in C(\Omega) : g \ge h\}$ are both convex sets, so by the separation theorem (Hahn-Banach) there is a linear function Q such that $\tilde{Q}(f) < \tilde{Q}(g)$ for any $f \in U, g \in V$. We can normalize $\tilde{Q}$ such that $\inf\{\tilde{Q}(f) : g \in V\} = \tilde{v}(h)$. Then $\tilde{v}(|f|) < \tilde{v}(h)$ implies $\tilde{Q}(f) \le \tilde{v}(h)$, so we must have $\tilde{Q}(f) \le \tilde{v}(|f|)$. It follows that $\tilde{Q}$ is induced by a measure $Q \le v$ (like in (2.14)). Then

$$\tilde{Q}(h) = \int_0^\infty Qh > t\, dt = \int_0^\infty vh > t\, dt$$

so that

$$Q\{h > t\} = v\{h > t\} \text{ for almost all } t.$$

It actually holds for all t because of assumption 3. of capacities (build a sequence (t_n) which satisfies it and converges to t). In particular, $Q\{h > 0\} = Q(\Omega) = v(\Omega) = 1$ so Q is a probability measure. For an increasing sequence (t_n), we have

$$v\{h \ge t\} \le \lim_{t_n \to t} v\{h > t_n\} = \lim_{t_n \to t} Q\{h > t_n\} = Q\{h \ge t\}$$

and the second equality is also proved. □

Proof of Lemma 2.5 (from Huber and Strassen 1973).
Let $A \in \mathcal{F}$ and $F_n \subset A$ an increasing sequence of closed sets with $v(F_n) = v(A)$. Define an upper semi-continuous function h such that $F_n = \{h \geq t_n\}$. Then, by Lemma 2.4, we can choose $Q \leq v$ with $Q(F_n) = v(F_n)$. Then

$$Q(A) \geq \lim Q(F_n) = \lim v(F_n) = v(A).$$

□