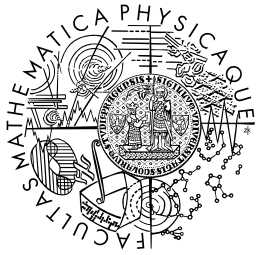


FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Filip Bočinec

Halfspace depth for location and scatter: robustness and minimax optimality

Department of Probability and Mathematical Statistics

Supervisor of the master thesis: doc. Mgr. Stanislav Nagy, Ph.D.

Study programme: Probability, Mathematical Statistics and Econometrics

Prague 2024

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

I want to express my deepest gratitude to my supervisor, doc. Mgr. Stanislav Nagy, Ph.D., for his invaluable mentoring and guidance. His expertise, patience, and encouragement have been instrumental in completing this thesis. I also extend my heartfelt thanks to my family and friends for their unwavering support and encouragement. Their love and belief in me have been a constant source of strength and motivation.

Title: Halfspace depth for location and scatter: robustness and minimax optimality

Author: Filip Bočinec

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. Mgr. Stanislav Nagy, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This thesis explores the concepts of location and scatter halfspace depth. Location halfspace depth is a well-established tool in nonparametric statistics, while scatter halfspace depth represents a newer concept that is currently undergoing active research. The primary goal of this work is to present the fundamental properties of halfspace depth for both location and scatter, with a special emphasis on the robustness of the corresponding medians. A significant portion of the thesis is dedicated to examining the minimax optimality of the location and scatter halfspace median. It provides a detailed framework concerning the rates of convergence and minimax optimal estimators. By employing this framework, the thesis demonstrates that both the location halfspace median and the scatter halfspace median achieve minimax optimality within Huber's contamination model. This finding underscores both the robustness and the rate optimality of these estimators.

Keywords: multivariate analysis, halfspace depth, scatter halfspace depth, robustness, minimax optimality

Contents

Introduction

Since the pioneering work of Tukey [1975], the concept of statistical depth has become an invaluable tool in multivariate and nonparametric statistics. Statistical depth functions provide robust methods for understanding and summarizing multivariate data, making them essential in modern data analysis. Notable applications include outlier detection [Serfling, 2006], depth-based classification and clustering [Jörnsten, 2004], and rank and sign tests [Hettmansperger and Oja, 1994].

The classical location halfspace depth, rooted in the notion of centrality, measures how central a point is with respect to a distribution. It generalizes the univariate concept of the median to higher dimensions, offering a robust location functional. This is particularly important in multivariate analysis, where outliers or skewed distributions can heavily influence traditional measures such as the mean.

Scatter halfspace depth, on the other hand, extends the idea of data depth to the scatter of the data. Traditional measures of scatter, such as covariance matrices, assume specific distributional forms and are sensitive to outliers, which can distort the analysis. Scatter halfspace depth offers a nonparametric alternative that does not rely on these assumptions. It measures the "centrality" of scatter matrices within the space of all positive definite matrices with respect to a distribution. This thesis explores both location halfspace depth and scatter halfspace depth, delving into their theoretical foundations. It is divided into three comprehensive chapters, each addressing different aspects of these concepts.

In the first chapter, we introduce location halfspace depth and discuss its main properties in detail. We further examine its robust characteristics, such as the influence function, gross error sensitivity, and breakdown point. Its robust properties make location halfspace depth a powerful tool in multivariate analysis.

In the second chapter, we turn our attention to scatter halfspace depth. We introduce its definition and discuss its essential properties, drawing analogies to location halfspace depth where appropriate. Notably, because the third chapter forms the main part of the thesis, several theorems in the first and second chapters only reference proofs from the literature to maintain a manageable scope for this work.

The third chapter constitutes the core of this thesis. We begin by introducing key concepts of minimax estimation theory, such as convergence rate and minimax optimal convergence rate. This framework is then used in Sections 3.4 and 3.5, where we demonstrate in detail that the location halfspace median and scatter halfspace median are minimax optimal estimators in Huber's contamination model. Huber's contamination model is a robust statistical framework that accounts for data contamination, making it highly relevant for real-world applications where data imperfections are common.

These findings indicate that the above-mentioned estimators are not only robust but also achieve an optimal convergence rate under contamination. This means that they provide relatively reliable estimates even when the data contains a significant amount of noise or outliers. These results have noteworthy implications for various applications, including financial data analysis, biomedi-

cal research, and any field that relies on robust multivariate statistical methods.

In summary, this thesis provides a thorough examination of location and scatter halfspace depths, highlighting their theoretical properties. It underscores the importance of statistical depth functions in modern data analysis and their potential to provide more accurate and insightful results in the presence of complex, multivariate data.

Notation

Symbol	Description
R^d	Euclidean space of dimension d
GL_d	Set of all non-singular $d \times d$ real matrices
PD_d	Set of all symmetric positive definite $d \times d$ real matrices
$\mathbf{I}_d$	$d \times d$ identity matrix
$x, \mathbf{A}$	Vector x, matrix $\mathbf{A}$ (bold font)
(a, b), $[a, b]$	Open interval, closed interval
,	Subset, proper subset
$\cdot, \cdot$, $\cdot$	Standard inner product, Euclidean norm
$\mathsf{int}(\cdot)$, $\mathsf{cl}(\cdot)$, $\mathsf{bd}(\cdot)$	Interior, closure and boundary
$B_d(x, r)$	Closed ball in R^d with center x and radius r
$\mathbf{1}_A(\cdot)$	Indicator function of set A
S^{d-1}	Unit sphere in R^d
$\cdot_{\mathsf{F}}$	Frobenius matrix norm
$\cdot_{\mathsf{op}}$	Operator matrix norm
λ_d	Lebesgue measure on R^d
(Ω, F, P)	Probability space considered in the following text
B^d	Borel σ-algebra on R^d
$P(\mathsf{R}^d)$	Space of all Borel probability measures on R^d
P_X	Distribution of random vector X
$X \quad P$	Random vector X with distribution P
$X \stackrel{D}{=} Y$	Random vectors X, Y with equal distributions
δ_x	Dirac measure concentrated in a point x
$N_d(\mu, \Sigma)$	d-variate Gaussian distribution with mean μ and covariance matrix Σ
Φ, φ	Distribution function and density of $N_1(0, 1)$
$P \quad Q$	P absolutely continuous w.r.t. Q
$\mathrm{d}P/\mathrm{d}Q$	Radon–Nikodym derivative of P w.r.t. Q
$\lvert\mathbf{A}\rvert$	Determinant of matrix $\mathbf{A}$
$\mathsf{tr}(\mathbf{A})$	Trace of matrix $\mathbf{A}$
$\sigma_{max}(\mathbf{S}), \sigma_{min}(\mathbf{S})$	Maximal and minimal eigenvalue of matrix $\mathbf{S} \quad \mathsf{PD}_d$
$\mathbf{S}^{1/2}$	Unique matrix $\mathbf{S}^{1/2} \quad \mathsf{PD}_d$ such that $\mathbf{S}^{1/2}\mathbf{S}^{1/2} = \mathbf{S} \quad \mathsf{PD}_d$
$a \quad b, a \quad b$	Maximum of $\{a, b\}$, minimum of $\{a, b\}$
a.e., a.s., w.r.t.	Almost everywhere, almost surely, with respect to

1. Location halfspace depth

We begin this thesis by defining a location parameter of a probability distribution.

Definition 1 (Location parameter) *Let P $\mathcal{P}(\mathbb{R}^d)$. Consider a functional $\Xi\colon P \to \mathbb{R}^d$ satisfying $\Xi(P_{\mathbf{A}X+b}) = \mathbf{A}\,\Xi(P_X)+b$ for any X P_X P, $\mathbf{A}$ GL_d and b $\mathbb{R}^d$ such that $P_{\mathbf{A}X+b}$ P. Then Ξ is called a* location functional *and $\Xi(P)$ is called a* location parameter *of P P.*

For the set $\mathcal{P}$ of all integrable d-variate random vectors, the mean value functional $\Xi(P) := \int_{\mathbb{R}^d} x\,\mathrm{d}P(x)$ is an example of a location functional. Another example is the univariate median functional defined on $\mathcal{P} = \mathcal{P}(\mathbb{R})$, i.e. $\Xi(P) = \arg\max_{x\ \mathbb{R}} \min\{P((-\infty, x]), P([x, \infty))\}$.

Considering a univariate probability distribution P $\mathcal{P}(\mathbb{R})$, one is able to tell how central any point x $\mathbb{R}$ is. This can be done by comparing the distribution function of P at x to $1/2$. That is, we can consider the median as the most central point, and points far from the median can be seen as the least central points. However, this approach strongly relies on the ordering of $\mathbb{R}$ and therefore cannot be straightforwardly generalized for multivariate distributions. The concept of location statistical depth offers a way to organize data points in a center-outward manner with respect to a given multivariate distribution. Therefore, it generalizes univariate rank and order statistics to multivariate data.

1.1 General notion of location statistical depth

As stated above, location statistical depth aims to quantify centrality. To do so, we need to define some notions of symmetry for d-variate random vectors (or equivalently for Borel probability measures on $\mathbb{R}^d$). In the following text, d $\mathbb{N}$ will always represent the dimension of the data points.

Definition 2 (Symmetry of random vector) *Let X P $\mathcal{P}(\mathbb{R}^d)$ be a random vector and consider a fixed point x_0 $\mathbb{R}^d$. We say that X (or P) is* centrally symmetric *about its center x_0 if*

$$X - x_0 \overset{D}{=} x_0 - X\,.$$

More generally, a random vector X (or its distribution P) is said to be angularly symmetric *about its center x_0 if*[1] *$(X - x_0)/\|X - x_0\|$ is centrally symmetric about $\mathbf{0}$. As an even more general notion, we say that X (or P) is* halfspace symmetric *about its center x_0 if $P(X \in H) \geq 1/2$ for every closed halfspace*[2] *H containing x_0.*

Note that every centrally symmetric distribution is also angularly symmetric, and every angularly symmetric distribution is halfspace symmetric. Based on the desirable properties a location statistical depth function should possess, Zuo and Serfling [2000a] introduced the subsequent definition.

[1] For ω Ω such that $X(\omega) = x_0$, this is considered to equal $\mathbf{0}$.

[2] By halfspace we mean any set in the form $\{x \mid \langle x, u\rangle \geq c\}$ for some u S^{d-1} and c $\mathbb{R}$.

Definition 3 (Location statistical depth) *A mapping $D\colon \mathbb{R}^d \times P(\mathbb{R}^d) \to [0, \infty)$ is called a* location statistical depth *(or simply* depth*) if it is bounded and satisfies the following properties:*

$(A1)$ *For any random vector $X \sim P_X \in P(\mathbb{R}^d)$, $\mathbf{A} \in \mathrm{GL}_d$, and $x, b \in \mathbb{R}^d$ it holds that $D(\mathbf{A}x + b; P_{\mathbf{A}X+b}) = D(x; P_X)$.*

$(A2)$ *$D(x_0; P) = \max_{x \in \mathbb{R}^d} D(x; P)$ holds for any symmetric $P \in P(\mathbb{R}^d)$ with center x_0.*

$(A3)$ *For any $P \in P(\mathbb{R}^d)$ with a deepest point x_0 and any $x \in \mathbb{R}^d$, the mapping $\alpha \to D(x_0 + \alpha(x - x_0); P)$ is non-increasing on $[0, 1]$.*

$(A4)$ *For any fixed $P \in P(\mathbb{R}^d)$, $D(x; P) \to 0$ as $\|x\| \to \infty$.*

By center, we mean centrality as described by one of the three types introduced in Definition 2.

The above definition should not be taken dogmatically. Numerous reasonable depth functions proposed in the literature may not satisfy all of the properties $(A1)$ – $(A4)$ (for a detailed analysis, see [Mosler and Mozharovskyi, 2022, Section 3]). Instead, these properties serve as guidelines. Property $(A1)$ is often referred to as *affine invariance*. It emphasizes that statistical procedures based on depth should not depend on the choice of the coordinate system. Properties $(A2)$ and $(A3)$ summarize the idea that the depth function should quantify the degree of centrality of the point of interest. Property $(A4)$ indicates that the measure of centrality of points far from the origin should be close to zero.

Definition 3 defines a population version of depth. In order to define the sample version, consider a random sample $X_1, \ldots, X_n$ generated by a distribution $P \in P(\mathbb{R}^d)$. We define the empirical probability measure

$$\widehat{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

Note that this is a random measure depending on the random sample. A sample version of depth is then defined as $D(\cdot; \widehat{P}_n)$. The point that maximizes depth can be thought of as a generalization of the univariate median. That is,

$$\arg\max_{x \in \mathbb{R}^d} D(x; \widehat{P}_n)$$

could serve as an estimator of the location parameter for d-variate random samples.

1.2 Definition of location halfspace depth

The location halfspace depth (introduced by Tukey [1975], therefore sometimes referred to as the Tukey depth) is defined as follows.

Definition 4 (Location halfspace depth) *Let $X \sim P \in P(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$. The* location halfspace depth *(or simply* halfspace depth*) of x w.r.t. P is defined as*

$$hD(x; P) := \inf_{u \in S^{d-1}} \mathsf{P}(\langle X, u \rangle \geq \langle x, u \rangle).$$

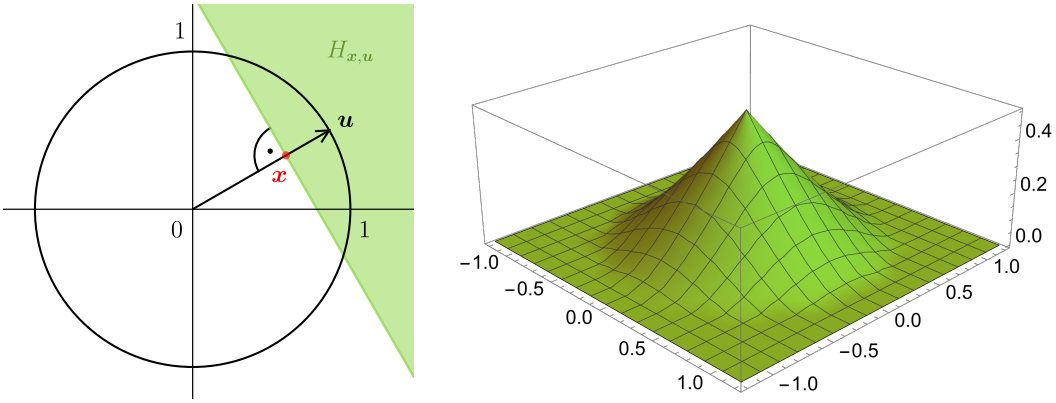
Note that for any fixed $P \in \mathcal{P}(\mathbb{R}^d)$, the function $hD(\cdot; P)$ is defined everywhere in $\mathbb{R}^d$ and it maps $\mathbb{R}^d$ to a subset of $[0, 1]$. For a point $x \in \mathbb{R}^d$ and a direction $u \in S^{d-1}$ denote

$$H_{x,u} := \{y \mid \langle y, u \rangle \geq \langle x, u \rangle\} \subset \mathbb{R}^d$$

the closed halfspace with x lying on its boundary and inner normal u. Then, the definition of halfspace depth can be rewritten as

$$hD(x; P) = \inf_{u \in S^{d-1}} P(H_{x,u}).$$

When calculating the depth of some point $x \in \mathbb{R}^d$, one is looking for a halfspace $H_{x,u}$ containing this point on its boundary with minimal probability. Every such halfspace is called a *minimizing halfspace*, and the corresponding direction u is called a *minimizing direction*.



(a) Minimizing halfspace $H_{x,u}$ at point x. Its boundary is perpendicular to the line connecting x and $\mathbf{0}$.

(b) 3D plot of the resulting halfspace depth.

Figure 1: The situation from Example 1.

Example 1 *Let P be the uniform distribution on the centered unit disc in $\mathbb{R}^2$, i.e. $\frac{dP}{d\lambda_2} = \frac{1}{\pi} \mathbf{1}_{B_2(\mathbf{0},1)}$. Fix a point $x \in \mathbb{R}^2$, $\|x\| < 1$, and consider a direction $u \in S^1$. Then, the probability $P(H_{x,u})$ equals the area of the disk segment, which is cut off from $B_2(\mathbf{0}, 1)$ by $H_{x,u}$, divided by π. See Figure 1a for reference. By basic tools of planimetry, we get*

$$P(H_{x,u}) = \frac{1}{\pi}\left(\arccos(\langle x, u \rangle) - \langle x, u \rangle \sqrt{1 - \langle x, u \rangle^2}\right). \tag{1.1}$$

Note that (1.1) is continuous in $\langle x, u \rangle$. In order to minimize it over u, we differentiate

$$\frac{d}{dt}\left(\arccos(t) - t\sqrt{1 - t^2}\right) = -2\sqrt{1 - t^2} < 0 \quad \text{for } |t| < 1.$$

By the Cauchy-Schwarz inequality, $\langle x, u \rangle \leq \|u\| \|x\| = \|x\|$. If $x = \mathbf{0}$, we obtain equality if and only if $u = x/\|x\|$ and if $x = \mathbf{0}$, then $\langle x, u \rangle = \|x\| = 0$ for all u. Summarizing these observations, we have

$$hD(x; P) = \inf_{u \in S^1} P(H_{x,u}) = \frac{1}{\pi}\left(\arccos \|x\| - \|x\| \sqrt{1 - \|x\|^2}\right) \quad \text{for } \|x\| < 1.$$

For a point x such that $\|x\| \geq 1$, its halfspace depth is obviously 0. The resulting halfspace depth is illustrated in Figure 1b.

A minimizing halfspace does not have to be unique. For instance, consider Example 1 and $x = \mathbf{0}$. Then, every direction $u \in S^1$ is minimizing. It is also possible that the value of halfspace depth is not attained in any direction, as shown in the following example.

Example 2 *Let P be the mixture of the uniform distribution on the centered unit disc and the Dirac measure concentrated in the point $z = (1, -1)^T$ with equal weights. See Figure 2 for reference. Consider the point $x = (1, 0)^T$. It is easy to see that $hD(x; P) = 0$. Indeed, let $u(\theta) = (\cos(\theta), \sin(\theta))^T$. Then $P(H_{x,u(\theta)}) \to 0$ as $\theta \to 0^+$, but $P(H_{x,u(\theta)}) > 0$ for all θ.*

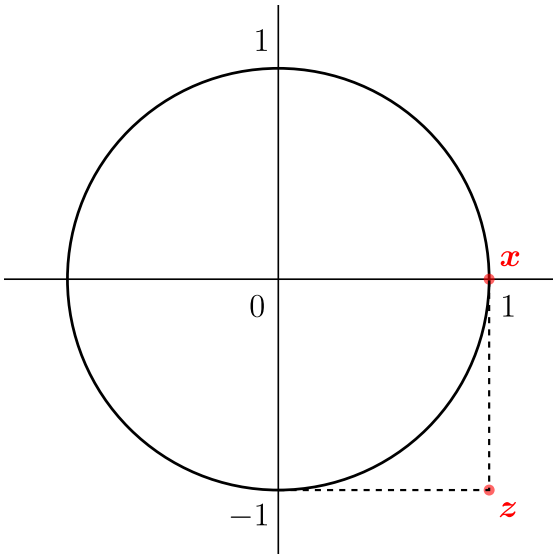


Figure 2: The situation from Example 2: For a mixture of the uniform distribution on the centered unit disc and the Dirac measure concentrated in the point z, there is no minimizing direction at point x.

For so-called *smooth* probability distributions, this situation can not happen.

Definition 5 (Smooth distribution) *Consider $P \in \mathcal{P}(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$. Then, P is said to be* smooth *at x if $P(\mathrm{bd}(H)) = 0$ for every halfspace $H \subset \mathbb{R}^d$ with x on its boundary. More generally, P is said to be* smooth *if it is smooth at every point $x \in \mathbb{R}^d$.*

If $P \in \mathcal{P}(\mathbb{R}^d)$ is smooth at $x \in \mathbb{R}^d$, then the minimizing halfspace at x always exists. This is an easy application of the Dominated convergence theorem (Theorem A1 in the Appendix). Indeed, fix a point x and consider a sequence $\{u_i\}_{i=1}^{\infty} \subset S^{d-1}$ such that $P(H_{x,u_i}) \to hD(x; P)$ as $i \to \infty$. As S^{d-1} is a compact set, the sequence $\{u_i\}_{i=1}^{\infty}$ has a cluster point. Without loss of generality, let $u_i \to u \in S^{d-1}$ as $i \to \infty$. Then,

$$P(H_{x,u}) - P(H_{x,u_i}) = \int_{\mathbb{R}^d} \left(\mathbf{1}_{H_{x,u}}(y) - \mathbf{1}_{H_{x,u_i}}(y)\right) \mathrm{d}P(y) \xrightarrow[i \to \infty]{} 0,$$

that is

$$hD(x; P) = P(H_{x,u}).$$

Regarding the assumptions of the Dominated convergence theorem, the difference in indicator functions is obviously bounded, and as $i \to \infty$, we have $u_i \to u$. Therefore, $\mathbf{1}_{H_{x,u}}(y) - \mathbf{1}_{H_{x,u_i}}(y) \to 0$ for all $y \in \mathbb{R}^d \setminus \mathrm{bd}(H_{x,u})$, that is P-a.e. because P is smooth at x.

It is also easy to see that all probability distributions $P \in \mathcal{P}(\mathbb{R}^d)$ which are absolutely continuous w.r.t. λ_d are smooth. But these two conditions are not equivalent. For example, consider the uniform distribution on the unit circle in $\mathbb{R}^2$. In the following section, we show that the location halfspace depth possesses more convenient properties for *smooth* distributions.

1.3 Basic properties of location halfspace depth

This section summarizes the well-known properties of the halfspace depth function from the literature. The goal of this work is not to establish fundamental properties but rather to emphasize robustness and minimax optimality of the location halfspace depth. Therefore, we do not provide proof for most theorems in this section.

The most important property of the location halfspace depth is that under certain conditions, it is a location statistical depth in the sense of Definition 3, as stated in the following theorem.

Theorem 6 *The location halfspace depth hD always satisfies properties $(A1)$, $(A3)$ and $(A4)$ from Definition 3. For a smooth P, it also satisfies property $(A2)$ for the halfspace symmetry (therefore also for the central and angular symmetry).*

This claim follows from several other fundamental properties of the halfspace depth that we are going to introduce before proving Theorem 6. The first easy consequence of its definition is quasiconcavity. A function $f\colon \mathbb{R}^d \to [0, \infty)$ is said to be quasiconcave if for any $x_1, x_2 \in \mathbb{R}^d$ and $\alpha \in [0, 1]$ it holds that

$$f(\alpha x_1 + (1 - \alpha)x_2) \geq \min\{f(x_1), f(x_2)\}.$$

Theorem 7 *For any fixed $P \in \mathcal{P}(\mathbb{R}^d)$, the halfspace depth $hD(\cdot; P)$ is quasiconcave.*

Proof. Refer to [Rousseeuw and Ruts, 1999, Proposition 1].

This property illuminates the geometry of the halfspace depth function. To be specific, we define α-depth regions as the upper-level sets of the halfspace depth function.

Definition 8 (α-depth region) *For any fixed $P \in \mathcal{P}(\mathbb{R}^d)$ and $\alpha \in [0, 1]$, the* α-depth region *of P is defined as*

$$D(\alpha; P) := \left\{x \in \mathbb{R}^d \mid hD(x; P) \geq \alpha\right\} \in \mathbb{R}^d.$$

With this notion, the quasiconcavity of $hD(\cdot; P)$ is equivalent to the convexity of all α-depth regions. Indeed, fix α $[0, 1]$ and x_1, x_2 $D(\alpha; P)$. Then, for any β $[0, 1]$ we have

$$hD(\beta x_1 + (1 - \beta)x_2; P) \geq \min\{hD(x_1; P), hD(x_2; P)\} \geq \alpha \qquad (1.2)$$

by the quasiconcavity of the halfspace depth. Hence $\beta x_1 + (1 - \beta)x_2$ $D(\alpha; P)$ and $D(\alpha; P)$ is a convex set. The other implication can be shown in a similar way.

Another property of interest is continuity. A function $f: \mathbb{R}^d \to [0, \infty)$ is defined to be upper semicontinuous if for any $t \geq 0$ the sets $\{x \quad \mathbb{R}^d \mid f(x) \geq t\}$ are closed. Every continuous function is obviously upper semicontinuous. The halfspace depth function is always upper semicontinuous. Further, it is continuous under certain conditions. Specifically, for a smooth distribution P, the function $hD(\cdot; P)$ is continuous, as demonstrated in the following theorem.

Theorem 9 *For any fixed P $P(\mathbb{R}^d)$, the function $hD(\cdot; P)$ is upper semicontinuous. If P is smooth, then $hD(\cdot; P)$ is continuous.*

Proof. Refer to [Donoho and Gasko, 1992, Lemma 6.1]. This lemma proves the continuity of the halfspace depth for absolutely continuous distributions. However, it is easy to see that its proof works for any smooth distribution.

The upper semi-continuity, together with Theorem 7, implies that the α-depth regions possess very plausible qualities to work with.

Theorem 10 *For any fixed P $P(\mathbb{R}^d)$ and α $[0, 1]$, the α-depth region $D(\alpha; P)$ is a closed convex set. If $\alpha = 0$, $D(\alpha; P)$ is also bounded, hence compact.*

Proof. This is a consequence of our Theorem 7, Theorem 9 and [Rousseeuw and Ruts, 1999, Proposition 5].

In Theorem 9, we noticed that the halfspace depth behaves especially nicely when the underlying probability distribution is smooth. This condition also implies that the halfspace depth function can be bounded from above by $1/2$, as formulated in the following theorem.

Theorem 11 *Let P $P(\mathbb{R}^d)$ be smooth. Then $hD(x; P) \leq 1/2$ holds for all x $\mathbb{R}^d$.*

Proof. Refer to [Rousseeuw and Ruts, 1999, Poposition 10]. This proposition establishes the upper bound for absolutely continuous distributions. However, it is evident that the proof applies to any smooth distribution.

Theorem 11 holds only for smooth probability distributions. Otherwise, there is no general upper bound smaller than 1. For example, consider $hD(\cdot; \delta_x)$ for some x $\mathbb{R}^d$. Then, it follows from the definition that $hD(x; \delta_x) = 1$. Now, we prove Theorem 6 to demonstrate the use of some of its basic properties mentioned above.

Proof of Theorem 6.

$(A1)$ Let $X \quad P_X \quad P(\mathsf{R}^d)$, $\mathbf{A} \quad \mathsf{GL}_d$ and $x, b \quad \mathsf{R}^d$. Then

$$\begin{aligned} hD(x; P_X) &= \inf_{u \ \mathsf{S}^{d-1}} \mathsf{P}(\ X, u \ \geq \ x, u\ ) \\ &= \inf_{u \ \mathsf{S}^{d-1}} \mathsf{P}\left(\frac{\left\langle X, \mathbf{A}^\mathsf{T} u \right\rangle}{\mathbf{A}^\mathsf{T} u} \geq \frac{\left\langle x, \mathbf{A}^\mathsf{T} u \right\rangle}{\mathbf{A}^\mathsf{T} u} \right) \\ &= \inf_{u \ \mathsf{S}^{d-1}} \mathsf{P}(\ \mathbf{A}X, u \ \geq \ \mathbf{A}x, u\ ) \\ &= \inf_{u \ \mathsf{S}^{d-1}} \mathsf{P}(\ \mathbf{A}X + b, u \ \geq \ \mathbf{A}x + b, u\ ) \\ &= hD(\mathbf{A}x + b; P_{\mathbf{A}X+b}). \end{aligned}$$

In the second equality we used that $\mathbf{A}^\mathsf{T}$ is non-singular, hence the mapping $u \to \left(\mathbf{A}^\mathsf{T} u\right) / \left\|\mathbf{A}^\mathsf{T} u\right\|$ maps S^{d-1} onto S^{d-1}.

$(A2)$ Consider $X \quad P_X$ for a smooth P_X and suppose that X is halfspace symmetric about $x_0 \quad \mathsf{R}^d$. Using the affine invariance, we can assume that $x_0 = \mathbf{0}$. By Theorem 11, it suffices to show that $hD(\mathbf{0}; P_X) = 1/2$. For any $u \quad \mathsf{S}^{d-1}$, we have (using the smoothness of P) $\mathsf{P}(\ X, u \ = 0) = 0$. Therefore, we obtain

$$\mathsf{P}(\ X, u \ \geq 0) = 1 - \mathsf{P}(\ X, u \ \leq 0)$$

which, using the halfspace symmetry, implies $\mathsf{P}(\ X, u \ \geq 0) = 1/2$ for all $u \quad \mathsf{S}^{d-1}$, hence $hD(\mathbf{0}; P_X) = 1/2$.

$(A3)$ Consider $X \quad P_X$ for a smooth P_X and let x_0 be a deepest point, i.e., let $x_0 = \arg\max_{x \ \mathsf{R}^d} hD(x; P_X)$. The function $hD(\cdot; P_X)$ is quasiconcave by Theorem 7. For a contradiction, suppose that there are $x \quad \mathsf{R}^d$ and $0 \leq \lambda_1 < \lambda_2 \leq 1$ such that

$$hD(x_0 + \lambda_1 (x - x_0); P_X) < hD(x_0 + \lambda_2 (x - x_0); P_X).$$

But due to the fact that $x_0 + \lambda_1 (x - x_0)$ lies in the line segment between x_0 and $x_0 + \lambda_2 (x - x_0)$, this contradicts the quasiconcavity.

$(A4)$ Fix $P \quad P(\mathsf{R}^d)$. Consider $\varepsilon > 0$. Then by Theorem 10, the ε-depth region $D(\varepsilon; P)$ is bounded. This implies that $hD(x; P) < \varepsilon$ for all x such that $\|x\|$ is large enough. Therefore, $hD(x; P) \to 0$ as $\|x\| \to \infty$.

Let us also show what the halfspace depth looks like for Gaussian distributions to demonstrate how it works.

Example 3 *Consider $X \quad N_d(\mu, \Sigma)$ for some $\mu \quad \mathsf{R}^d$ and $\Sigma \quad \mathsf{PD}_d$. Recall that the distribution function of the univariate standard Gaussian distribution is denoted by Φ. For all $x \quad \mathsf{R}^d$ we have*

$$hD(x; N_d(\mu, \Sigma)) = hD(\Sigma^{-1/2}(x - \mu); N_d(\mathbf{0}, \mathbf{I}_d)) \tag{1.3}$$

by a ne invariance. Therefore, we should first calculate $hD(y; N_d(\mathbf{0}, \mathbf{I}_d))$ *for all* $y \in \mathbb{R}^d$. *For* $Y \sim N_d(\mathbf{0}, \mathbf{I}_d)$ *and all* $u \in S^{d-1}$ *we have* $\langle Y, u \rangle \sim N_1(0, 1)$. *Hence,*

$$P(\langle Y, u \rangle \geq \langle y, u \rangle) = 1 - \Phi(\langle y, u \rangle) \geq 1 - \Phi(\|y\|), \tag{1.4}$$

where the inequality is due to the Cauchy-Schwarz theorem and for $u = y/\|y\|$, *we obtain equality. Therefore, this choice of* u *is the minimizing direction and* $hD(y; N_d(\mathbf{0}, \mathbf{I}_d)) = 1 - \Phi(\|y\|)$. *Returning back to the case of a general Gaussian distribution, by combining* (1.3) *and* (1.4), *we obtain that*

$$hD(x; N_d(\mu, \Sigma)) = 1 - \Phi\left(\left\|\Sigma^{-1/2}(x - \mu)\right\|\right).$$

Subsequently, we can see that the halfspace depth of Gaussian distribution is uniquely maximized for the mean μ *with the maximum depth* $1/2$. *Moreover, for* $\alpha \in (0, 1/2)$ *we have*

$$D(\alpha; P) = \left\{x \in \mathbb{R}^d \mid \sqrt{(x - \mu)^\top \Sigma^{-1}(x - \mu)} \leq \Phi^{-1}(1 - \alpha)\right\}.$$

These are the so-called Mahalanobis ellipsoids centered about μ. *The* $(1/2)-$*depth region* $D(1/2; P)$ *equals the singleton* $\{\mu\}$ *and* $D(\alpha; P) = \emptyset$ *for* $\alpha > 1/2$. *This corresponds to the fact that the Gaussian distribution is centrally symmetric about its mean.*

Calculating the depth function analytically is not an easy task in general. For some classes of distributions, it can be done relatively easily. Those are, for example, elliptically symmetric distributions and symmetric stable distributions, see [Chen and Tyler, 2004, Theorem 3.1]. For uniform distributions on convex bodies, this problem is closely related to the concept of floating bodies from geometry, as explored by Nagy et al. [2019].

1.4 Sample location halfspace depth and halfspace median

For halfspace depth to be useful from a statistical perspective, we must be able to estimate it based on a random sample from a distribution. Therefore, let us define the sample halfspace depth as follows.

Definition 12 (Sample halfspace depth) *Consider a random sample* $\{X_i\}_{i=1}^n$ *from a distribution* $P \in \mathcal{P}(\mathbb{R}^d)$ *and* $x \in \mathbb{R}^d$. *The* sample halfspace depth *of* x *w.r.t.* $\{X_i\}_{i=1}^n$ *is defined as*

$$\begin{aligned} hD(x; \{X_i\}_{i=1}^n) := hD(x; \widehat{P}_n) &= \min_{u \in S^{d-1}} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{\langle X_i, u \rangle \geq \langle x, u \rangle\}} \\ &= \min_{u \in S^{d-1}} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i \in H_{x,u}\}}. \end{aligned}$$

This definition implies that the sample halfspace depth of x is the smallest number of observations that can lie in a halfspace with x on its boundary, divided by n. A fundamental quality of estimators lies in how they behave for large sample sizes n. It turns out that the sample halfspace depth is strongly uniformly consistent.

Theorem 13 *For a random sample* $\{X_i\}_{i=1}^n \quad P \quad P(\mathbb{R}^d)$ *we have*

$$\sup_{x \ \mathbb{R}^d} |hD(x; \{X_i\}_{i=1}^n) - hD(x; P)| \xrightarrow[n \to \infty]{a.s.} 0.$$

Proof. Refer to [Donoho and Gasko, 1992, Section 6].

As mentioned above, the halfspace depth tells us how deep a given point is w.r.t. a distribution. It is, therefore, natural to define the deepest point as a generalization of the one-dimensional median. Such a point always exists, as stated in the following theorem.

Theorem 14 *Consider a probability measure* $P \quad P(\mathbb{R}^d)$. *Then, there exists at least one point* $y \quad \mathbb{R}^d$ *such that* $hD(y; P) = \sup_{x \ \mathbb{R}^d} hD(x; P)$.

Proof. Refer to [Rousseeuw and Ruts, 1999, Proposition 7].

Therefore, we can define the following notions.

Definition 15 (Maximal depth) *For* $P \quad P(\mathbb{R}^d)$, *we define the* maximal depth

$$\Pi(P) := \max_{x \ \mathbb{R}^d} hD(x; P)$$

and the set of points of maximal halfspace depth

$$\mathsf{Med}(P) := D(\Pi(P); P).$$

Using Theorems 10 and 14, the set $\mathsf{Med}(P)$ is non-empty, convex, and compact. Therefore, the following definition is well grounded.

Definition 16 (Halfspace median) *Define the functional* $T \colon P(\mathbb{R}^d) \to \mathbb{R}^d$ *so that it maps* $P \quad P(\mathbb{R}^d)$ *to the centroid*[3] *of* $\mathsf{Med}(P)$, *i.e. the centroid of*

$$\arg\max_{x \ \mathbb{R}^d} hD(x; P).$$

The value $T(P)$ *is called the* halfspace median *(or* Tukey's median*) of* P. *For a random sample* $\{X_i\}_{i=1}^n \quad P \quad P(\mathbb{R}^d)$, *we define the* sample halfspace median *(also called* sample Tukey's median*) as* $T(\widehat{P}_n)$.

From Theorem 11, we know that the maximal depth is bounded from above by $1/2$ for smooth distributions. Conversely, it turns out that the lower bound of $1/(d+1)$ for the maximum depth holds for all distributions.

Theorem 17 *For any* $P \quad P(\mathbb{R}^d)$ *we have* $\Pi(P) \geq \frac{1}{d+1}$.

[3]Centroid of a bounded, non-empty convex set $A \quad \mathbb{R}^d$ is defined as the expectation of the uniform distribution on A.

Proof. This claim is a consequence of Helly's theorem, which is a basic result in the field of discrete geometry. Proof can be found in [Rousseeuw and Ruts, 1999, Proposition 9].

The previous theorem implies that for any smooth P $P(\mathsf{R}^1)$, we have $\Pi(P) = 1/2$. To conclude this section, we summarize the basic properties of the halfspace median in the following theorem.

Theorem 18 *Let P $P(\mathsf{R}^d)$, then it holds that*

- *$T(P)$ is a location parameter of P in terms of Definition 1.*
- *If $d = 1$, then $T(P)$ coincides with the standard univariate median[4].*
- *The halfspace depth of $T(P)$ is at least $1/(d+1)$. If P is smooth, then the halfspace depth of $T(P)$ is bounded from above by $1/2$.*
- *If $\widehat{P}_n$ is the empirical measure corresponding to P, then $T(\widehat{P}_n) \xrightarrow[n \to \infty]{a.s.} T(P)$.*

Proof. The first property, affine equivariance, easily follows from the affine invariance of the halfspace depth (see Theorem 6). The second property arises from the fact that if $d = 1$, then $hD(x; P) = \min\{F(x), 1 - F(x^-)\}$, where F is the distribution function of P. The third property is derived from Theorem 11 and Theorem 17. The strong consistency follows from Lemma S.2.5 in the supplement to [Paindaveine and Van Bever, 2018].

1.5 Robustness of halfspace median

In this section, we discuss the robustness properties of the halfspace median. When inferring using parametric statistics, one assumes a parametric model. The properties of the estimators usually rely strongly on the fulfillment of the model assumptions. For example, the exact t-test assumes the normality of the underlying distribution. Without normality, the t-test works only asymptotically. The area of robust statistics explores how much statistical methods are affected by outliers or other small departures from model assumptions.

Take, for example, the problem of estimating the expectation (center) of a Gaussian distribution. The maximum likelihood theory yields the sample mean as an estimator. This estimator is unbiased and consistent. Furthermore, by the Cramér–Rao bound, we see that it is an efficient estimator. Thus, assuming a Gaussian model, it exhibits excellent properties. However, let us now assume that we are not entirely certain whether the data come from a Gaussian distribution. In the following text, we will demonstrate that in such situations, the sample halfspace median has good robustness properties against such violations of assumptions.

[4] This holds if we define the univariate median of X P $P(\mathsf{R})$ as the centerpoint of all points x R such that $\mathsf{P}(X \leq x) \geq 1/2$ and $\mathsf{P}(X \geq x) \geq 1/2$.

Assume that we have a statistical functional $\Xi\colon P(\mathbb{R}^d) \to \mathbb{R}^d$. In this section, we consider several measures of its robustness. The departure from model assumptions will be represented by contaminated distributions.

Definition 19 (Contaminated distribution) *Let $P, Q \in P(\mathbb{R}^d)$, $\varepsilon \in [0,1]$ and $x \in \mathbb{R}^d$. We consider the distribution $(1-\varepsilon)P + \varepsilon Q \in P(\mathbb{R}^d)$ defined as*

$$((1-\varepsilon)P + \varepsilon Q)(B) = (1-\varepsilon)P(B) + \varepsilon Q(B) \quad \text{for } B \in \mathcal{B}^d.$$

Specifically, we denote

$$P_{(\varepsilon,Q)} := (1-\varepsilon)P + \varepsilon Q,$$
$$P_{(\varepsilon,x)} := (1-\varepsilon)P + \varepsilon \delta_x.$$

One can define the influence function and gross error sensitivity using the notion of contaminated distributions.

Definition 20 (Influence function, gross error sensitivity) *Consider a location statistical functional $\Xi\colon P(\mathbb{R}^d) \to \mathbb{R}^d$ and a probability measure $P \in P(\mathbb{R}^d)$. The* influence function *of Ξ at P is a function of $x \in \mathbb{R}^d$ defined as*

$$IF(x; \Xi, P) := \lim_{\varepsilon \to 0^+} \frac{\Xi(P_{(\varepsilon,x)}) - \Xi(P)}{\varepsilon},$$

provided that the limit exists for all $x \in \mathbb{R}^d$. The gross error sensitivity *of Ξ at P is defined as*

$$\gamma(\Xi, P) := \sup_{x \in \mathbb{R}^d} IF(x; \Xi, P) \ .$$

The influence function (being one-sided Gateaux derivative of Ξ) describes how the contamination by a point x with infinitesimally small mass ε changes the value of the statistical functional, normalized by ε. In general, one seeks functionals with bounded influence function to get desirable robustness properties. On the other hand, gross error sensitivity evaluates the maximum effect that an infinitesimally small point-mass contamination can have, indicating its reliability under extreme conditions.

Both previous measures indicate robustness only w.r.t. a point-mass contamination. A stronger approach is to measure the maximum impact that any form of contamination can exert on a functional. This leads to a definition of maximum bias and breakdown point.

Definition 21 (Maximum bias, breakdown point) *Consider a statistical functional $\Xi\colon P(\mathbb{R}^d) \to \mathbb{R}^d$, a probability measure $P \in P(\mathbb{R}^d)$ and $\varepsilon \in [0,1]$. The* maximum bias *function of Ξ at P is defined as*

$$B(\varepsilon; \Xi, P) := \sup_{Q \in P(\mathbb{R}^d)} \left\| \Xi(P_{(\varepsilon,Q)}) - \Xi(P) \right\|.$$

The breakdown point *of Ξ at P is defined as*

$$\varepsilon(\Xi, P) := \inf \{ \varepsilon \in (0,1] \mid B(\varepsilon; \Xi, P) = \infty \}.$$

The breakdown point is the maximal proportion ε of contamination the statistical functional can handle before producing an arbitrarily large bias.

Recall that the functional corresponding to the halfspace median is denoted by T. The *support* of a probability measure $P \in \mathcal{P}(\mathbb{R}^d)$, usually denoted by $\mathrm{supp}(P)$, is defined as the complement of the union of all open sets which are P-null sets, i.e. the smallest closed set A such that $P(\mathbb{R}^d \setminus A) = 0$. Using this notion, we say that P has *contiguous support* if $\mathrm{supp}(P)$ cannot be separated by a zero P-mass slab of a non-empty interior. Here, by slab, we mean any set of the form $\left\{x \in \mathbb{R}^d \mid a < \langle x, u \rangle < b\right\}$ for some $u \in \mathbb{S}^{d-1}$ and $a, b \in \mathbb{R}$. If we assume that a probability measure P has contiguous support and is smooth at some $x \in \mathrm{Med}(P)$, then $hD(\cdot; P)$ is strictly monotone. That is, for all $\alpha \in (0, \Pi(P))$, we have

$$\mathrm{int}\left(D(\alpha; P)\right) = \left\{x \in \mathbb{R}^d \mid hD(x; P) > \alpha\right\}.$$

This has been proven by Laketa and Nagy [2022, Section 4.3]. Also, recall that P is said to be halfspace symmetric about x if $P(H) \geq 1/2$ for every halfspace H with x on its boundary (see Definition 2). Zuo and Serfling [2000b, Theorem 2.1] have shown that for every probability measure P halfspace symmetric about x, we have $\mathrm{Med}(P) = \{x\}$ unless P is concentrated on a line and its distribution has more than one univariate median. That is, P is supported on some one-dimensional affine subspace of $\mathbb{R}^d$, and $\mathrm{Med}(P)$ is a line segment in this subspace, symmetric around x.

In order to determine the influence function of T at P, we need to calculate the halfspace median for the contaminated distribution $P_{(\varepsilon,x)}$. This is not an easy task in general, but as shown in the following theorem, it can be done after imposing an additional assumption on P.

Definition 22 *Consider $P \in \mathcal{P}(\mathbb{R}^d)$, $\alpha \in (0, \Pi(P)]$ and $u \in \mathbb{S}^{d-1}$. Then, $r(\alpha, u)$ is defined as the radius of the α-depth region $D(\alpha; P)$ along the direction $u \in \mathbb{S}^{d-1}$, i.e. the length of*

$$D(\alpha; P) \cap \{T(P) + s\,u \mid s \geq 0\}.$$

In the subsequent theorem, we restrict ourselves to $d > 1$, which is the interesting case. For $d = 1$, the halfspace median coincides with the usual univariate median, and its robustness properties are well known. See, for example, [Huber and Ronchetti, 2009, Section 3.2].

Theorem 23 *Let $d > 1$. Consider a probability measure $P \in \mathcal{P}(\mathbb{R}^d)$, which is halfspace symmetric about $x_0 \in \mathbb{R}^d$, smooth at x_0 and has contiguous support. Then for any $\varepsilon \in [0, 1/3)$ we have*

$$T(P_{(\varepsilon,x)}) = \begin{cases} x & \text{for } x \in \mathrm{int}\left(D\left(\frac{1-3\varepsilon}{2(1-\varepsilon)}; P\right)\right), \\ x_0 + \frac{1}{2}\, r\left(\frac{1-3\varepsilon}{2(1-\varepsilon)}, \frac{x-x_0}{\|x-x_0\|}\right) \frac{x-x_0}{\|x-x_0\|} & \text{for } x \notin \mathrm{int}\left(D\left(\frac{1-3\varepsilon}{2(1-\varepsilon)}; P\right)\right). \end{cases}$$

Note *This theorem is due to Chen and Tyler [2002, Theorem 3.1]. However, the authors do not assume contiguous support of P. Without this assumption, the theorem does not hold, as demonstrated in Example 4. Additionally, Chen and Tyler [2002] assume that P is absolutely continuous w.r.t. the Lebesgue measure. This is not necessary. We replace this assumption with smoothness of P at its center. Finally, that paper does not explicitly state that $d > 1$ is required for the proof to be valid.*

Proof. Using the affine equivariance of T, we can assume $x_0 = \mathbf{0}$. If $x = \mathbf{0}$, the claim is obvious. Let $x = \mathbf{0}$.

- By the assumption of halfspace symmetry and the smoothness of P at $\mathbf{0}$, we have $P(H_{\mathbf{0},u}) = 1/2$ for all $u \in S^{d-1}$. Evidently, there exists $u \in S^{d-1}$ such that $H_{\mathbf{0},u}$ does not include x. Therefore, $hD(\mathbf{0}; P_{(\varepsilon,x)}) = \frac{1-\varepsilon}{2}$.

- Consider any $y \neq \{\alpha x \mid \alpha \in [0,1]\}$. Then, there is $u \in S^{d-1}$ such that $H_{y,u}$ does not include the set $\{\alpha x \mid \alpha \in [0,1]\}$. Because P has contiguous support and $H_{y,u}$ does not include $\mathbf{0}$, we have $P(H_{y,u}) < 1/2$. Therefore, $P_{(\varepsilon,x)}(H_{y,u}) = (1-\varepsilon)\, P(H_{y,u}) < \frac{1-\varepsilon}{2}$, hence $hD(y; P_{(\varepsilon,x)}) < \frac{1-\varepsilon}{2}$.

- On the other hand, for any $y \in \{\alpha x \mid \alpha \in (0,1)\}$, we can choose a sequence $\{u_i\}_{i=1}^{\infty} \subset S^{d-1}$ such that for all $i \in \mathbb{N}$, H_{y,u_i} contains $\mathbf{0}$, does not contain x and the distance between $\mathrm{bd}\,(H_{y,u_i})$ and $\mathbf{0}$ tends to 0 as $i \to \infty$. This is possible only because we assume that $d > 1$. As a result, $P_{(\varepsilon,x)}(H_{y,u_i}) = (1-\varepsilon)P(H_{y,u_i}) \to \frac{1-\varepsilon}{2}$ as $i \to \infty$ and $hD(y; P_{(\varepsilon,x)}) \leq \frac{1-\varepsilon}{2}$.

We have shown that $\Pi\left(P_{(\varepsilon,x)}\right) \geq \frac{1-\varepsilon}{2}$, $\mathrm{Med}\left(P_{(\varepsilon,x)}\right) \subset \{\alpha x \mid \alpha \in [0,1]\}$ and $hD(\cdot; P_{(\varepsilon,x)})$ is bounded from above by $\frac{1-\varepsilon}{2}$ on $\{\alpha x \mid \alpha \in [0,1)\}$. Now, we consider three different cases.

- If $x \in \mathrm{int}\left(D\left(\frac{1-3\varepsilon}{2(1-\varepsilon)}; P\right)\right)$, then $hD(x; P_{(\varepsilon,x)}) > (1-\varepsilon)\frac{1-3\varepsilon}{2(1-\varepsilon)} + \varepsilon = \frac{1-\varepsilon}{2}$ since by the assumptions $hD(\cdot; P)$ is strictly monotone. Therefore, x is the only point of the maximal depth, hence $T(P_{(\varepsilon,x)}) = x$, as claimed.

- Let $x \in \mathrm{bd}\left(D\left(\frac{1-3\varepsilon}{2(1-\varepsilon)}; P\right)\right)$. Using the strict monotonicity of $hD(\cdot; P)$, we have that $hD(x; P) = \frac{1-3\varepsilon}{2(1-\varepsilon)}$. Therefore, $hD(x; P_{(\varepsilon,x)}) = \frac{1-\varepsilon}{2}$, hence $\Pi\left(P_{(\varepsilon,x)}\right) = \frac{1-\varepsilon}{2}$. We already know that $hD(\mathbf{0}; P_{(\varepsilon,x)}) = \frac{1-\varepsilon}{2}$ and since the depth regions are convex, we have $\mathrm{Med}\left(P_{(\varepsilon,x)}\right) = \{\alpha x \mid \alpha \in [0,1]\} = D\left(\frac{1-3\varepsilon}{2(1-\varepsilon)}; P\right) \cap \{\alpha x \mid \alpha \in [0,1]\}$. Because the centroid of a line segment equals its center point, the claim follows readily.

- Finally, if $x \notin D\left(\frac{1-3\varepsilon}{2(1-\varepsilon)}; P\right)$, then $hD(x; P) < \frac{1-3\varepsilon}{2(1-\varepsilon)}$. As a consequence, $hD(x; P_{(\varepsilon,x)}) < \frac{1-3\varepsilon}{2} + \varepsilon = \frac{1-\varepsilon}{2}$. Consider $y \in \{\alpha x \mid \alpha \in (0,1)\}$. Much like before, if $y \notin D\left(\frac{1-3\varepsilon}{2(1-\varepsilon)}; P\right)$, then $hD(y; P_{(\varepsilon,x)}) < \frac{1-\varepsilon}{2}$. On the other hand, if $y \in D\left(\frac{1-3\varepsilon}{2(1-\varepsilon)}; P\right)$, then $(1-\varepsilon)\, P(H) \geq \frac{1-3\varepsilon}{2} = \frac{1-\varepsilon}{2} - \varepsilon$ for all halfspaces H with y on boundary. Also, $(1-\varepsilon)\, P(H) \geq \frac{1-\varepsilon}{2}$ for all such halfspaces that contain $\mathbf{0}$ and hence do not contain x. As a consequence, $P_{(\varepsilon,x)}(H) \geq \frac{1-\varepsilon}{2}$ for all halfspaces H with y on boundary, therefore $hD(y; P_{(\varepsilon,x)}) = \frac{1-\varepsilon}{2}$. We have used the previous observation that $hD(\cdot; P_{(\varepsilon,x)})$ is bounded from above by $\frac{1-\varepsilon}{2}$ on $\{\alpha x \mid \alpha \in [0,1)\}$. Summarizing this, we have $\mathrm{Med}\left(P_{(\varepsilon,x)}\right) = D\left(\frac{1-3\varepsilon}{2(1-\varepsilon)}; P\right) \cap \{\alpha x \mid \alpha \in [0,1]\}$. The claim follows just like in the previous case.

Example 4 *Let $P \in \mathcal{P}(\mathbb{R}^2)$ be the uniform distribution on the union of two rectangles with vertices*

$$(-8,-6)^\top, (-5,-6)^\top, (-5,6)^\top, (-8,6)^\top \text{ and } (5,-6)^\top, (8,-6)^\top, (8,6)^\top, (5,6)^\top,$$

see Figure 3. Then, P is evidently halfspace symmetric about the origin $\mathbf{0}$ and is smooth at $\mathbf{0}$. Consider $\varepsilon = 1/4$ and $x = (0,4)^\top$. We have $1-\varepsilon = 3/4$ and $\varepsilon/(1-\varepsilon) = 1/3$. By considering all halfplanes with x on their boundary, it is easy to see that $hD(x;P) = 1/6$ and

$$x \quad \text{bd}\left(D\left(\frac{1-3\varepsilon}{2(1-\varepsilon)}; P\right)\right) = \text{bd}\left(D\left(\frac{1}{6}; P\right)\right),$$

hence

$$r\left(\frac{1-3\varepsilon}{2(1-\varepsilon)}, \frac{x}{x}\right) = \|x\| = 4. \tag{1.5}$$

Further, let the points a and b be defined as in Figure 3 and denote by D the convex hull of the set $\{\mathbf{0}, a, b, x\}$. Then, using the definition of $P_{(\varepsilon,x)}$, after considering all possible halfplanes, we realize that

- $hD\left(\mathbf{0}; P_{(\varepsilon,x)}\right) = (1-\varepsilon)/2 = \frac{3}{8}$,
- $hD\left(x; P_{(\varepsilon,x)}\right) = (1-\varepsilon)hD(x;P) + \varepsilon = \frac{3}{8}$,
- $hD\left(a; P_{(\varepsilon,x)}\right) = hD\left(b; P_{(\varepsilon,x)}\right) = (1-\varepsilon)/2 = \frac{3}{8}$ *and*
- $hD\left(y; P_{(\varepsilon,x)}\right) < \frac{1-\varepsilon}{2}$ *for all* $y \notin D$.

Therefore, using the quasiconcavity of the halfspace depth, we deduce that

$$D = D\left(\frac{3}{8}; P_{(\varepsilon,x)}\right).$$

It is also easy to see that for all points y between $\mathbf{0}$ and x, it holds that $hD(y;P_{(\varepsilon,x)}) \le \frac{1-\varepsilon}{2} = \frac{3}{8}$. Therefore, using the quasiconcavity of $hD(\cdot;P_{(\varepsilon,x)})$ again, it holds that

$$\Pi\left(P_{(\varepsilon,x)}\right) = \frac{3}{8} \quad \text{and} \quad \text{Med}\left(P_{(\varepsilon,x)}\right) = D.$$

As a consequence, $T(P_{(\varepsilon,x)})$ is the centroid of D, marked as a red point in Figure 3. However, this point does not lie exactly in the middle between $\mathbf{0}$ and x. This, together with (1.5), shows that the assumption of contiguous support is necessary in Theorem 23.

From the preceding Theorem 23, we observe that as we move x farther away from x_0, the Tukey median of the contaminated distribution $P_{(\varepsilon,x)}$ (as a function of the contaminating point x), initially increases linearly along a ray and subsequently decreases to a fixed value. This behavior will be illustrated in Example 5, where P is a Gaussian distribution. Theorem 23 also allows us, under certain assumptions, to determine the influence function and gross error sensitivity for the Tukey median.

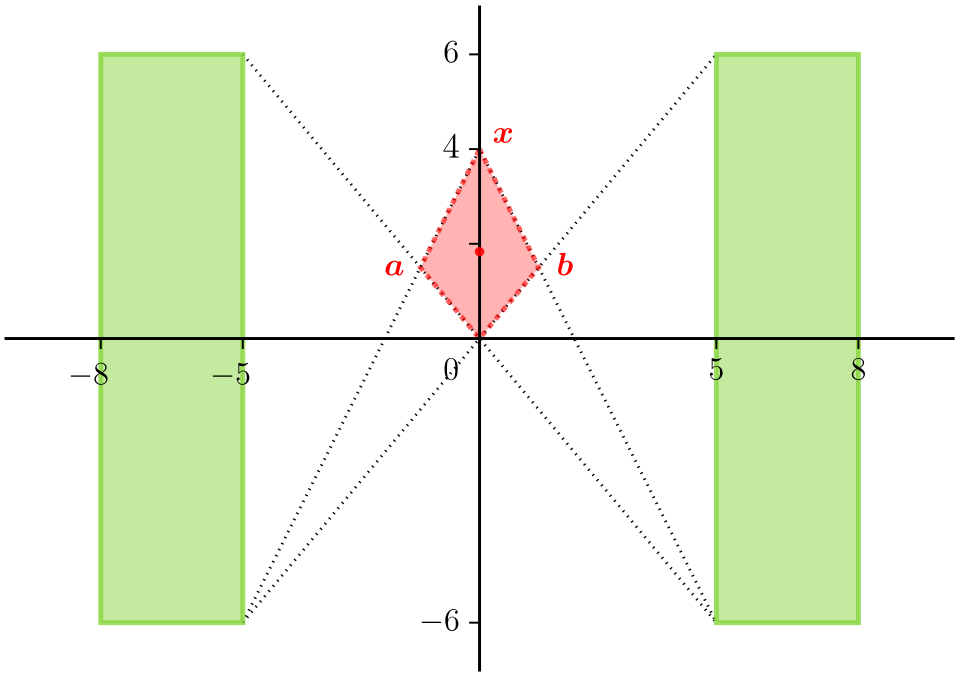


Figure 3: The situation from Example 4. The Tukey median of a uniform distribution on the union of two green rectangles, contaminated by $\boldsymbol{x}$, is marked by a red dot. This point does not lie exactly in the middle between $\mathbf{0}$ and $\boldsymbol{x}$, which contradicts [Chen and Tyler, 2002, Theorem 3.1].

Theorem 24 *Under the conditions of Theorem 23 it holds that*

$$IF(\boldsymbol{x};T,P) = \begin{cases} \mathbf{0} & \text{for } \boldsymbol{x} = \boldsymbol{x}_0, \\ \lim_{\alpha \to 0+} \left(\frac{1}{2\alpha}\, r\left(\frac{1}{2} - \alpha, \frac{\boldsymbol{x}-\boldsymbol{x}_0}{\boldsymbol{x}-\boldsymbol{x}_0}\right)\right) \frac{\boldsymbol{x}-\boldsymbol{x}_0}{\boldsymbol{x}-\boldsymbol{x}_0} & \text{for } \boldsymbol{x} = \boldsymbol{x}_0, \end{cases} \tag{1.6}$$

provided that the limit exists.

This theorem states that, under the specified assumptions, the influence function remains constant along rays originating from the center of symmetry. This implies boundedness of the influence function, indicating good robustness properties of Tukey's median.

Proof. For $\boldsymbol{x} = \boldsymbol{x}_0$, the claim is obvious. Consider $\boldsymbol{x} = \boldsymbol{x}_0$. Then, for ε small enough, we have $\boldsymbol{x} \;/\; \mathsf{int}\left(D\left(\frac{1-3\varepsilon}{2(1-\varepsilon)}; P\right)\right)$. Therefore, by Theorem 23, we have

$$\begin{aligned} IF(\boldsymbol{x};T,P) &= \lim_{\varepsilon \to 0+} \frac{1}{2\,\varepsilon}\, r\left(\frac{1-3\varepsilon}{2(1-\varepsilon)}, \frac{\boldsymbol{x}-\boldsymbol{x}_0}{\boldsymbol{x}-\boldsymbol{x}_0}\right) \frac{\boldsymbol{x}-\boldsymbol{x}_0}{\boldsymbol{x}-\boldsymbol{x}_0} \\ &= \lim_{\varepsilon \to 0+} \frac{1}{2\,\varepsilon}\, r\left(\frac{1}{2} - \frac{\varepsilon}{(1-\varepsilon)}, \frac{\boldsymbol{x}-\boldsymbol{x}_0}{\boldsymbol{x}-\boldsymbol{x}_0}\right) \frac{\boldsymbol{x}-\boldsymbol{x}_0}{\boldsymbol{x}-\boldsymbol{x}_0}, \end{aligned}$$

which is evidently equivalent to (1.6).

The formula for the gross error sensitivity follows immediately.

Theorem 25 *Let the conditions of Theorem 23 be satisfied and further assume that the limit*

$$\lim_{\alpha \to 0+} \left(\frac{1}{\alpha} r\left(\frac{1}{2} - \alpha, u \right) \right)$$

exists for all $u \in S^{d-1}$. *Then,*

$$\gamma(T, P) = \sup_{u \in S^{d-1}} \lim_{\alpha \to 0+} \left(\frac{1}{2\alpha} r\left(\frac{1}{2} - \alpha, u \right) \right).$$

Proof. The claim follows from Theorem 24.

Now we will demonstrate the application of the preceding theorems for the Tukey median of a contaminated Gaussian distribution.

Example 5 *Consider* $P = N_d(\mu, \Sigma)$. *The assumptions of Theorem 23 are obviously satisfied for* $x_0 = \mu$. *Recall that the distribution function of the univariate standard Gaussian distribution and its density are denoted by* Φ *and* φ, *respectively. From Example 3 we know that*

$$D\left(\frac{1}{2} - \alpha; N_d(\mu, \Sigma) \right) = \left\{ y \in \mathbb{R}^d \mid (y - \mu)^\top \Sigma^{-1} (y - \mu) \le \left(\Phi^{-1}\left(\frac{1}{2} + \alpha \right) \right)^2 \right\}.$$

For $x = \mu$, *the width of this ellipsoid along the direction* $(x - \mu)/\|x - \mu\|$ *is*

$$2\, r\left(\frac{1}{2} - \alpha, \frac{x - \mu}{\|x - \mu\|} \right) = 2\, \Phi^{-1}\left(\frac{1}{2} + \alpha \right) \frac{\left\| \Sigma^{1/2}(x - \mu) \right\|}{\|x - \mu\|}.$$

Therefore, by Theorem 23 we have

$$T(P_{(\varepsilon, x)}) = \mu + \frac{1}{2} \Phi^{-1}\left(\frac{1}{2} + \alpha \right) \frac{\left\| \Sigma^{1/2}(x - \mu) \right\|}{\|x - \mu\|^2} (x - \mu)$$

and by Theorem 24

$$\begin{aligned} IF(x; T, N_d(\mu, \Sigma)) &= \lim_{\alpha \to 0+} \left(\frac{1}{2\alpha} \Phi^{-1}\left(\frac{1}{2} + \alpha \right) \right) \frac{\left\| \Sigma^{1/2}(x - \mu) \right\|}{\|x - \mu\|^2} (x - \mu) \\ &= \frac{\left\| \Sigma^{1/2}(x - \mu) \right\|}{2\, \varphi(\Phi^{-1}(1/2))\, \|x - \mu\|^2} (x - \mu) = \frac{\left\| \Sigma^{1/2}(x - \mu) \right\|}{2 \sqrt{2\pi}\, \|x - \mu\|^2} (x - \mu), \end{aligned}$$

where the limit was evaluated using the L'Hôpital's rule. For $x = \mu$, *we have* $IF(\mu; T, N_d(\mu, \Sigma)) = 0$. *Furthermore, because the spectral norm of* $\Sigma^{1/2}$ *equals* $\sqrt{\sigma_{max}(\Sigma)}$, *we have* $\left\| \Sigma^{1/2}(x - \mu) \right\| / \|x - \mu\| \le \sqrt{\sigma_{max}(\Sigma)}$. *For* x *such that* $x - \mu$ *is the corresponding eigenvector of* Σ *(or* $\Sigma^{1/2}$*), we obtain equality. Therefore*

$$\gamma(T, N_d(\mu, \Sigma)) = \sqrt{\frac{\sigma_{max}(\Sigma)}{8\pi}}.$$

This means that a larger maximal eigenvalue of the covariance matrix Σ *suggests greater variability in the data, potentially making the estimator more sensitive to outliers or extreme values.*

Now, we would like to determine the effect of ε-contamination by an arbitrary probability distribution Q. In this case, it is not possible to find any general expression for the halfspace median of the contaminated distribution $P_{(\varepsilon,Q)}$. However, as shown in the following theorem, the maximum possible bias for halfspace symmetric distributions is produced by point-mass contamination. For a general distribution, we are at least able to bound the maximum bias. This allows us to determine the breakdown point of the halfspace median. The subsequent theorem and its proof is taken from [Chen and Tyler, 2002, Theorem 4.1].

Theorem 26 *For any* $P \in \mathcal{P}(\mathbb{R}^d)$, $d > 1$ *and* $\varepsilon \in \left(0, \frac{\Pi(P)}{1+\Pi(P)}\right)$ *we have*

$$B(\varepsilon; T, P) \leq \sup_{u \in S^{d-1}} r\left(\Pi(P) - \frac{\varepsilon}{(1-\varepsilon)}, u\right) =: \left\| D\left(\Pi(P) - \frac{\varepsilon}{(1-\varepsilon)}; P\right)\right\|_r, \quad (1.7)$$

hence $\frac{\Pi(P)}{1+\Pi(P)} \leq \varepsilon(T, P)$. *If* P *is smooth, then we also have* $\varepsilon(T, P) \leq 1/3$. *Finally, if* P *satisfies conditions of Theorem 23 and is smooth, then for all* $\varepsilon \in (0, 1/3)$ *it holds that*

$$B(\varepsilon; T, P) = \left\| D\left(\frac{1}{2} - \frac{\varepsilon}{(1-\varepsilon)}; P\right)\right\|_r \quad (1.8)$$

and $\varepsilon(T, P) = 1/3$.

Proof.

- Fix $Q \in \mathcal{P}(\mathbb{R}^d)$. For all halfspaces H with $T(P)$ on its boundary, we have $P(H) \geq \Pi(P)$. Therefore, $hD(T(P); P_{(\varepsilon,Q)}) \geq (1-\varepsilon)\Pi(P)$. This implies $\Pi(P_{(\varepsilon,Q)}) \geq (1-\varepsilon)\Pi(P)$. Consider $y \notin D\left(\Pi(P) - \frac{\varepsilon}{(1-\varepsilon)}; P\right)$, then

$$\begin{aligned} hD(y; P_{(\varepsilon,Q)}) &\leq (1-\varepsilon) hD(y; P) + \varepsilon \\ &< (1-\varepsilon)\left(\Pi(P) - \frac{\varepsilon}{1-\varepsilon}\right) + \varepsilon \\ &= (1-\varepsilon)\Pi(P). \end{aligned}$$

  This means that $\mathrm{Med}\left(P_{(\varepsilon,Q)}\right) \notin D\left(\Pi(P) - \frac{\varepsilon}{(1-\varepsilon)}; P\right)$. Since this depth region is bounded by Theorem 10, this proves the upper bound (1.7) for an arbitrary distribution P. Consequently we have $\frac{\Pi(P)}{1+\Pi(P)} \leq \varepsilon(T, P)$.

- For P which is smooth at all points, we can consider $Q = \delta_x$ for $x \in \mathbb{R}^d$ and $\varepsilon > 1/3$. Then, we have $hD(x; P_{(\varepsilon,x)}) \geq \varepsilon > 1/3$. On the other hand, consider arbitrary $y \in \mathbb{R}^d, y \neq x$, and find $u \in S^{d-1}$ such that $\langle u, x - y\rangle = 0$. Then, using the smoothness of P, we have that at least one of $H_{y,u}$, $H_{y,-u}$ has P-mass less than or equal to $1/2$. Without loss of generality, let $P(H_{y,u}) \leq 1/2$. Now, consider a sequence $\{u_i\}_{i=1}^{\infty} \subset S^{d-1}$ such that $x \notin H_{y,u_i}$ and $u_i \xrightarrow[i\to\infty]{} u$. Using the smoothness of P, we have $P_{(\varepsilon,x)}(H_{y,u_i}) \xrightarrow[i\to\infty]{} (1-\varepsilon)P(H_{y,u}) \leq (1-\varepsilon)/2 < 1/3$. This implies that $hD(y; P_{(\varepsilon,x)}) < 1/3 < hD(x; P_{(\varepsilon,x)})$. Therefore, x is the halfspace median of $P_{(\varepsilon,x)}$ and by letting $\|x\| \to \infty$, we can produce an arbitrarily large bias. Therefore, $\varepsilon(T, P) \leq 1/3$.

- Recall that under conditions of Theorem 23 we have $\Pi(P) = 1/2$, hence $\frac{\Pi(P)}{1+\Pi(P)} = 1/3$. Therefore, for any $\varepsilon \in (0, 1/3)$ and $x \in \operatorname{int}\left(D\left(\frac{1}{2} - \frac{\varepsilon}{(1-\varepsilon)}; P\right)\right)$ we can use Theorem 23 to show that $T(P_{(\varepsilon,x)}) = x$. As a consequence, considering $Q = \delta_x$ proves equality (1.8). The equality $\varepsilon(T, P) = 1/3$ now follows from the first and the second part.

The previous theorem implies that the greater the maximal depth of a point w.r.t. a given distribution, the greater the breakdown point of Tukey's median. By Theorem 17 we know that for any $P \in \mathcal{P}(\mathbb{R}^d)$, the lower bound $\Pi(P) \geq 1/(d+1)$ holds. Together with the previous theorem, this implies that $\varepsilon(T, P) \geq 1/(d+2)$. However, the argument of Theorem 26 can be refined so that we obtain $\varepsilon(T, P) \geq 1/(d+1)$, see [Chen and Tyler, 2002, Theorem 4.2]. Note that this implies that for any smooth $P \in \mathcal{P}(\mathbb{R}^2)$ we have $\varepsilon(T, P) = 1/3$.

All robustness measures introduced above are based on the population distribution P. However, we can also investigate the robustness of the halfspace median based on finite samples. For example, a finite-sample version of the influence function is the sensitivity curve, see [Huber and Ronchetti, 2009, Section 1.5]. For a dataset $X = \{x_1, \ldots, x_n\}$, Donoho and Gasko [1992] define the finite-sample breakdown point as

$$\varepsilon^*(T, X) = \min_{m \in \mathbb{N}} \left\{ \frac{m}{n+m} \,\middle|\, \sup\left\{ T(X \cup Y^m) - T(X) \;\middle|\; Y^m = \{y_1, \ldots, y_m\} \subset \mathbb{R}^d \right\} = \infty \right\},$$

where $T(X)$ is the halfspace median calculated from X and $T(X \cup Y^m)$ is the halfspace median calculated from $X \cup Y^m$. That is, we contaminate out dataset X by m points Y^m. If, by some choice of Y^m, this shifts the halfspace median arbitrarily far from the original halfspace median, we say that T breaks down. Then, $\varepsilon^*(T, X)$ is the smallest contamination fraction $m/(n+m)$ such that T can break down. Donoho and Gasko [1992, Proposition 3.3] show that if X corresponds to a random sample from any absolutely continuous, centrally symmetric distribution in $\mathbb{R}^d, d > 2$, then $\varepsilon^*(T, X)$ converges to $1/3$ almost surely as $n \to \infty$. Furthermore, Donoho and Gasko [1992, Proposition 3.4] prove that $\varepsilon^*(T, X) \geq 1/(d+1)$ for any X in a general position (this means that no more that k points lie in any $(k-1)$-dimensional affine subspace, $k = 1, \ldots, d$). This corresponds to Theorem 26 above.

2. Scatter halfspace depth

In the first chapter of this thesis, we discussed the location halfspace depth. We showed that it allows us to order points with respect to some fixed distribution, and the deepest point can serve as a location estimator. This estimator possesses outstanding properties, especially those regarding robustness. However, in mathematical statistics, one is often interested in higher-order characteristics of probability distributions. A natural question is whether it is possible to use concepts similar to the ones discussed in the first chapter also in these situations. Therefore, in the second part of this thesis, we deal with scatter halfspace depth for matrices. We begin by defining the scatter parameter of a distribution.

Definition 27 (Scatter parameter) *Let P $\mathcal{P}(\mathsf{R}^d)$. Consider a functional $\mathbf{\Psi}\colon \mathcal{P} \to \mathsf{PD}_d$ satisfying $\mathbf{\Psi}(P_{\mathbf{A}X+b}) = \mathbf{A}\mathbf{\Psi}(P_X)\mathbf{A}^\mathsf{T}$ for any X P_X $\mathcal{P}$, $\mathbf{A}$ GL_d and b R^d such that $P_{\mathbf{A}X+b}$ $\mathcal{P}$. Then $\mathbf{\Psi}$ is called a* scatter functional *and $\mathbf{\Psi}(P)$ is called a* scatter parameter *of P $\mathcal{P}$. The uniquely determined matrix $\mathbf{\Psi}(P)^{1/2}$ PD_d is called a* scale parameter *of P.*

The functional assigning the covariance matrix to the appropriately integrable random vector is an example of a scatter functional. Another example (for $d = 1$) is the median squared deviation (2.1) defined below.

2.1 Axiomatic approach to scatter statistical depth

At the beginning of this work, we defined general location statistical depth. One of the desired properties stress that location depth should behave nicely with (centrally/angularly/halfspace) symmetric distributions. Analogously, scatter statistical depth should reflect the "shape" of the underlying distribution. Therefore, it should work well with so-called elliptically symmetric distributions, as defined below. Recall that by MSD we mean the *median squared deviation*, i.e.

$$\mathsf{MSD}(Z) = \mathsf{median}\left((Z - \mathsf{median}(Z))^2\right) \tag{2.1}$$

for any Z P $\mathcal{P}(\mathsf{R})$. In this definition, if the "median interval" is not a singleton, we always take its midpoint as the median (similarly as for the Tukey median).

Definition 28 (Elliptically symmetric distribution) *Consider a random vector X P $\mathcal{P}(\mathsf{R}^d)$. Then, we say that X (or its distribution P) is* spherically symmetric *about the origin if and only if $\mathbf{O}X \overset{D}{=} X$ for all $d \times d$ orthogonal matrices $\mathbf{O}$. More generally, the random vector X is said to have an* elliptically symmetric distribution *with location μ R^d and scatter $\mathbf{\Sigma}$ PD_d if and only if $X \overset{D}{=} \mu + \mathbf{\Sigma}^{1/2}Z$, where*

- *Z $\mathcal{P}(\mathsf{R}^d)$ is spherically symmetric about the origin,*
- *$\mathsf{MSD}(Z_1) = 1$, where Z_1 is the first component of Z.*

Note that the definition of a spherically symmetric distribution implies that $\mathsf{median}(Z_k^2) = 1$ for any $k \in \{1, \ldots, d\}$. As a consequence of this condition, it can be shown that the location $\boldsymbol{\mu}$ and the scatter matrix $\boldsymbol{\Sigma}$ of an elliptically symmetric distribution are uniquely determined. It is also easy to see that every elliptically symmetric distribution with location $\boldsymbol{\mu}$ is centrally symmetric about its center $\boldsymbol{\mu}$.

An example of a spherically symmetric (about the origin) distribution is the d-variate Gaussian distribution $N_d(\mathbf{0}, \alpha \mathbf{I}_d)$ for any $\alpha > 0$. Indeed, if a random variable Z has univariate standard Gaussian distribution, then we see that

$$\begin{aligned}
\mathsf{MSD}(Z) &= \mathsf{median}(Z^2) = \arg\max_{z>0} \left(\min\left\{\mathsf{P}(Z^2 \le z), \mathsf{P}(Z^2 \ge z)\right\}\right) \\
&= \arg\max_{z>0} \left(\min\left\{\Phi\left(\sqrt{z}\right) - \Phi\left(-\sqrt{z}\right), 1 - \Phi\left(\sqrt{z}\right) + \Phi\left(-\sqrt{z}\right)\right\}\right) \\
&= \arg\max_{z>0} \left(\min\left\{2\,\Phi\left(\sqrt{z}\right) - 1, 2 - 2\,\Phi\left(\sqrt{z}\right)\right\}\right) \\
&= \arg\max_{z>0} \left(\min\left\{\Phi\left(\sqrt{z}\right) - \frac{1}{2}, 1 - \Phi\left(\sqrt{z}\right)\right\}\right) = \left(\Phi^{-1}\left(\frac{3}{4}\right)\right)^2 =: \beta.
\end{aligned}$$

Therefore, $\mathsf{MSD}\left(Z/\sqrt{\beta}\right) = 1$. As a consequence, for all $\boldsymbol{\mu} \in \mathsf{R}^d$ and $\boldsymbol{\Sigma} \in \mathsf{PD}_d$, the Gaussian distribution $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}/\beta)$ is elliptically symmetric with location $\boldsymbol{\mu}$ and scatter $\boldsymbol{\Sigma}$. Note that β is defined such that it satisfies

$$\Phi\left(\sqrt{\beta}\right) = \frac{3}{4} \tag{2.2}$$

and this constant will be used in the following text.

The subsequent text deals with the space PD_d of all symmetric positive definite $d \times d$ matrices. Therefore, it is necessary to introduce a topology in this space. We will consider the Frobenius topology induced by the Frobenius distance, which is itself inherited from the Frobenius norm, i.e.

$$d_F(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) := \|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_{\mathsf{F}} = \sqrt{\mathsf{tr}\left((\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)^{\mathsf{T}}\right)}.$$

A real function $f\colon \mathsf{PD}_d \to [0, \infty)$ will be called F-(semi)continuous if it is (semi)continuous w.r.t. the Frobenius topology on PD_d. Similarly, a set $A \subset \mathsf{PD}_d$ is called F-(bounded/closed) if it is bounded/closed w.r.t. the Frobenius topology on PD_d. We also use the operator matrix norm in Section 3.5 about the minimax optimality of the scatter halfspace median. This norm (also called the spectral norm) is defined for any $d \times d$ symmetric matrix $\mathbf{A}$ as

$$\|\mathbf{A}\|_{\mathsf{op}} = \sup_{u \in \mathsf{S}^{d-1}} \|\mathbf{A}u\|,$$

which is equal to the square root of the largest singular value of $\mathbf{A}$. Because $\mathbf{A}$ is symmetric, $\|\mathbf{A}\|_{\mathsf{op}}$ is equal to

$$\sqrt{\max\{|\lambda_1|, \ldots, |\lambda_d|\}}$$

where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $\mathbf{A}$.

As Zuo and Serfling [2000a] define general location depth, Paindaveine and Van Bever [2018] similarly suggest the following scatter statistical depth definition.

Definition 29 (Scatter statistical depth) *A mapping $D^{\mathrm{sc}}\colon \mathsf{PD}_d \times P(\mathsf{R}^d) \to [0, \infty)$ is called a* scatter statistical depth *(or simply* scatter depth*) if it is bounded and satisfies the following properties:*

$(B1)$ *For any random vector $X \sim P_X \in P(\mathsf{R}^d)$, $\boldsymbol{\Sigma} \in \mathsf{PD}_d$, $\mathbf{A} \in \mathsf{GL}_d$ and $b \in \mathsf{R}^d$ it holds that $D^{\mathrm{sc}}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top; P_{\mathbf{A}X+b}) = D^{\mathrm{sc}}(\boldsymbol{\Sigma}; P_X)$.*

$(B2)$ $D^{\mathrm{sc}}(\boldsymbol{\Sigma}_0; P) = \max_{\boldsymbol{\Sigma} \in \mathsf{PD}_d} D^{\mathrm{sc}}(\boldsymbol{\Sigma}; P)$ *holds true for any elliptically symmetric distribution $P \in P(\mathsf{R}^d)$ with scatter $\boldsymbol{\Sigma}_0$.*

$(B3)$ *If for any fixed $P \in P(\mathsf{R}^d)$ a matrix $\boldsymbol{\Sigma}_0 \in \mathsf{PD}_d$ maximizes $D^{\mathrm{sc}}(\cdot; P)$, then for an arbitrary $\boldsymbol{\Sigma} \in \mathsf{PD}_d$ the mapping $\alpha \to D^{\mathrm{sc}}(\boldsymbol{\Sigma}_0 + \alpha(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0); P)$ is non-increasing on $[0, 1]$,*

$(B4)$ *Consider any fixed $P \in P(\mathsf{R}^d)$ and $\{\boldsymbol{\Sigma}_n\}_{n=1}^{\infty} \subset \mathsf{PD}_d$ such that either $d_F(\boldsymbol{\Sigma}_n, \boldsymbol{\Sigma}) \xrightarrow[n\to\infty]{} 0$ for some singular $d \times d$ matrix $\boldsymbol{\Sigma}$, or $d_F(\boldsymbol{\Sigma}_n, \mathbf{I}_d) \xrightarrow[n\to\infty]{} \infty$. Then $D^{\mathrm{sc}}(\boldsymbol{\Sigma}_n; P) \xrightarrow[n\to\infty]{} 0$.*

Properties $(B1)-(B4)$ are somewhat analogous to those for location statistical depth. Namely, $(B1)$ states that depth should transform accordingly after the affine transformation of the underlying distribution. Property $(B2)$ expresses that the scatter estimator based on maximal depth should be Fisher-consistent for elliptically symmetric distributions. Properties $(B3)$ and $(B4)$ claim that scatter depth should possess some monotonicity properties, and the depth of matrices near the "boundary" of the space PD_d should be close to zero.

Again, we can define the sample version of the scatter depth D^{sc} based on a random sample $X_1, \ldots, X_n \sim P \in P(\mathsf{R}^d)$ as $D^{\mathrm{sc}}(\boldsymbol{\Sigma}; \widehat{P}_n)$ and a matrix maximizing this function can be used as an estimator of the scatter parameter of P.

2.2 Definition of scatter halfspace depth

The scatter halfspace depth was first introduced in [Chen et al., 2018]. There, it was studied only for centred Gaussian distributions with fixed location $\mathbf{0}$. The following general definition is provided in [Paindaveine and Van Bever, 2018].

Definition 30 (Scatter halfspace depth) *Let $X \sim P \in P(\mathsf{R}^d)$ and $\boldsymbol{\Sigma} \in \mathsf{PD}_d$. Further, consider some location functional $\boldsymbol{\Xi}$. The* scatter halfspace depth *of $\boldsymbol{\Sigma}$ w.r.t. P is defined as*

$$hD^{\mathrm{sc}}_{\boldsymbol{\Xi}}(\boldsymbol{\Sigma}; P) := \inf_{u \in \mathsf{S}^{d-1}} \min\left\{\mathsf{P}\left(|\langle X - \boldsymbol{\Xi}(P), u\rangle| \le \sqrt{u^\top \boldsymbol{\Sigma} u}\right), \mathsf{P}\left(|\langle X - \boldsymbol{\Xi}(P), u\rangle| \ge \sqrt{u^\top \boldsymbol{\Sigma} u}\right)\right\}.$$

To interpret the scatter halfspace depth geometrically, we introduce the subsequent definition.

Definition 31 *For* $u \in \mathbb{S}^{d-1}$, $t \geq 0$ *and* $\Sigma \in \mathrm{PD}_d$, *we define*

$$\begin{aligned} H^{\mathrm{in}}_{u,t} &= \left\{x \in \mathbb{R}^d \mid |\langle x, u\rangle| \leq t\right\}, \\ H^{\mathrm{out}}_{u,t} &= \left\{x \in \mathbb{R}^d \mid |\langle x, u\rangle| \geq t\right\}, \\ H^{\mathrm{in}}_{u,\Sigma} &= \left\{x \in \mathbb{R}^d \mid |\langle x, u\rangle| \leq \sqrt{u^\top \Sigma u}\right\}, \\ H^{\mathrm{out}}_{u,\Sigma} &= \left\{x \in \mathbb{R}^d \mid |\langle x, u\rangle| \geq \sqrt{u^\top \Sigma u}\right\}. \end{aligned} \tag{2.3}$$

Using this notion, we can rewrite

$$hD^{\mathrm{sc}}_{\Xi}(\Sigma; P) = \inf_{u \in \mathbb{S}^{d-1}} \min\left\{P\left(H^{\mathrm{in}}_{u,\Sigma} + \Xi(P)\right), P\left(H^{\mathrm{out}}_{u,\Sigma} + \Xi(P)\right)\right\}, \tag{2.4}$$

where $A + z = \{x + z \mid x \in A\}$ for any $A \subseteq \mathbb{R}^d$ and $z \in \mathbb{R}^d$. Denote by E_Σ the surface

$$\Sigma^{1/2}\mathbb{S}^{d-1} = \left\{\Sigma^{1/2}u \in \mathbb{R}^d \mid u \in \mathbb{S}^{d-1}\right\}$$

of an ellipsoid $\Sigma^{1/2}B_d(\mathbf{0}, 1)$. Then, it can be shown (see [Brabenec, 2021, Section 1.3] for reference) that the set $H^{\mathrm{in}}_{u,\Sigma}$ consists exactly of all points in between two supporting hyperplanes of $\Sigma^{1/2}B_d(\mathbf{0}, 1)$ with normal vector u. Similarly, $H^{\mathrm{out}}_{u,\Sigma}$ consists of all points which are not in between these two hyperplanes. Our situation is shown in Figure 4. From (2.4) we can see that if the matrix Σ has a small scatter halfspace depth w.r.t. P, then in some direction $u \in \mathbb{S}^{d-1}$, at least one of the sets $H^{\mathrm{in}}_{u,\Sigma} + \Xi(P), H^{\mathrm{out}}_{u,\Sigma} + \Xi(P)$ has small P-probability. Therefore, when looking for a matrix with a high scatter halfspace depth w.r.t. P, one has to find Σ such that in every direction $u \in \mathbb{S}^{d-1}$, both $H^{\mathrm{in}}_{u,\Sigma} + \Xi(P)$ and $H^{\mathrm{out}}_{u,\Sigma} + \Xi(P)$ have high probability P. This is the reason why the scatter halfspace depth captures the geometric characteristics of the elliptically symmetric distribution well. For a smooth P, $hD^{\mathrm{sc}}_{\Xi}(\cdot; P)$ is obviously bounded from above by $1/2$.

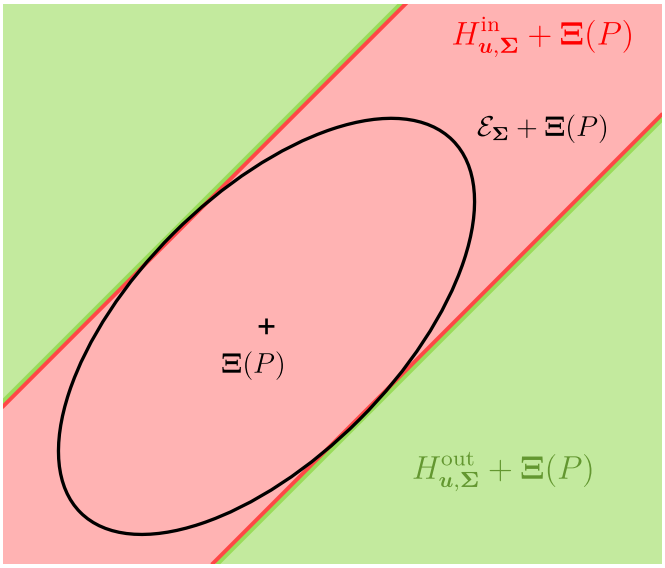


Figure 4: Interpretation of the definition of the scatter halfspace depth of Σ w.r.t. the distribution P: When determining the depth of Σ, we are looking for a direction $u \in \mathbb{S}^{d-1}$ in which at least one of the sets $H^{\mathrm{in}}_{u,\Sigma} + \Xi(P)$ and $H^{\mathrm{out}}_{u,\Sigma} + \Xi(P)$ has the smallest probability P.

For a better understanding, consider the univariate case $d = 1$. Then, for a fixed probability distribution $P \in \mathcal{P}(\mathbb{R})$, $X \sim P$ and $\sigma^2 > 0$, we have

$$hD^{sc}_{\Xi}(\sigma^2; P) = \min\left\{\mathsf{P}\left((X - \Xi(P))^2 \leq \sigma^2\right), \mathsf{P}\left((X - \Xi(P))^2 \geq \sigma^2\right)\right\}. \quad (2.5)$$

Note that if Ξ corresponds to the univariate median, then the point σ^2, which maximizes (2.5), is precisely the median squared deviation (2.1). Further, we derive what the scatter halfspace depth of a matrix looks like with respect to a Gaussian distribution.

Example 6 *Let $X \sim P_X \sim N_d(\mu, \Sigma)$ for an arbitrary $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathrm{PD}_d$. First, note that if Ξ is a location functional, then it must hold that $\Xi(P_X) = \mu$. Indeed, by the definition of the location functional, we have for all $\mu \in \mathbb{R}^d$ that*

$$\Xi(P_X) = -\Xi(P_{-X}) = -\Xi(P_{X-2\mu}) = -\Xi(P_X) + 2\mu,$$

which implies $\Xi(P_X) = \mu$. Therefore, $X - \Xi(P_X) \sim N_d(\mathbf{0}, \Sigma)$ and

$$X - \Xi(P_X), u \sim N_d\left(\mathbf{0}, u^\top \Sigma u\right).$$

Consider some $\Gamma \in \mathrm{PD}_d$. Then, we can write

$$\mathsf{P}\left(|\, X - \Xi(P_X), u \,| \leq \sqrt{u^\top \Gamma u}\right) = \mathsf{P}\left(\left|\frac{X - \Xi(P_X), u}{\sqrt{u^\top \Sigma u}}\right| \leq \sqrt{\frac{u^\top \Gamma u}{u^\top \Sigma u}}\right)$$
$$= 2\,\Phi\left(\sqrt{\frac{u^\top \Gamma u}{u^\top \Sigma u}}\right) - 1.$$

Similarly,

$$\mathsf{P}\left(|\, X - \Xi(P_X), u \,| \geq \sqrt{u^\top \Gamma u}\right) = 2\left(1 - \Phi\left(\sqrt{\frac{u^\top \Gamma u}{u^\top \Sigma u}}\right)\right).$$

As a consequence, we have

$$hD^{sc}_{\Xi}(\Gamma; P_X) = 2 \inf_{u \in \mathbb{S}^{d-1}} \min\left\{\Phi\left(\sqrt{\frac{u^\top \Gamma u}{u^\top \Sigma u}}\right) - \frac{1}{2}, 1 - \Phi\left(\sqrt{\frac{u^\top \Gamma u}{u^\top \Sigma u}}\right)\right\}$$
$$= 2 \min\left\{\Phi\left(\sqrt{\inf_{u \in \mathbb{S}^{d-1}} \frac{u^\top \Gamma u}{u^\top \Sigma u}}\right) - \frac{1}{2}, 1 - \Phi\left(\sqrt{\sup_{u \in \mathbb{S}^{d-1}} \frac{u^\top \Gamma u}{u^\top \Sigma u}}\right)\right\}.$$

Now, note that

$$\inf_{u \in \mathbb{S}^{d-1}} \frac{u^\top \Gamma u}{u^\top \Sigma u} = \inf_{u \neq 0} \frac{u^\top \Gamma u}{u^\top \Sigma u} \overset{v = \Sigma^{1/2} u}{=} \inf_{v \neq 0} \frac{v^\top \Sigma^{-\frac{1}{2}} \Gamma \Sigma^{-\frac{1}{2}} v}{\|v\|^2} = \inf_{v \in \mathbb{S}^{d-1}} v^\top \Sigma^{-\frac{1}{2}} \Gamma \Sigma^{-\frac{1}{2}} v.$$

By the Cauchy-Schwarz inequality, this is equal to the smallest eigenvalue of $\Sigma^{-\frac{1}{2}} \Gamma \Sigma^{-\frac{1}{2}}$ denoted by $\sigma_{min}(\Sigma^{-\frac{1}{2}} \Gamma \Sigma^{-\frac{1}{2}})$. Analogically

$$\sup_{u \in \mathbb{S}^{d-1}} \frac{u^\top \Gamma u}{u^\top \Sigma u} = \sigma_{max}(\Sigma^{-\frac{1}{2}} \Gamma \Sigma^{-\frac{1}{2}}).$$

Therefore, we can write

$$hD_{\Xi}^{\mathrm{sc}}(\mathbf{\Gamma}; P_X) = 2\min\left\{\Phi\left(\sqrt{\sigma_{min}(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Gamma}\mathbf{\Sigma}^{-\frac{1}{2}})}\right) - \frac{1}{2}, 1 - \Phi\left(\sqrt{\sigma_{max}(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Gamma}\mathbf{\Sigma}^{-\frac{1}{2}})}\right)\right\}.$$

This is obviously maximized for $\mathbf{\Gamma}$ satisfying

$$\Phi\left(\sqrt{\sigma_{min}(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Gamma}\mathbf{\Sigma}^{-\frac{1}{2}})}\right) = \Phi\left(\sqrt{\sigma_{max}(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Gamma}\mathbf{\Sigma}^{-\frac{1}{2}})}\right) = \frac{3}{4}$$

which implies

$$\sigma_{min}(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Gamma}\mathbf{\Sigma}^{-\frac{1}{2}}) = \sigma_{max}(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Gamma}\mathbf{\Sigma}^{-\frac{1}{2}}) = \beta, \tag{2.6}$$

where β is defined in (2.2)*. Note that $\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Gamma}\mathbf{\Sigma}^{-\frac{1}{2}}$ PD_d, therefore it is diagonalizable. From* (2.6) *we have*

$$\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Gamma}\mathbf{\Sigma}^{-\frac{1}{2}} = \mathbf{U}^{\mathsf{T}}\mathbf{\Lambda}\mathbf{U} = \beta\,\mathbf{I}_d,$$

where $\mathbf{U}$ is an orthogonal matrix and $\mathbf{\Lambda}$ is a diagonal matrix with all diagonal entries equal to β. This implies that

$$\mathbf{\Gamma} = \beta\mathbf{\Sigma},$$

which is the scatter parameter of P_X. To sum up, the scatter halfspace depth w.r.t. the Gaussian distribution with a covariance matrix $\mathbf{\Sigma}$ is uniquely maximized for its scatter parameter $\beta\mathbf{\Sigma}$. This will hold for any elliptically symmetric distribution, as stated in Theorem 36.

2.3 Basic properties of scatter halfspace depth

In this section, we introduce basic properties of the scatter halfspace depth. Similarly as in Chapter 1.3, we do not provide proofs, only references. This is because the purpose of this work is not to prove basic properties; rather, the work focuses mainly on minimax optimality of the scatter halfspace depth.

One of the most important properties of the scatter halfspace depth is that it is affine invariant, as stated in the following theorem.

Theorem 32 *The scatter halfspace depth is a ne invariant for any location functional Ξ. That is, for any X P_X $P(\mathrm{R}^d)$, $\mathbf{\Sigma}$ PD_d, $\mathbf{A}$ GL_d and b R^d, it holds that*

$$hD_{\Xi}^{\mathrm{sc}}(\mathbf{A}\mathbf{\Sigma}\mathbf{A}^{\mathsf{T}}; P_{\mathbf{A}X+b}) = hD_{\Xi}^{\mathrm{sc}}(\mathbf{\Sigma}; P_X).$$

Recall that by $P_{\mathbf{A}X+b}$ we denote the distribution of $\mathbf{A}X + b$.

Proof. Refer to [Paindaveine and Van Bever, 2018, Theorem 2.1].

This property is useful because it implies that the scatter halfspace depth does not depend on our choice of the coordinate system. Next, similarly as for the location halfspace depth, we define a scatter α-depth region.

Definition 33 (Scatter α-depth region) *For any $P \in \mathcal{P}(\mathbb{R}^d)$ and $\alpha \in [0,1]$, the* scatter α-depth region *of P is defined as*

$$D^{\mathrm{sc}}(\alpha; P) := \{\Sigma \in \mathrm{PD}_d \mid hD^{\mathrm{sc}}_{\Xi}(\Sigma; P) \geq \alpha\} \subset \mathrm{PD}_d.$$

An easy consequence of Theorem 32 is that if $X \sim P_X \in \mathcal{P}(\mathbb{R}^d)$, $\mathbf{A} \in \mathrm{GL}_d$ and $b \in \mathbb{R}^d$, then

$$D^{\mathrm{sc}}(\alpha; P_{\mathbf{A}X+b}) = \mathbf{A} D^{\mathrm{sc}}(\alpha; P_X) \mathbf{A}^\top$$

holds for any $\alpha \in [0,1]$. The subsequent theorem shows that the scatter halfspace depth is also quasiconcave. This is an analogous statement to Theorem 7, which establishes quasiconcavity of the location halfspace depth.

Theorem 34 *For any $P \in \mathcal{P}(\mathbb{R}^d)$, the function $hD^{\mathrm{sc}}(\cdot; P)$ is quasiconcave. That is, for any $\Sigma_1, \Sigma_2 \in \mathrm{PD}_d$ and $\alpha \in [0,1]$ we have*

$$hD^{\mathrm{sc}}_{\Xi}(\alpha\Sigma_1 + (1-\alpha)\Sigma_2; P) \geq \min\left\{hD^{\mathrm{sc}}_{\Xi}(\Sigma_1; P), hD^{\mathrm{sc}}_{\Xi}(\Sigma_2; P)\right\}.$$

Proof. Refer to [Paindaveine and Van Bever, 2018, Theorem 3.3].

This theorem implies that $D^{\mathrm{sc}}(\alpha; P)$ is always a convex subset of PD_d, which can be shown just like in (1.2) for location halfspace depth regions $D(\alpha; P)$. Further, we are interested in whether the scatter halfspace depth is upper F-semicontinuous or even F-continuous.

Theorem 35 *For any $P \in \mathcal{P}(\mathbb{R}^d)$, the function $hD^{\mathrm{sc}}(\cdot; P)$ is always upper F-semicontinuous. If P is smooth at $\Xi(P)$, then $hD^{\mathrm{sc}}(\cdot; P)$ is also F-continuous.*

Proof. Refer to [Paindaveine and Van Bever, 2018, Theorem 3.1].

The previous theorem implies that the sets $D^{\mathrm{sc}}(\alpha; P), \alpha \in [0,1]$, are always F-closed. Paindaveine and Van Bever [2018, Theorem 3.2] show that for $\alpha > 0$, these sets are also F-bounded. The authors of that paper claim that all bounded sets are also totally bounded in (PD_d, d_F). Therefore, we have that for $\alpha > 0$, the set $D^{\mathrm{sc}}(\alpha; P)$ is F-totally bounded and F-closed. But the space (PD_d, d_F) is not complete, hence we can not conclude that $D^{\mathrm{sc}}(\alpha; P)$ is F-compact, in contrast with the case of the location halfspace depth. Therefore, Paindaveine and Van Bever [2018] also analyse properties of the scatter halfspace depth w.r.t. the so-called geodesic topology, see [Paindaveine and Van Bever, 2018, Section 4]. This is used, for example, to prove Theorem 40 stated below.

Similarly as for the location halfspace depth, arguably the most important property is that under certain conditions, the scatter halfspace depth is a scatter statistical depth in the sense of Definition 29.

Theorem 36 *For the scatter halfspace depth, properties $(B1)$, $(B2)$ and $(B3)$ of Definition 29 are always met. Property $(B4)$ is satisfied for P smooth at $\Xi(P)$.*

Proof. Property $(B1)$ is implied by Theorem 32. For the proof of property $(B2)$, refer to [Paindaveine and Van Bever, 2018, Theorem 5.1]. Property $(B3)$ is implied by Theorem 34, see [Paindaveine and Van Bever, 2018, Section 5] for details. Finally, the property $(B4)$ is proven in [Brabenec, 2021, Theorem 16 and Theorem 19].

2.4 Sample scatter halfspace depth and scatter halfspace median

Analogously as for the location halfspace depth, we estimate the scatter halfspace depth using the sample scatter halfspace depth, which is defined as follows.

Definition 37 (Sample scatter halfspace depth) *Consider a random sample* $\{X_i\}_{i=1}^n$ P $P(\mathsf{R}^d)$, $\boldsymbol{\Sigma}$ PD_d *and some location functional* $\boldsymbol{\Xi}$. *The* sample scatter halfspace depth *of* $\boldsymbol{\Sigma}$ *w.r.t.* $\{X_i\}_{i=1}^n$ *is defined as*

$$hD_{\boldsymbol{\Xi}}^{\mathsf{SC}}(\boldsymbol{\Sigma}; \{X_i\}_{i=1}^n) \coloneqq hD_{\boldsymbol{\Xi}}^{\mathsf{SC}}(\boldsymbol{\Sigma}; \widehat{P}_n)$$
$$= \inf_{u \in \mathsf{S}^{d-1}} \min\left\{ \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{|\, x_i - \widehat{\theta}_n, u \,| \leq \sqrt{u^\top \boldsymbol{\Sigma} u}\}}, \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{|\, x_i - \widehat{\theta}_n, u \,| \geq \sqrt{u^\top \boldsymbol{\Sigma} u}\}} \right\},$$

where $\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\Xi}(\widehat{P}_n)$.

It turns out that $hD_{\boldsymbol{\Xi}}^{\mathsf{SC}}(\boldsymbol{\Sigma}; \{X_i\}_{i=1}^n)$ is (under a certain condition for $\boldsymbol{\Xi}$) a strongly uniformly consistent estimator of $hD_{\boldsymbol{\Xi}}^{\mathsf{SC}}(\boldsymbol{\Sigma}; P)$. This claim is stated in the subsequent theorem.

Theorem 38 *Suppose we have a random sample* $\{X_i\}_{i=1}^n$ P $P(\mathsf{R}^d)$. *If* $\widehat{\boldsymbol{\theta}}_n \xrightarrow[n\to\infty]{a.s.} \boldsymbol{\Xi}(P)$, *then*

$$\sup_{\boldsymbol{\Sigma} \in \mathsf{PD}_d} |hD_{\boldsymbol{\Xi}}^{\mathsf{SC}}(\boldsymbol{\Sigma}; \{X_i\}_{i=1}^n) - hD_{\boldsymbol{\Xi}}^{\mathsf{SC}}(\boldsymbol{\Sigma}; P)| \xrightarrow[n\to\infty]{a.s.} 0. \tag{2.7}$$

Proof. Refer to [Paindaveine and Van Bever, 2018, Theorem 2.2].

A natural choice of $\boldsymbol{\Xi}$ is the Tukey median functional T. Then, by Theorem 18 it holds that $T(\widehat{P}_n) \xrightarrow[n\to\infty]{a.s.} T(P)$. As a consequence, $hD_T^{\mathsf{SC}}(\cdot; \{X_i\}_{i=1}^n)$ is always a strongly uniformly consistent estimator of $hD_T^{\mathsf{SC}}(\cdot; P)$. In Chapter 3, we consider a model consisting of centered Gaussian distributions with an unknown covariance matrix. In that situation, we can use a constant location functional $\mathbf{0}$, which is obviously consistent. Consequently, in this case, the convergence (2.7) holds true.

Further, we define the maximal scatter halfspace depth and the set of matrices of maximal scatter depth as follows.

Definition 39 (Maximal scatter depth) *For* P $P(\mathsf{R}^d)$, *we define the* maximal scatter halfspace depth *as*

$$\Pi^{\mathsf{SC}}(P) \coloneqq \sup_{\boldsymbol{\Sigma} \in \mathsf{PD}_d} hD_{\boldsymbol{\Xi}}^{\mathsf{SC}}(\boldsymbol{\Sigma}; P)$$

and the set of matrices of maximal scatter halfspace depth as

$$\mathsf{Med}^{sc}(P) := D^{\mathsf{SC}}(\Pi^{\mathsf{SC}}(P); P).$$

For the location halfspace depth, there is always a point with maximal location halfspace depth. This is not the case with the scatter halfspace depth, as noted in [Brabenec, 2021, Section 2.2]. Yet, under the assumption that P is smooth at its location $\Xi(P)$, the deepest matrix always exists. This is stated in the following theorem.

Theorem 40 *Let P $P(\mathsf{R}^d)$. If Ξ is a location functional such that P is smooth at $\Xi(P)$, then there exists Σ PD_d such that*

$$hD^{\mathsf{SC}}_{\Xi}(\Sigma; P) = \Pi^{\mathsf{SC}}(P),$$

therefore $\mathsf{Med}^{sc}(P) =$.

Proof. Refer to [Paindaveine and Van Bever, 2018, Theorem 4.3].

Similarly as for the deepest points of the location halfspace depth, we can have several deepest matrices of the scatter halfspace depth. Paindaveine and Van Bever [2018, Theorem 5.1] show that under some mild assumptions, the scatter halfspace depth w.r.t. an elliptically symmetric distribution is uniquely maximized for its scatter parameter.

Definition 41 (Scatter halfspace median) *Let P $P(\mathsf{R}^d)$. Any matrix Σ PD_d that maximizes $hD^{\mathsf{SC}}_{\Xi}(\cdot; P)$ is called a* scatter halfspace median[1] *of P, denoted by $\Sigma^{\mathsf{SC}}(P)$.*

To conclude this section, let us note that a *sample scatter halfspace median*

$$\widehat{\Sigma}^{\mathsf{SC}}_n := \operatorname*{arg\,max}_{\Sigma\ \mathsf{PD}_d} hD^{\mathsf{SC}}_{\Xi}(\Sigma; \{X_i\}^n_{i=1})$$

always exists, because the sample scatter halfspace depth attains only finitely many values.

2.5 Robustness of scatter halfspace median

Since the concept of the scatter halfspace depth and the scatter halfspace median matrix is relatively new compared to the location halfspace depth, not much is known about its robustness properties so far. In this section, we briefly present one of the results of Louvet and Van Bever [2024]. That article primarily discusses the influence function of the scatter halfspace depth. Only in Section 4 of that article, the influence function of the scatter halfspace median is derived under certain conditions, as presented further. The following theorem can be found in [Louvet and Van Bever, 2024, Theorem 5].

[1] If we only need one representative of this set of matrices, we can, for example, choose its centroid, similarly as for the location halfspace depth.

Theorem 42 *Let $P \in \mathcal{P}(\mathbb{R}^d)$ correspond to the elliptically symmetric distribution with location $\boldsymbol{\mu} = \mathbf{0}$ and scatter $\boldsymbol{\Sigma}_0 \in \mathrm{PD}_d$. That is, for $X \sim P$, the canonical representation is $X = \boldsymbol{\Sigma}_0^{1/2} Z$, where $Z \in \mathcal{P}(\mathbb{R}^d)$ is spherically symmetric about the origin. Denote the cumulative distribution function of Z_1 by F and assume that this distribution is absolutely continuous with density $f = F'$. Consider a functional $\Xi \not\equiv \mathbf{0}$ and let $f(1) > 0$. Then, $\boldsymbol{\Sigma}_0$ is the unique scatter halfspace median of P, i.e. the unique maximizer of $hD^{\mathrm{sc}}_{\Xi}(\cdot; P)$. Furthermore, let $x \in \mathbb{R}^d$ and $\varepsilon \in (0, 1)$.*

- *If $|\langle u, x \rangle| < \sqrt{u^\top \boldsymbol{\Sigma}_0 u}$ for all $u \in \mathbb{S}^{d-1}$, then*

$$\operatorname*{arg\,max}_{\boldsymbol{\Sigma} \in \mathrm{PD}_d} hD^{\mathrm{sc}}_{\Xi}(\boldsymbol{\Sigma}; P_{(\varepsilon,x)}) = \left(F^{-1}\left(\frac{3 - 4\varepsilon}{4 - 4\varepsilon}\right)\right)^2 \boldsymbol{\Sigma}_0.$$

  As a consequence,

$$IF(x; \boldsymbol{\Sigma}^{\mathrm{sc}}, P) := \lim_{\varepsilon \to 0^+} \frac{\boldsymbol{\Sigma}^{\mathrm{sc}}(P_{(\varepsilon,x)}) - \boldsymbol{\Sigma}^{\mathrm{sc}}(P)}{\varepsilon} = \frac{-\boldsymbol{\Sigma}_0}{2f(1)}.$$

- *If $|\langle u, x \rangle| \geq \sqrt{u^\top \boldsymbol{\Sigma}_0 u}$ for some $u \in \mathbb{S}^{d-1}$, then*

$$\sup_{\boldsymbol{\Sigma} \in \mathrm{PD}_d} hD^{\mathrm{sc}}_{\Xi}(\boldsymbol{\Sigma}; P_{(\varepsilon,x)}) = hD^{\mathrm{sc}}_{\Xi}(\boldsymbol{\Sigma}_0; P_{(\varepsilon,x)}) = \frac{1 - \varepsilon}{2}.$$

Proof. Refer to [Louvet and Van Bever, 2024, Theorem 5].

This theorem assumes for simplicity that the distribution P is centered at the origin. Thus, we do not need to estimate the location of the distribution. Furthermore, the assumption that $f(1) > 0$ ensures that there is exactly one deepest matrix (scatter halfspace median) w.r.t. P, which is essential for us to be able to say anything about the influence function. If there were multiple deepest matrices, the situation would become complicated. From Theorem 42, we can see that if the contaminating point is inside the ellipsoid $\boldsymbol{\Sigma}_0^{1/2} B_d(\mathbf{0}, 1)$, then the only deepest matrix will still be $\boldsymbol{\Sigma}_0$, just rescaled by an appropriate constant depending on the level of contamination ε. However, if the point is outside that ellipsoid, then the deepest matrix is still $\boldsymbol{\Sigma}_0$, but it may not be the only one. This again leads to a complicated situation because to determine the influence function in this case, we would need to be able to select exactly one representative of those deepest matrices.

To summarize, from Theorem 42, we can see that the influence function is bounded (and constant), at least for contamination by points close to the center of symmetry. This, however, indicates good robust properties only partially.

3. Minimax optimality

3.1 Convergence rate

Consider a general parametric statistical model $M = \{P_\theta \mid \theta \in \Theta\} \subseteq \mathcal{P}(\mathbb{R}^d)$, where Θ is some parametric space. Our task is to estimate the parameter θ based on a random sample $X_1, \ldots, X_n \sim P_\theta$. To do so, we construct an estimator $\widehat{\theta}_n = \Xi(\widehat{P}_n)$ using some statistical functional Ξ. Besides properties of $\widehat{\theta}_n$ such as consistency or unbiasedness, we are often interested in how fast the estimator $\widehat{\theta}_n$ converges to θ.

Example 7 *Let $M = \left\{N_d(\theta, \mathbf{I}_d) \mid \theta \in \mathbb{R}^d\right\}$ be the statistical model of interest. We consider the sample mean $\widehat{\theta}_n = \overline{X}_n$. Then $\sqrt{n}(\theta - \widehat{\theta}_n)$ is distributed as $N_d(\mathbf{0}, \mathbf{I}_d)$ and consequently $n\left\|\theta - \widehat{\theta}_n\right\|^2 \sim \chi^2_d$. Note that $\mathsf{E}\left(n\left\|\theta - \widehat{\theta}_n\right\|^2\right) = d$. Utilizing Markov's inequality, we have that for any $a > 0$, it holds that*

$$\mathsf{P}\left(n\left\|\theta - \widehat{\theta}_n\right\|^2 \geq a\right) \leq \frac{d}{a} \quad \forall n \in \mathbb{N}.$$

Let $0 < c < 1$ and consider $a = d/c$. Rewriting the inequality above, we get

$$\mathsf{P}\left(n\left\|\theta - \widehat{\theta}_n\right\|^2 \geq \frac{d}{c}\right) \leq c.$$

As a consequence, for all $n, d \in \mathbb{N}$ it holds that

$$\mathsf{P}\left(\left\|\theta - \widehat{\theta}_n\right\|^2 < \frac{1}{c}\frac{d}{n}\right) \geq 1 - c. \tag{3.1}$$

Inequality (3.1) indicates that we need twice as many observations to obtain twice as accurate (in terms of the squared Euclidean norm) estimate with the same probability. On the other hand, the higher the dimension, the higher the number of observations required to obtain equally precise estimates.

The quantity d/n (considered as a function of n) in the example above is the so-called *convergence rate*. This depends on our choice of measuring the accuracy of estimates. Generally, this is done using a *loss function*.

Definition 43 (Loss function) *A* loss function *is any measurable map*

$$L\colon \Theta \times \Theta \to [0, \infty).$$

*We consider a symmetric loss function such that $L(\theta_1, \theta_2) = 0$ if and only if $\theta_1 = \theta_2$, and we suppose that it satisfies the τ-*triangle inequality *(sometimes also called weak triangle inequality), i.e.*

$$L(\theta_1, \theta_2) \leq \tau\,(L(\theta_1, \theta_3) + L(\theta_3, \theta_2)) \tag{3.2}$$

for some $\tau \geq 1$ and all $\theta_1, \theta_2, \theta_3 \in \Theta$.

A loss function is usually chosen as some metric or its square. If $L(\cdot,\cdot) = d(\cdot,\cdot)^2$ for some metric d on Θ, then (3.2) is satisfied for $\tau = 2$. To see this, note that the following inequality holds

$$d(\theta_1,\theta_2)^2 \leq (d(\theta_1,\theta_3) + d(\theta_3,\theta_2))^2 \leq 2\left(d(\theta_1,\theta_3)^2 + d(\theta_3,\theta_2)^2\right). \tag{3.3}$$

In this chapter, we drop the indices in sequences $\{a_n\}_{n=1}^{\infty}$, i.e., we write only $\{a_n\}$. By the term *rate*, we mean a positive real sequence. All rates will be indexed by n, the random sample size. To define the convergence rate, we consider the following notation.

Definition 44 (Order of rates) *Consider two rates $\{a_n\}, \{b_n\}$. We say that $\{a_n\}$ is* of lower order *than $\{b_n\}$ (writing $\{a_n\} \prec \{b_n\}$) if there exists an absolute constant $C > 0$ such that $a_n \leq C\, b_n$ for all $n \in \mathbb{N}$. We say that $\{a_n\}$ is* of the same order *as $\{b_n\}$ if $\{a_n\} \prec \{b_n\}$ and $\{b_n\} \prec \{a_n\}$. This is denoted by $\{a_n\} \asymp \{b_n\}$.*

Now, we are ready to define the convergence rate of an estimator.

Definition 45 (Convergence rate) *Consider an estimator $\widehat{\theta}_n$, a loss function L and a rate $\{R_n\}$. We say that $\{R_n\}$ is the* convergence rate *of the estimator $\widehat{\theta}_n$ in terms of the loss L if:*

1. *For any $\delta > 0$ there exists an absolute constant $C > 0$ such that*

$$\sup_{n \in \mathbb{N}} \sup_{\theta \in \Theta} P_\theta\left(L(\widehat{\theta}_n, \theta) \geq C R_n\right) \leq \delta, \tag{3.4}$$

   where the estimator $\widehat{\theta}_n$ is based on a random sample $X_1, \ldots, X_n \sim P_\theta$.

2. *If the property (3.4) above is satisfied for some other rate $\{R_n\}$, then $\{R_n\} \prec \{R_n\}$.*

Note *We use $C, D, C', \ldots$ to represent universal constants. Note that these constants may vary in different sections and theorems.*

This work considers convergence rates dependent on the sample size n, the dimension d, and the proportion of contamination ε (this will be explained in Section 3.3). Therefore, in Definition 45, by C we mean a constant not depending on n, d and ε. To summarize the definition above, a rate $\{R_n\}$ is said to be the convergence rate of an estimator $\widehat{\theta}_n$ in terms of some loss L if $\{R_n\}$ is a rate of the lowest order such that $\left\{L(\widehat{\theta}_n, \theta)/R_n\right\}$ is bounded in probability (or tight) uniformly over all $\theta \in \Theta$ and $n \in \mathbb{N}$. From the definition, we can see that a convergence rate is not uniquely defined. This is summarized in the following theorem.

Theorem 46 *Let $\widehat{\theta}_n$ be an estimator and L be a loss function.*

1. *If both $\{R_n\}$ and $\{Q_n\}$ are convergence rates of $\widehat{\theta}_n$ in terms of the loss L, then $\{R_n\} \asymp \{Q_n\}$.*

2. *If $\{R_n\}$ is a convergence rate of $\widehat{\theta}_n$ in terms of the loss L and $\{Q_n\} \prec \{R_n\}$, then $\{Q_n\}$ is also a convergence rate of $\widehat{\theta}_n$ in terms of the loss L.*

Proof.

1. By Definition 45 we have that $\{R_n\} \in \{Q_n\}$ and $\{Q_n\} \in \{R_n\}$, hence $\{R_n\} \sim \{Q_n\}$.

2. Let $C_1, C_2 > 0$ be such that for all $n \in \mathbb{N}$ it holds that $R_n \leq C_1 Q_n$ and $Q_n \leq C_2 R_n$. Let $\delta > 0$. Then, by the definition of the convergence rate, there exists $C > 0$ such that (3.4) is satisfied. Therefore, it also holds that

$$\sup_{n \in \mathbb{N}} \sup_{\theta \in \Theta} P_\theta \left(L(\widehat{\theta}_n, \theta) \geq C\, C_1 Q_n \right) \leq \sup_{n \in \mathbb{N}} \sup_{\theta \in \Theta} P_\theta \left(L(\widehat{\theta}_n, \theta) \geq C R_n \right) \leq \delta.$$

   It is also easy to see that if the property (3.4) in Definition 45 is satisfied for some other rate $\{R_n\}$, then for all $n \in \mathbb{N}$ it holds that $Q_n \leq C_2 R_n \leq C_2 C_3 R_n$ for some $C_3 > 0$. Hence $\{Q_n\} \in \{R_n\}$ and therefore $\{Q_n\}$ is a convergence rate of $\widehat{\theta}$ in terms of L.

From the definition, it is also easy to see that estimators with lower order convergence rates (in terms of the loss L) converge faster to the true value of θ. Therefore, one can choose among different estimators based on their convergence rates. This task is called the *minimax estimation*.

Definition 47 (Minimax optimal estimator) *Let L be some loss function. Consider an estimator $\widehat{\theta}_n$ with a convergence rate $\{R_n\}$ in terms of the loss L. This estimator (and its convergence rate) is called* minimax optimal *if, for all other estimators $\widetilde{\theta}_n$ with the corresponding convergence rate $\{Q_n\}$, it holds that $\{R_n\} \sim \{Q_n\}$.*

In this chapter, we will derive convergence rates for the location halfspace median and scatter halfspace median in the Huber contamination model and show that these estimators are indeed minimax optimal in such settings.

The general strategy to find estimators that attain a minimax optimal convergence rate is based on two steps. First, we find some lower bound on the minimax optimal rate. That is, we search for some rate $\{Q_n\}$ such that for all estimators with convergence rate $\{R_n\}$ it holds that $\{Q_n\} \in \{R_n\}$. This can be solved, for example, by converting this problem into a hypothesis testing task; refer to [Tsybakov, 2009, Chapter 2] for details. In the second step, we have to find some estimator with convergence rate $\{R_n\}$ such that $\{Q_n\} \sim \{R_n\}$, to show that $\{Q_n\}$ is attainable. The following theorem will be useful to obtain the lower bound of the minimax optimal rate.

Theorem 48 *Let $\{Q_n\}$ be a rate and L be some loss function. Assume that there exist $C, \delta > 0$ such that*

$$\inf_{n \in \mathbb{N}} \inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} P_\theta \left(L(\widehat{\theta}_n, \theta) \geq C Q_n \right) \geq \delta, \tag{3.5}$$

where the second infimum is taken over all estimators $\widehat{\theta}_n$ based on a random sample $X_1, \ldots, X_n \sim P_\theta$. Then, for every estimator $\widehat{\theta}_n$ with a convergence rate $\{R_n\}$ in terms of L, it holds that $\{Q_n\} \in \{R_n\}$.

Proof. For a contradiction, assume that there exists a convergence rate $\{R_n\}$ of some estimator $\widehat{\theta}_n$ such that $\{Q_n\}$ $\{R_n\}$ does not hold. By the assumption (3.5), let $C, \delta > 0$ be such that

$$\inf_{n \in \mathbb{N}} \sup_{\theta \in \Theta} P_\theta \left(L(\widehat{\theta}_n, \theta) \geq C Q_n \right) \geq \delta. \tag{3.6}$$

Because $\{R_n\}$ is a convergence rate of the estimator $\widehat{\theta}_n$ in terms of L, by definition (3.4) there exists $C_2 > 0$ such that

$$\sup_{n \in \mathbb{N}} \sup_{\theta \in \Theta} P_\theta \left(L(\widehat{\theta}_n, \theta) \geq C_2 R_n \right) \leq \delta/2. \tag{3.7}$$

Because we assume that $\{Q_n\}$ $\{R_n\}$ is not true, there exists some $m \in \mathbb{N}$ such that $Q_m > C_2/C\, R_m$. However, this implies that $C\, Q_m > C_2\, R_m$ and together with (3.6), we can bound

$$\sup_{\theta \in \Theta} P_\theta \left(L(\widehat{\theta}_m, \theta) \geq C_2 R_m \right) \geq \sup_{\theta \in \Theta} P_\theta \left(L(\widehat{\theta}_m, \theta) \geq C Q_m \right) \geq \delta.$$

This contradicts (3.7); thus, the theorem is proved.

Also, note that if we define the risk function as

$$R(\widehat{\theta}_n, \theta) = \mathsf{E}_\theta \left(L(\widehat{\theta}_n, \theta) \right),$$

then the other approach to minimax estimation is to minimize the maximal risk

$$\sup_{\theta \in \Theta} R(\widehat{\theta}_n, \theta)$$

over all estimators $\widehat{\theta}_n$. If the condition (3.5) is satisfied for some $C, \delta > 0$, then by the Markov inequality we have

$$\inf_{n \in \mathbb{N}} \inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathsf{E}_\theta \left(L(\widehat{\theta}_n, \theta) \right) \geq \delta\, C\, Q_n.$$

Therefore, to bound the maximal risk from below, it is sufficient to find $C, \delta > 0$ such that (3.5) holds. This is the most common method in the literature; see [Tsybakov, 2009, Section 2.2]. However, to show that the location and scatter halfspace median are minimax optimal, it will be convenient to work directly with probabilities instead of risks.

3.2 Distance of probability measures

In this chapter, the following notions regarding the distance of probability measures will be useful.

Definition 49 (Total variation distance, Kullback-Leibler divergence) *Consider two probability measures $P, Q \in \mathcal{P}(\mathbb{R}^d)$. Then, the* total variation distance *of P and Q is defined as*

$$\mathsf{TV}\,(P, Q) = \sup_{F} |P(F) - Q(F)|,$$

where the supremum is taken over all measurable sets $F \quad \mathsf{R}^d$. *If* $P \quad Q$, *then the Kullback-Leibler divergence from* Q *to* P *is defined as*

$$D(P||Q) = \int_{\mathsf{R}^d} \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P.$$

It is easy to see that $\mathsf{TV}(\cdot,\cdot)$ is a metric on $\mathsf{P}(\mathsf{R}^d)$. Also, note that for any two probability measures $P, Q \quad \mathsf{P}(\mathsf{R}^d)$ there exists another probability measure $\mu \quad \mathsf{P}(\mathsf{R}^d)$ which dominates both P and Q in the sense that $P \quad \mu$ and $Q \quad \mu$. For example, we can take $\mu = \frac{1}{2}P + \frac{1}{2}Q$ or some other convex combination. This allows us to formulate Scheffé's Theorem, which associates the total variation distance with L^1 distance of densities.

Theorem 50 (Scheffé's Theorem) *Let* $P, Q \quad \mathsf{P}(\mathsf{R}^d)$ *and* $\mu \quad \mathsf{P}(\mathsf{R}^d)$ *be such that* $P, Q \quad \mu$. *Denote* $p = \mathrm{d}P/\,\mathrm{d}\mu$ *and* $q = \mathrm{d}Q/\,\mathrm{d}\mu$. *Then, it holds that*

$$\mathsf{TV}(P,Q) = \frac{1}{2}\int_{\mathsf{R}^d} |p - q|\,\mathrm{d}\mu = 1 - \int_{\mathsf{R}^d} (p \quad q)\,\mathrm{d}\mu.$$

Proof. At first, note that the following equalities hold

$$\begin{aligned}(p-q)\,\mathbf{1}_{\{p\geq q\}} + (q-p)\,\mathbf{1}_{\{p<q\}} &= |p-q|\,,\\ (p-q)\,\mathbf{1}_{\{p\geq q\}} &= p - p \quad q,\\ (q-p)\,\mathbf{1}_{\{p<q\}} &= q - p \quad q.\end{aligned}$$

By integrating these over R^d w.r.t. μ we get

$$\begin{aligned}\int_{p\geq q}(p-q)\,\mathrm{d}\mu + \int_{p<q}(q-p)\,\mathrm{d}\mu &= \int_{\mathsf{R}^d}|p-q|\,\mathrm{d}\mu,\\ \int_{p\geq q}(p-q)\,\mathrm{d}\mu &= 1 - \int_{\mathsf{R}^d} p \quad q\,\mathrm{d}\mu,\\ \int_{p<q}(q-p)\,\mathrm{d}\mu &= 1 - \int_{\mathsf{R}^d} p \quad q\,\mathrm{d}\mu.\end{aligned} \tag{3.8}$$

The latter two integrals are equal. This implies

$$\int_{p\geq q}(p-q)\,\mathrm{d}\mu = \int_{p<q}(q-p)\,\mathrm{d}\mu = \frac{1}{2}\int_{\mathsf{R}^d}|p-q|\,\mathrm{d}\mu. \tag{3.9}$$

Denoting $A = \left\{x \quad \mathsf{R}^d \mid p(x) \geq q(x)\right\}$, for any measurable $F \quad \mathsf{R}^d$ it holds that

$$\begin{aligned}|P(F) - Q(F)| = \left|\int_F (p-q)\,\mathrm{d}\mu\right| &= \int_{F\cap A}(p-q)\,\mathrm{d}\mu + \int_{F\cap A^{\mathsf{c}}}(q-p)\,\mathrm{d}\mu\\ &\leq \int_{F\cap A}(p-q)\,\mathrm{d}\mu \leq \int_A (p-q)\,\mathrm{d}\mu = |P(A) - Q(A)|\,.\end{aligned}$$

Therefore, the supremum in the definition of the total variation distance is attained for the set A (or A^{c} by symmetry). Hence

$$\begin{aligned}\mathsf{TV}(P,Q) = |P(A) - Q(A)| &= \int_{p\geq q}(p-q)\,\mathrm{d}\mu\\ &\overset{(3.9)}{=} \frac{1}{2}\int_{\mathsf{R}^d}|p-q|\,\mathrm{d}\mu \overset{(3.8)}{=} 1 - \int_{\mathsf{R}^d}(p \quad q)\,\mathrm{d}\mu.\end{aligned}$$

The total variation distance and Kullback-Leibler divergence are closely linked through Pinsker's inequality.

Theorem 51 (Pinsker's inequality) *For any P, Q $P(\mathsf{R}^d)$ it holds that*

$$\mathsf{TV}(P,Q)^2 \leq D(P||Q)/2.$$

Proof. Refer to [Tsybakov, 2009, Lemma 2.5].

The following theorem determines the Kullback-Leibler divergence of two multivariate Gaussian distributions.

Theorem 52 *For any μ_1, μ_2 R^d and Σ_1, Σ_2 PD_d it holds that*

$$D(N_d(\mu_1,\Sigma_1)\,||\,N_d(\mu_2,\Sigma_2)) = \frac{1}{2}\left(\log\frac{|\Sigma_2|}{|\Sigma_1|} + \mathsf{tr}\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_1-\mu_2)^\mathsf{T}\Sigma_2^{-1}(\mu_1-\mu_2) - d\right).$$

In particular, for $\Sigma_1 = \Sigma_2 = \mathbf{I}_d$ we obtain

$$D(N_d(\mu_1,\mathbf{I}_d)\,||\,N_d(\mu_2,\mathbf{I}_d)) = \frac{1}{2}\,\|\mu_1-\mu_2\|^2.$$

Proof. For $i = 1, 2$, denote by P_i the distribution $N_d(\mu_i, \Sigma_i)$ and by p_i its probability density function w.r.t. λ_d. It is easy to see that $\mathrm{d}P_1/\mathrm{d}P_2 = p_1/p_2$. Note that

$$\log\left(\frac{p_1(x)}{p_2(x)}\right) = \frac{1}{2}\left(\log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) - (x-\mu_1)^\mathsf{T}\Sigma_1^{-1}(x-\mu_1) + (x-\mu_2)^\mathsf{T}\Sigma_2^{-1}(x-\mu_2)\right)$$

for any x R^d. Let X P_1, then $D(P_1||P_2)$ can be obviously calculated as

$$D(P_1||P_2) = \mathsf{E}\left[\log\left(\frac{p_1(X)}{p_2(X)}\right)\right]$$
$$= \frac{1}{2}\left(\log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) - \mathsf{E}(X-\mu_1)^\mathsf{T}\Sigma_1^{-1}(X-\mu_1) + \mathsf{E}(X-\mu_2)^\mathsf{T}\Sigma_2^{-1}(X-\mu_2)\right).$$

Note that for a random vector Y $P(\mathsf{R}^d)$ with mean μ, variance matrix Σ, and a matrix $\mathbf{A}$ PD_d, it holds that

$$\mathsf{E}\left[Y^\mathsf{T}\mathbf{A}Y\right] = \mu^\mathsf{T}\mathbf{A}\mu + \mathsf{tr}\left(\mathbf{A}\Sigma\right).$$

Using this, we obtain

$$D(P_1||P_2) = \frac{1}{2}\left(\log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) - \mathsf{tr}\left(\mathbf{I}_d\right) + \mathsf{tr}\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_1-\mu_2)^\mathsf{T}\Sigma_2^{-1}(\mu_1-\mu_2)\right)$$
$$= \frac{1}{2}\left(\log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) + \mathsf{tr}\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_1-\mu_2)^\mathsf{T}\Sigma_2^{-1}(\mu_1-\mu_2) - d\right).$$

3.3 Huber's contamination model

Recall that we are dealing with a statistical model $M = \{P_\theta \mid \theta \in \Theta\} \subseteq P(\mathbb{R}^d)$. However, in this chapter, our task is to estimate θ from Huber's contamination model.

Definition 53 (Huber's contamination model) *Let $\varepsilon \in [0,1]$ and $Q \in P(\mathbb{R}^d)$. We denote by*

$$P_{(\varepsilon,\theta,Q)} = (1-\varepsilon)P_\theta + \varepsilon Q$$

the distribution P_θ that is ε-contaminated by Q. For a fixed $\varepsilon \in [0,1)$, the statistical model

$$H_\varepsilon = \left\{P_{(\varepsilon,\theta,Q)} \mid \theta \in \Theta, Q \in P(\mathbb{R}^d)\right\}$$

is called Huber's contamination model.

Huber's contamination model assumes that the data may contain both "clean" observations from the assumed distribution P_θ and "contaminating" observations from some other distribution Q (outliers, faulty observations, ...).

Note *For $\varepsilon > 0$, the parameter θ is usually not identifiable in the model H_ε. This means that for $\theta_1, \theta_2 \in \Theta, \theta_1 = \theta_2$, one is able to find $Q_1, Q_2 \in P(\mathbb{R}^d)$ such that $P_{(\varepsilon,\theta_1,Q_1)} = P_{(\varepsilon,\theta_2,Q_2)}$. This will be used to prove Theorem 55.*

As indicated in Section 3.1, we must first find a lower bound to determine a minimax optimal convergence rate in Huber's contamination model H_ε. However, it turns out that if we can find a lower bound $\{R_n\}$ for a minimax optimal rate in the non-contaminated model M, then $\{R_n \cdot \omega(\varepsilon,\Theta)\}$ is a lower bound for a minimax optimal rate in model H_ε, where $\omega(\varepsilon,\Theta)$ is the modulus of continuity of model M.

Definition 54 (Modulus of continuity) *Given a statistical model M, a loss function L and $\varepsilon \in [0,1)$, the quantity*

$$\omega(\varepsilon,\Theta) = \sup\{L(\theta_1,\theta_2) \mid \mathrm{TV}(P_{\theta_1},P_{\theta_2}) \leq \varepsilon/(1-\varepsilon); \theta_1,\theta_2 \in \Theta\}$$

is called the modulus of continuity *of model M.*

The concept of the modulus of continuity dates back to the pioneering work of Donoho and Liu [1991]. The interpretation is that the quantity $\omega(\varepsilon,\Theta)$ assesses the ability of the loss function L to distinguish between two distributions $P_{\theta_1}, P_{\theta_2}$ that are close in terms of the total variation. In every model M with identifiable parameter θ, $\omega(0,\Theta)$ equals 0.

The following theorem is one of the most important results of this work. It asserts that if we have a lower bound for the minimax optimal convergence rate in model M, we can determine a lower bound for the minimax optimal convergence rate in model H_ε using the modulus of continuity $\omega(\varepsilon,\Theta)$. Recall that $\widehat{\theta}_n = \Xi\left(\widehat{P}_n\right)$ represents an estimator based on a statistical functional Ξ and an empirical measure $\widehat{P}_n$. Further on, this empirical measure will correspond to a random sample $X_1, \ldots, X_n$ from a distribution P_θ (in model M) or $P_{(\varepsilon,\theta,Q)}$ (in model H_ε). A lower index always indicates the underlying generating distribution, compare (3.10) and (3.11).

Theorem 55 *Let $\delta > 0$ be an absolute constant, L be a loss function satisfying the τ-triangle inequality, and $\{R_n^0\}$ be a rate such that in model M it holds that*

$$\inf_{n \in \mathbb{N}} \inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} P_\theta \left(L\left(\widehat{\theta}_n, \theta\right) \geq R_n^0 \right) \geq \delta. \tag{3.10}$$

Then, for all $\varepsilon \in [0,1)$, in model H_ε it holds that

$$\inf_{n \in \mathbb{N}} \inf_{\widehat{\theta}_n} \sup_{\theta, Q} P_{(\varepsilon,\theta,Q)} \left(L\left(\widehat{\theta}_n, \theta\right) \geq R_n^\varepsilon \right) \geq \left(\delta \quad \frac{1}{2} \right), \tag{3.11}$$

where $\{R_n^\varepsilon\} \in \{R_n^0 \quad \omega(\varepsilon, \Theta)\}$.

Proof. The proof is based on the proof of [Chen et al., 2018, Theorem 5.1]. However, the proof is elaborated with more detailed steps, and some ambiguities are clarified; especially the last part of the proof, where the τ-triangle inequality is used to obtain the desired lower bound.

The assertion is trivial if $\varepsilon = 0$. Consider fixed $n \in \mathbb{N}, \varepsilon \in (0,1)$ and assume that (3.10) holds. We consider two cases.

Case 1: At first, assume $R_n^0 \geq \omega(\varepsilon, \Theta)$, hence $R_n^0 \wedge \omega(\varepsilon, \Theta) = R_n^0$. By setting $Q = P_\theta$, we have $P_{(\varepsilon,\theta,Q)} = P_\theta$. Therefore

$$\inf_{\widehat{\theta}_n} \sup_{\theta, Q} P_{(\varepsilon,\theta,Q)} \left(L\left(\widehat{\theta}_n, \theta\right) \geq R_n^0 \right) \geq \inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} P_\theta \left(L\left(\widehat{\theta}_n, \theta\right) \geq R_n^0 \right) \geq \delta \geq \left(\delta \quad \frac{1}{2} \right).$$

Hence, (3.11) holds with $R_n^\varepsilon = R_n^0$. The interesting case is the second one.

Case 2: Now, assume $R_n^0 < \omega(\varepsilon, \Theta)$, hence $R_n^0 \wedge \omega(\varepsilon, \Theta) = \omega(\varepsilon, \Theta)$. The idea is to find $\theta_1, \theta_2 \in \Theta$ and $Q_1, Q_2 \in \mathcal{P}(\mathbb{R}^d)$ such that $P_{(\varepsilon,\theta_1,Q_1)} = P_{(\varepsilon,\theta_2,Q_2)}$ but $L(\theta_1, \theta_2)$ is close to $\omega(\varepsilon, \Theta)$.

By Definition 54 of the modulus of continuity, find θ_1, θ_2 such that

$$0 < \mathrm{TV}(P_{\theta_1}, P_{\theta_2}) \leq \varepsilon/(1-\varepsilon)$$

and $L(\theta_1, \theta_2)$ is arbitrarily close to $\omega(\varepsilon, \Theta)$. Then, there exists $\varepsilon \in (0, \varepsilon]$ such that

$$\mathrm{TV}(P_{\theta_1}, P_{\theta_2}) = \frac{\varepsilon}{1-\varepsilon}.$$

Let $\mu = \frac{1}{2} P_{\theta_1} + \frac{1}{2} P_{\theta_2}$ so that $P_{\theta_1}, P_{\theta_2} \ll \mu$ and define densities

$$p_{\theta_j} = \frac{\mathrm{d}P_{\theta_j}}{\mathrm{d}\mu}, \qquad j \in \{1, 2\}.$$

Using these, we can define measures Q_1, Q_2 by their densities

$$\frac{\mathrm{d}Q_1}{\mathrm{d}\mu} = \frac{(p_{\theta_2} - p_{\theta_1}) \mathbf{1}\{p_{\theta_2} \geq p_{\theta_1}\}}{\mathrm{TV}(P_{\theta_1}, P_{\theta_2})},$$
$$\frac{\mathrm{d}Q_2}{\mathrm{d}\mu} = \frac{(p_{\theta_1} - p_{\theta_2}) \mathbf{1}\{p_{\theta_1} \geq p_{\theta_2}\}}{\mathrm{TV}(P_{\theta_1}, P_{\theta_2})}.$$

Now, we verify that Q_1, Q_2 are probability measures. This is done by the same analysis as was carried out in the proof of Scheffé's Theorem 50, see equalities (3.8) and (3.9). Namely, we have

$$\begin{aligned}\int_{\mathsf{R}^d} (p_{\theta_2} - p_{\theta_1})\, \mathbf{1}\{p_{\theta_2} \geq p_{\theta_1}\}\, \mathrm{d}\mu &= \int_{\mathsf{R}^d} (p_{\theta_1} - p_{\theta_2})\, \mathbf{1}\{p_{\theta_1} \geq p_{\theta_2}\}\, \mathrm{d}\mu \\ &= \frac{1}{2}\int_{\mathsf{R}^d} |p_{\theta_1} - p_{\theta_2}|\, \mathrm{d}\mu = \mathsf{TV}(P_{\theta_1}, P_{\theta_2}),\end{aligned}$$

where the last equality also follows from Scheffé's Theorem 50. Therefore, Q_1 and Q_2 are indeed probability measures. Now, consider contaminated probability distributions

$$P_j = (1 - \varepsilon\ )P_{\theta_j} + \varepsilon\ Q_j, \qquad j \quad \{1,2\}.$$

Recall that $\mathsf{TV}(P_{\theta_1}, P_{\theta_2}) = \varepsilon\ /(1 - \varepsilon\ )$. Hence, the density of P_1 w.r.t. μ is

$$\begin{aligned}\frac{\mathrm{d}P_1}{\mathrm{d}\mu} &= (1 - \varepsilon\ )p_{\theta_1} + \frac{\varepsilon}{\mathsf{TV}(P_{\theta_1}, P_{\theta_2})}(p_{\theta_2} - p_{\theta_1})\, \mathbf{1}\{p_{\theta_2} \geq p_{\theta_1}\} \\ &= (1 - \varepsilon\ )\,(p_{\theta_1} + (p_{\theta_2} - p_{\theta_1})\, \mathbf{1}\{p_{\theta_2} \geq p_{\theta_1}\}) \\ &= (1 - \varepsilon\ )\,(p_{\theta_1} \quad p_{\theta_2}).\end{aligned}$$

Analogously

$$\frac{\mathrm{d}P_2}{\mathrm{d}\mu} = (1 - \varepsilon\ )\,(p_{\theta_1} \quad p_{\theta_2}),$$

hence $P_1 = P_2$. To sum up, we have found parameters θ_1, θ_2 and distributions Q_1, Q_2 such that $L(\theta_1, \theta_2)$ is arbitrarily close to $\omega(\varepsilon, \Theta)$ but the contaminated distributions P_1 and P_2 generated by these parameters are identical, denoted by $P_1 = P_2 = P$.

Now, we can prove the desired inequality (3.11). First, note that for all estimators $\widehat{\theta}_n$ we have, using the τ-triangle inequality of L (3.3), that

$$L(\theta_1, \theta_2) \leq \tau\left(L\left(\widehat{\theta}_n, \theta_1\right) + L\left(\widehat{\theta}_n, \theta_2\right)\right) \leq 2\tau\left(L\left(\widehat{\theta}_n, \theta_1\right) \quad L\left(\widehat{\theta}_n, \theta_2\right)\right),$$

which implies

$$\frac{1}{2\tau}L(\theta_1, \theta_2) \leq L\left(\widehat{\theta}_n, \theta_1\right) \quad L\left(\widehat{\theta}_n, \theta_2\right). \tag{3.12}$$

As a consequence, we can bound

$$\begin{aligned}&\inf_{\widehat{\theta}_n} \sup_{\theta, Q} P_{(\varepsilon,\theta,Q)}\left(L\left(\widehat{\theta}_n, \theta\right) \geq \frac{1}{2\tau}L(\theta_1, \theta_2)\right) \\ &\quad \geq \inf_{\widehat{\theta}_n} \max_{j \ \{1,2\}} P_j\left(L\left(\widehat{\theta}_n, \theta_j\right) \geq \frac{1}{2\tau}L(\theta_1, \theta_2)\right) \\ &\quad = \inf_{\widehat{\theta}_n} \max_{j \ \{1,2\}} P\ \left(L\left(\widehat{\theta}_n, \theta_j\right) \geq \frac{1}{2\tau}L(\theta_1, \theta_2)\right) \\ &\quad \geq \inf_{\widehat{\theta}_n} \max_{j \ \{1,2\}} P\ \left(L\left(\widehat{\theta}_n, \theta_j\right) \geq \left(L\left(\widehat{\theta}_n, \theta_1\right) \quad L\left(\widehat{\theta}_n, \theta_2\right)\right)\right).\end{aligned} \tag{3.13}$$

Because $P_j \quad \mathsf{H}_\varepsilon$ for $j = 1, 2$, the first inequality follows by taking the supremum over a smaller set. For the equality, we used that $P_1 = P_2 = P$, and in the second inequality, we used (3.12). However, for any estimator $\widehat{\theta}_n$, at least one of the events

$$\left[L\left(\widehat{\theta}_n, \theta_1\right) \geq L\left(\widehat{\theta}_n, \theta_2\right)\right],\ \left[L\left(\widehat{\theta}_n, \theta_2\right) \geq L\left(\widehat{\theta}_n, \theta_1\right)\right]$$

has P -probability at least $1/2$. Therefore, (3.13) can be further bounded from below by $1/2$. In conclusion, because n was chosen arbitrarily, we have shown that

$$\inf_{n} \inf_{\widehat{\theta}_n} \sup_{\theta, Q} P_{(\varepsilon,\theta,Q)} \left(L\left(\widehat{\theta}_n, \theta\right) \geq \frac{1}{2\tau} L\left(\theta_1, \theta_2\right) \right) \geq \frac{1}{2}.$$

Because $\tau \geq 1$ is some fixed constant, it holds that

$$\frac{1}{2\tau} L(\theta_1, \theta_2) \quad L(\theta_1, \theta_2)$$

and θ_1, θ_2 can be chosen such that $L(\theta_1, \theta_2)$ is arbitrarily close to $\omega(\varepsilon, \Theta)$. This implies that (3.11) holds with

$$R_n^{\varepsilon} \quad \omega(\varepsilon, \Theta),$$

which concludes the proof.

Note *Chen et al. [2018] use Le Cam's two point testing method (see [Yu, 1997, Lemma 1]) to bound* (3.13) *from below. As we have shown, this is not necessary.*

To sum up, we have shown that to bound a minimax optimal rate in model H_{ε} from below, it suffices to find the lower bound for minimax optimal rate in model M and determine the modulus of continuity of model M.

3.4 Optimality of location halfspace median

Motivated by the previous theory for general parametric models, in this section, we consider a specific parametric model

$$M^{\text{loc}} = \left\{ N_d(\mu, \Sigma) \mid \mu \quad \mathbb{R}^d, \Sigma \quad \text{PD}_d, \sigma_{max}(\Sigma) \leq M \right\},$$

where $M > 0$ is fixed. We consider all Gaussian distributions with covariance matrices having a bounded spectrum. The parameter of interest is the mean $\theta = \mu \quad \mathbb{R}^d = \Theta$. This is a location parameter, hence the superscript "loc" in M^{loc}. Estimators of μ are denoted by $\widehat{\mu}_n$ and the location halfspace median is denoted by $\widehat{\mu}_n^{\text{hs}}$. We consider the squared Euclidean loss, i.e.

$$L(\mu_1, \mu_2) = \ \mu_1 - \mu_2 \ ^2.$$

Recall that this loss function satisfies the τ-triangle inequality for $\tau = 2$ (see inequality (3.3)). The corresponding Huber's contamination model for $\varepsilon \quad [0, 1)$ is

$$H_{\varepsilon}^{\text{loc}} = \left\{ (1 - \varepsilon) N_d(\mu, \Sigma) + \varepsilon Q \mid \mu \quad \mathbb{R}^d, \Sigma \quad \text{PD}_d, \sigma_{max}(\Sigma) \leq M, Q \quad P(\mathbb{R}^d) \right\}.$$

In this section, by $P_{(\mu,\Sigma)}$ we denote the probability distribution $N_d(\mu, \Sigma)$. Further, by $P_{(\mu,\Sigma,Q)}$ we denote the contaminated probability distribution

$$(1 - \varepsilon) N_d(\mu, \Sigma) + \varepsilon Q.$$

The amount of contamination ε is omitted in this notation and will be evident from the context.

First, we focus on establishing a lower bound for a minimax optimal convergence rate in model M^{loc}. The following theorem is one of the essential tools for determining lower bounds in the minimax estimation; see [Tsybakov, 2009, Section 2.7] for reference.

Theorem 56 (Fano's inequality) *Let $M \in \mathbb{N}$, $M \geq 2$ and $\{P_1, \ldots, P_M\} \subset \mathcal{P}(\mathbb{R}^d)$ with the corresponding parameters $\theta_1, \ldots, \theta_M \in \mathbb{R}^p$. If $a, b > 0$ are such that $\|\theta_i - \theta_j\| \geq a$ and $D(P_i||P_j) \leq b$ for all $i \neq j$, then for all $n \in \mathbb{N}$ it holds that*

$$\inf_{\widehat{\theta}_n} \frac{1}{M} \sum_{i=1}^{M} P_i\left(\left\|\theta_i - \widehat{\theta}_n\right\| \geq a/2\right) \geq 1 - \frac{n\,b + \log(2)}{\log(M)}, \tag{3.14}$$

where the infimum is taken over all estimators $\widehat{\theta}_n$ of θ based on a random sample $X_1, \ldots, X_n$. In (3.14), the probability

$$P_i\left(\left\|\theta_i - \widehat{\theta}_n\right\| \geq a/2\right)$$

considers that the estimator $\widehat{\theta}_n$ is calculated using a random sample $X_1, \ldots, X_n$ from the distribution P_i.

Proof. The claim follows from the proof of [Yu, 1997, Lemma 3]. The proof utilizes advanced methods from information theory.

Using Fano's inequality, we can deduce the following result.

Theorem 57 *In model M^{loc}, there exists an absolute constant $\delta > 0$ such that*

$$\inf_{n \in \mathbb{N}} \inf_{\widehat{\mu}_n} \sup_{P_{(\mu,\Sigma)} \in M^{\text{loc}}} P_{(\mu,\Sigma)}\left(\left\|\widehat{\mu}_n - \mu\right\|^2 \geq R_n^0\right) \geq \delta \tag{3.15}$$

holds for some $\{R_n^0\} \subset d/n$. The infimum is taken over all estimators $\widehat{\mu}_n$ of μ based on a random sample $X_1, \ldots, X_n \sim P_{(\mu,\Sigma)}$.

Note *Chen et al. [2018] assure that the claim of Theorem 57 is a classic result of minimax estimation theory. The authors provide [Ma and Wu, 2015] as a reference for this assertion. However, I did not find the stated claim in this paper or in other literature. Therefore, the proof of this theorem is the original result of this work. The guide [Scarlett and Cevher, 2021] was helpful for correctly applying Fano's inequality.*

Proof. Let us consider

$$R_n^0 = \frac{\kappa}{64}\frac{d}{n} \quad \text{where} \quad \kappa = \frac{\log(2)}{4} > 0. \tag{3.16}$$

We will prove that (3.15) holds with $\delta = 1/4$ using Fano's inequality. Obviously, for all $n \in \mathbb{N}$, we can bound

$$\inf_{\widehat{\mu}_n} \sup_{P_{(\mu,\Sigma)} \in M^{\text{loc}}} P_{(\mu,\Sigma)}\left(\left\|\widehat{\mu}_n - \mu\right\|^2 \geq R_n^0\right) \geq \inf_{\widehat{\mu}_n} \sup_{\mu \in \Theta_0} P_{(\mu,\mathbf{I}_d)}\left(\left\|\widehat{\mu}_n - \mu\right\|^2 \geq R_n^0\right),$$

where

$$\Theta_0 = \left\{ \mu \quad \mathbb{R}^d \mid \quad \mu \quad \leq \sqrt{\frac{\kappa\, d}{n}} \right\}. \tag{3.17}$$

Therefore, to prove this theorem, it is sufficient to show that for all $n \quad \mathbb{N}$, the inequality

$$\inf_{\widehat{\mu}_n} \sup_{\mu \quad \Theta_0} P_{(\mu, \mathbf{I}_d)} \left(\quad \widehat{\mu}_n - \mu \quad ^2 \geq R_n^0 \right) \geq \frac{1}{4} \tag{3.18}$$

holds.

Fix $n \quad \mathbb{N}$ and let $\{\mu_1, \ldots, \mu_M\} \quad \Theta_0$ be a set of the maximum cardinality such that for all $i = j$ we have

$$\mu_i - \mu_j \quad > \frac{1}{4}\sqrt{\frac{\kappa\, d}{n}}. \tag{3.19}$$

The set $\{\mu_1, \ldots, \mu_M\}$ is a so-called $\frac{1}{4}\sqrt{\frac{\kappa\, d}{n}}$-packing of Θ_0. Then:

1. By Theorem 52 and our choice of Θ_0 in (3.17), we can bound

$$D(P_{(\mu_i, \mathbf{I}_d)} || P_{(\mu_j, \mathbf{I}_d)}) = \frac{1}{2} \quad \mu_i - \mu_j \quad ^2 \leq \frac{1}{2} \left(\quad \mu_i \quad + \quad \mu_j \quad \right)^2 \leq 2\frac{\kappa\, d}{n}. \tag{3.20}$$

2. By Theorem A5 in the Appendix, we can bound the cardinality of any $\frac{1}{4}\sqrt{\frac{\kappa\, d}{n}}$-packing of Θ_0 from below by 4^d, i.e.

$$M \geq 4^d. \tag{3.21}$$

Using this, we bound the left side of (3.18) from below as follows

$$\begin{aligned}
&\inf_{\widehat{\mu}_n} \sup_{\mu \quad \Theta_0} P_{(\mu, \mathbf{I}_d)} \left(\quad \widehat{\mu}_n - \mu \quad ^2 \geq R_n^0 \right) \\
&\qquad \geq \inf_{\widehat{\mu}_n} \max_{i \quad \{1, \ldots, M\}} P_{(\mu_i, \mathbf{I}_d)} \left(\quad \widehat{\mu}_n - \mu_i \quad ^2 \geq R_n^0 \right) \\
&\qquad \geq \inf_{\widehat{\mu}_n} \frac{1}{M} \sum_{i=1}^{M} P_{(\mu_i, \mathbf{I}_d)} \left(\quad \widehat{\mu}_n - \mu_i \quad ^2 \geq R_n^0 \right) \\
&\qquad \overset{(3.16)}{=} \inf_{\widehat{\mu}_n} \frac{1}{M} \sum_{i=1}^{M} P_{(\mu_i, \mathbf{I}_d)} \left(\quad \widehat{\mu}_n - \mu_i \quad \geq \frac{1}{2} \left(\frac{1}{4}\sqrt{\frac{\kappa\, d}{n}} \right) \right) \\
&\qquad \overset{\text{Fano}}{\geq} 1 - \frac{2n\frac{\kappa\, d}{n} + \log(2)}{\log(M)} \\
&\qquad \overset{(3.21)}{\geq} 1 - \frac{2\, n\frac{\kappa\, d}{n} + \log(2)}{\log(4^d)} = 1 - \frac{\kappa}{\log(2)} - \frac{1}{2d} \\
&\qquad \overset{d \geq 1}{\geq} 1 - \frac{\kappa}{\log(2)} - \frac{1}{2} \overset{(3.16)}{=} \frac{1}{2} - \frac{1}{4} = \frac{1}{4}.
\end{aligned}$$

In the first inequality, we take the supremum over a smaller set. In the second inequality, we use that a maximum is always bounded from below by an average. In the third inequality, we use Fano's inequality (stated in Theorem 56) for $P_i = P_{(\mu_i, \mathbf{I}_d)}$ and $\theta_i = \mu_i$, $i = 1, \ldots, M$. The assumptions of Fano's inequality

are satisfied by (3.19), (3.20) and (3.21). This proves the theorem.

Now we can use Theorem 55 to determine a lower bound of a minimax optimal convergence rate in model $H_\varepsilon^{\text{loc}}$.

Theorem 58 *There exists an absolute constant $\delta > 0$ such that for any $\varepsilon \in [0, 1)$*

$$\inf_{n \in \mathbb{N}} \inf_{\widehat{\mu}_n} \sup_{P_{(\mu,\Sigma,Q)} \in H_\varepsilon^{\text{loc}}} P_{(\mu,\Sigma,Q)} \left(\|\widehat{\mu}_n - \mu\|^2 \geq R_n^\varepsilon \right) \geq \delta \tag{3.22}$$

holds for some $\{R_n^\varepsilon\} \subset (d/n) \cdot \varepsilon^2$. The infimum is taken over all estimators $\widehat{\mu}_n$ of μ based on a random sample $X_1, \ldots, X_n \sim P_{(\mu,\Sigma,Q)}$.

Proof. Without loss of generality, let $M \geq 1$. To use Theorem 55, we must first bound the modulus of continuity $\omega(\varepsilon, \Theta)$. Note that by Pinsker's inequality given in Theorem 51 and using the Kullback-Leibler divergence of two Gaussian distributions (see Theorem 52), we have for any $\mu_1, \mu_2 \in \mathbb{R}^d$ that

$$\text{TV}\left(P_{(\mu_1,\mathbf{I}_d)}, P_{(\mu_2,\mathbf{I}_d)}\right)^2 \leq \frac{1}{2} D(P_{(\mu_1,\mathbf{I}_d)} || P_{(\mu_2,\mathbf{I}_d)}) = \frac{1}{4} \|\mu_1 - \mu_2\|^2.$$

Therefore, if $\|\mu_1 - \mu_2\|/2 \leq \varepsilon$, then

$$\text{TV}\left(P_{(\mu_1,\mathbf{I}_d)}, P_{(\mu_2,\mathbf{I}_d)}\right) \leq \varepsilon \leq \frac{\varepsilon}{1-\varepsilon}.$$

As a consequence

$$\begin{aligned}\omega(\varepsilon, \Theta) &= \sup\left\{ \|\mu_1 - \mu_2\|^2 \mid \text{TV}\left(P_{(\mu_1,\Sigma_1)}, P_{(\mu_2,\Sigma_2)}\right) \leq \frac{\varepsilon}{1-\varepsilon} \right\} \\ &\geq \sup\left\{ \|\mu_1 - \mu_2\|^2 \mid \|\mu_1 - \mu_2\|/2 \leq \varepsilon \right\} = 4\varepsilon^2.\end{aligned}$$

The inequality follows from the fact that the first supremum is taken over all $\mu_1, \mu_2 \in \mathbb{R}^d$ and $\Sigma_1, \Sigma_2 \in \text{PD}_d$ such that $\sigma_{max}(\Sigma_1), \sigma_{max}(\Sigma_2) \leq M$ and the second supremum is taken only over $\Sigma_1 = \Sigma_2 = \mathbf{I}_d$ and μ_1, μ_2 such that $\|\mu_1 - \mu_2\|/2 \leq \varepsilon$. We have shown that $\omega(\varepsilon, \Theta) \geq 4\varepsilon^2$, hence $(d/n) \cdot \omega(\varepsilon, \Theta) \geq (d/n) \cdot (4\varepsilon^2)$. By Theorem 55 and Theorem 57, (3.22) holds for some $\{R_n^\varepsilon\} \subset (d/n) \cdot \omega(\varepsilon, \Theta)$, hence also for some $\{R_n^\varepsilon\} \subset (d/n) \cdot \varepsilon^2$.

We demonstrated that a minimax optimal convergence rate in model $H_\varepsilon^{\text{loc}}$ is bounded from below by $(d/n) \cdot \varepsilon^2$. Next, we will show that this rate is indeed attained for the location halfspace median $\widehat{\mu}_n^{\text{hs}}$. This indicates that the location halfspace median is indeed minimax optimal in model $H_\varepsilon^{\text{loc}}$.

Theorem 59 *Assume that $\varepsilon \in (0, 1/5)$. Then, for any $\delta \in (0, 1/2)$ there exist absolute constants $C, D > 0$ such that for all $n \in \mathbb{N}$ satisfying*

$$\left(\frac{d}{n} + \frac{\log(1/\delta)}{n}\right) < D \tag{3.23}$$

we have

$$\inf_{P_{(\mu,\mathbf{\Sigma},Q)}\ H_\varepsilon^{\mathrm{loc}}} P_{(\mu,\mathbf{\Sigma},Q)}\left(\left\|\widehat{\mu}_n^{\mathrm{hs}} - \mu\right\|^2 < C\left(\left(\frac{d}{n} \quad \varepsilon^2\right) + \frac{\log(1/\delta)}{n}\right)\right) \geq 1 - 2\delta,$$

hence

$$\sup_{P_{(\mu,\mathbf{\Sigma},Q)}\ H_\varepsilon^{\mathrm{loc}}} P_{(\mu,\mathbf{\Sigma},Q)}\left(\left\|\widehat{\mu}_n^{\mathrm{hs}} - \mu\right\|^2 \geq C\left(\left(\frac{d}{n} \quad \varepsilon^2\right) + \frac{\log(1/\delta)}{n}\right)\right) < 2\delta.$$

Proof. The proof is based on the proof of [Chen et al., 2018, Theorem 2.1]. However, the theorem is proven down to finer details, and some ambiguities are clarified. Especially the last part of the proof, where the Lipschitz continuity of the quantile function of the standard Gaussian distribution is utilized, is rigorously argued. The proof is divided into four parts.

Part 1: Auxiliary observations

First, we prepare the following observations that will be useful in deriving the intended bound.

(L1) By the affine equivariance of the location halfspace median, we can assume that $\mu = \mathbf{0}$. Also, without loss of generality, we can take $\mathbf{\Sigma} = \mathbf{I}_d$. Otherwise, we would transform the random sample by the linear mapping $x \to \mathbf{\Sigma}^{-1/2}x$ and note that

$$\left\|\mathbf{\Sigma}^{-1/2}\widehat{\mu}_n^{\mathrm{hs}}\right\|^2 = \left(\widehat{\mu}_n^{\mathrm{hs}}\right)^{\mathsf{T}}\mathbf{\Sigma}^{-1}\widehat{\mu}_n^{\mathrm{hs}} \geq \left\|\widehat{\mu}_n^{\mathrm{hs}}\right\|^2/M.$$

The inequality holds since all eigenvalues of $\mathbf{\Sigma}$ are bounded from above by M, hence all eigenvalues of $\mathbf{\Sigma}^{-1}$ are bounded from below by M^{-1}. Therefore, it suffices to find C under the assumption $\mathbf{\Sigma} = \mathbf{I}_d$ and then multiply this constant by M.

(L2) Recall that $hD(x; N_d(\mathbf{0}, \mathbf{I}_d)) = 1 - \Phi(\ x\ )$ for all $x \quad \mathsf{R}^d$. See Example 3 for derivation.

(L3) Considering a random sample $\{X_i\}_{i=1}^n$ from a contaminated distribution $P_{(\mathbf{0},\mathbf{I}_d,Q)}$, we can decompose $\{X_i\}_{i=1}^n = \{Y_i\}_{i=1}^{n_1} \quad \{Z_i\}_{i=1}^{n_2}$ where, marginally, $n_2 \quad \mathsf{Binomial}(n, \varepsilon)$, $n_1 = n - n_2$, and conditionally on n_1, n_2 we have that $\{Y_i\}_{i=1}^{n_1}$ is a random sample from $N_d(\mathbf{0}, \mathbf{I}_d)$ and $\{Z_i\}_{i=1}^{n_2}$ is a random sample from Q.

(L4) By Theorem A2 in the Appendix, we have with probability at least $1 - \delta$ that

$$\begin{aligned}&\sup_{x \ \mathsf{R}^d} |hD(x; N_d(\mathbf{0}, \mathbf{I}_d)) - hD(x; \{Y_i\}_{i=1}^{n_1})| \\ &\qquad \leq \sup_{H \ H_d} \left|P_{(\mathbf{0},\mathbf{I}_d)}(H) - \widehat{P}_{n_1}(H)\right| \leq \sqrt{\frac{1440\pi e}{1 - e^{-1}}}\sqrt{\frac{d+1}{n_1}} + \sqrt{\frac{\log(1/\delta)}{2n_1}},\end{aligned}$$

where $\widehat{P}_{n_1}$ is the empirical distribution of $\{Y_i\}_{i=1}^{n_1}$ and H_d is the system of all closed halfspaces in R^d, i.e.

$$H_d := \left\{H_{x,u} \mid x \quad \mathsf{R}^d, u \quad \mathsf{S}^{d-1}\right\}. \tag{3.24}$$

(L5) By the definition of the sample halfspace depth, it follows that

$$n_1\, hD(x; \{Y_i\}_{i=1}^{n_1}) \geq n\, hD(x; \{X_i\}_{i=1}^{n}) - n_2 \geq n_1\, hD(x; \{Y_i\}_{i=1}^{n_1}) - n_2 \tag{3.25}$$

holds for all $x \in \mathbb{R}^d$. For example, to see the first inequality, note that

$$\inf_{u \in S^{d-1}} \sum_{i=1}^{n_1} \mathbf{1}_{\{ Y_i,u \geq x,u \}} \geq \inf_{u \in S^{d-1}} \sum_{i=1}^{n} \mathbf{1}_{\{ X_i,u \geq x,u \}} - n_2. \tag{3.26}$$

This inequality holds because, for a fixed $u \in S^{d-1}$, the left side of (3.26) represents the number of observations Y_i in $H_{x,u}$, which is always greater than or equal to the number of observations X_i in $H_{x,u}$ without n_2. This is because some of the observations Z_i can also lie in $H_{x,u}$. The second inequality is proven analogously.

(L6) By Theorem A4 in the Appendix, if

$$\sqrt{\frac{\log(1/\delta)}{2\,n}} < 1/5 \tag{3.27}$$

holds for $\delta > 0$, then

$$\frac{n_2}{n_1} \leq \frac{\varepsilon}{1-\varepsilon} + \frac{25}{12}\sqrt{\frac{\log(1/\delta)}{2\,n}} < \frac{2}{3}$$

holds with probability at least $1 - \delta$.

Part 2: Choice of constant D

Now, we define a constant D so that the proof of the statement works. In particular, we need to ensure that (3.27) holds and also that

$$80\sqrt{\frac{3\pi e}{1-e^{-1}}}\sqrt{\frac{d}{n}} + \frac{7}{2}\sqrt{\frac{\log(1/\delta)}{n}} < \frac{1}{5} \tag{3.28}$$

is true, which will be useful later. For the entire subsequent proof, fix $\delta \in (0, 1/2)$ and consider $n \in \mathbb{N}$ such that the assumption (3.23) is satisfied for

$$D = \left(\frac{\frac{1}{5}}{80\sqrt{\frac{3\pi e}{1-e^{-1}}} + \frac{7}{2}} \right)^2 > 0.$$

Because $a + b < D$ implies $a \cdot b < D$ for all $a, b > 0$, it is easy to see that also

$$\sqrt{\frac{d}{n}} \cdot \sqrt{\frac{\log(1/\delta)}{n}} < \sqrt{D}. \tag{3.29}$$

Therefore, by our choice of D and (3.29), we can bound

$$80\sqrt{\frac{3\pi e}{1-e^{-1}}}\sqrt{\frac{d}{n}} + \frac{7}{2}\sqrt{\frac{\log(1/\delta)}{n}} < \left(80\sqrt{\frac{3\pi e}{1-e^{-1}}} + \frac{7}{2} \right) \sqrt{D} = \frac{1}{5}.$$

As a consequence, inequality (3.28) holds. Also, inequality (3.28) implies that

$$\sqrt{\frac{\log(1/\delta)}{2n}} < \frac{7}{2}\sqrt{\frac{\log(1/\delta)}{n}} < \frac{1}{5},$$

hence, using the assumption $\varepsilon < 1/5$, (L6) gives that

$$\mathsf{P}\left[\frac{n_2}{n_1} \leq \frac{\varepsilon}{1-\varepsilon} + \frac{25}{12}\sqrt{\frac{\log(1/\delta)}{2n}} < \frac{2}{3}\right] \geq 1-\delta. \tag{3.30}$$

Further, note that

$$\frac{n_2}{n_1} < \frac{2}{3} \qquad 3n_2 < 2n_1 \qquad 3(n-n_1) < 2n_1 \qquad n_1 > 3n/5. \tag{3.31}$$

To sum up, we find out that for our fixed δ and a large enough n, at least $3/5$ of all observations are non-contaminating with probability at least $1-\delta$. Also, the inequality (3.28) holds. Both of these facts will be used later in the proof.

Part 3: Bound for the depth of $\hat{\mu}_n^{\mathsf{hs}}$

We derive the following series of inequalities using the above observations and our choice of D. These hold with probability at least $1-\delta$ conditionally on the marginal counts n_1, n_2. We have

$$\begin{aligned}
1-\Phi\left(\left\|\hat{\mu}_n^{\mathsf{hs}}\right\|\right) &\overset{(L2)}{=} hD(\hat{\mu}_n^{\mathsf{hs}}; N_d(\mathbf{0}, \mathbf{I}_d)) \\
&\overset{(L4)}{\geq} hD(\hat{\mu}_n^{\mathsf{hs}}; \{Y_i\}_{i=1}^{n_1}) - \sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{d+1}{n_1}} - \sqrt{\frac{\log(1/\delta)}{2n_1}} \\
&\overset{(L5)}{\geq} \frac{n}{n_1} hD(\hat{\mu}_n^{\mathsf{hs}}; \{X_i\}_{i=1}^{n}) - \frac{n_2}{n_1} - \sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{d+1}{n_1}} - \sqrt{\frac{\log(1/\delta)}{2n_1}} \\
&\geq \frac{n}{n_1} hD(\mathbf{0}; \{X_i\}_{i=1}^{n}) - \frac{n_2}{n_1} - \sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{d+1}{n_1}} - \sqrt{\frac{\log(1/\delta)}{2n_1}} \\
&\overset{(L5)}{\geq} hD(\mathbf{0}; \{Y_i\}_{i=1}^{n_1}) - \frac{n_2}{n_1} - \sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{d+1}{n_1}} - \sqrt{\frac{\log(1/\delta)}{2n_1}} \\
&\overset{(L4)}{\geq} hD(\mathbf{0}; N_d(\mathbf{0}, \mathbf{I}_d)) - \frac{n_2}{n_1} - 2\sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{d+1}{n_1}} - \sqrt{\frac{2\log(1/\delta)}{n_1}} \\
&\overset{(L2)}{=} \frac{1}{2} - \frac{n_2}{n_1} - 2\sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{d+1}{n_1}} - \sqrt{\frac{2\log(1/\delta)}{n_1}},
\end{aligned} \tag{3.32}$$

where the third inequality follows from the fact that $\hat{\mu}_n^{\mathsf{hs}}$ is the maximizer of $hD(\cdot; \{X_i\}_{i=1}^n)$. Rewriting (3.32), we have that

$$\mathsf{P}\left[\Phi\left(\left\|\hat{\mu}_n^{\mathsf{hs}}\right\|\right) \leq \frac{1}{2} + \frac{n_2}{n_1} + 2\sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{d+1}{n_1}} + \sqrt{\frac{2\log(1/\delta)}{n_1}} \,\middle|\, n_1, n_2\right] \geq 1-\delta.$$

However, by taking the expectation on both sides (and considering its monotonicity), we obtain

$$\mathsf{P}\left[\Phi\left(\left\|\hat{\mu}_n^{\mathsf{hs}}\right\|\right) \leq \frac{1}{2} + \frac{n_2}{n_1} + 2\sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{d+1}{n_1}} + \sqrt{\frac{2\log(1/\delta)}{n_1}}\right] \geq 1-\delta.$$

Now, we combine this result with inequality (3.30). Note that for any two random events A, B we have that $\mathsf{P}(A \cap B) \geq \mathsf{P}(A) + \mathsf{P}(B) - 1$. Therefore, it holds that

$$\mathsf{P}\left[\Phi\left(\left\|\widehat{\boldsymbol{\mu}}_n^{\mathsf{hs}}\right\|\right) \leq \frac{1}{2} + \frac{n_2}{n_1} + 2\sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{d+1}{n_1}} + \sqrt{\frac{2\log(1/\delta)}{n_1}}, \right.$$
$$\left. \frac{n_2}{n_1} \leq \frac{\varepsilon}{1-\varepsilon} + \frac{25}{12}\sqrt{\frac{\log(1/\delta)}{2n}} \leq \frac{2}{3}\right] \geq 1 - 2\delta. \quad (3.33)$$

Now, under the condition of the second random event in (3.33), we can further upper bound

$$\begin{aligned}
\Phi\left(\left\|\widehat{\boldsymbol{\mu}}_n^{\mathsf{hs}}\right\|\right) &\overset{(3.32)}{\leq} \frac{1}{2} + \frac{n_2}{n_1} + 2\sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{d+1}{n_1}} + \sqrt{\frac{2\log(1/\delta)}{n_1}} \\
&\overset{\Phi(0)=\frac{1}{2}}{=} \Phi(0) + \frac{n_2}{n_1} + 2\sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{d+1}{n_1}} + \sqrt{\frac{2\log(1/\delta)}{n_1}} \\
&\overset{(3.30)}{\leq} \Phi(0) + \frac{\varepsilon}{1-\varepsilon} + \frac{25}{12}\sqrt{\frac{\log(1/\delta)}{2n}} + 2\sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{d+1}{n_1}} + \sqrt{\frac{2\log(1/\delta)}{n_1}} \\
&\overset{(3.31)}{\leq} \Phi(0) + \frac{\varepsilon}{1-\varepsilon} + \frac{25}{12}\sqrt{\frac{\log(1/\delta)}{2n}} + 2\sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{5(d+1)}{3n}} + \sqrt{\frac{10\log(1/\delta)}{3n}} \\
&= \Phi(0) + \frac{\varepsilon}{1-\varepsilon} + 40\sqrt{\frac{6\pi e}{1-e^{-1}}}\sqrt{\frac{d+1}{n}} + \sqrt{\frac{1585\log(1/\delta)}{288n}} \\
&\overset{d+1\leq 2d}{\leq} \Phi(0) + \frac{\varepsilon}{1-\varepsilon} + 40\sqrt{\frac{6\pi e}{1-e^{-1}}}\sqrt{\frac{2d}{n}} + \sqrt{\frac{1585}{288}}\sqrt{\frac{\log(1/\delta)}{n}} \\
&\overset{\varepsilon<1/5}{<} \Phi(0) + \frac{5}{4}\varepsilon + 80\sqrt{\frac{3\pi e}{1-e^{-1}}}\sqrt{\frac{d}{n}} + \frac{7}{2}\sqrt{\frac{\log(1/\delta)}{n}}.
\end{aligned}$$

In the last inequality, we also used the fact that $\sqrt{1585/288} < 7/2$. Ultimately, we have

$$\mathsf{P}\left[\Phi\left(\left\|\widehat{\boldsymbol{\mu}}_n^{\mathsf{hs}}\right\|\right) - \Phi(0) \leq \frac{5}{4}\varepsilon + 80\sqrt{\frac{3\pi e}{1-e^{-1}}}\sqrt{\frac{d}{n}} + \frac{7}{2}\sqrt{\frac{\log(1/\delta)}{n}}, \right.$$
$$\left. \frac{n_2}{n_1} \leq \frac{\varepsilon}{1-\varepsilon} + \frac{25}{12}\sqrt{\frac{\log(1/\delta)}{2n}} \leq \frac{2}{3}\right] \geq 1 - 2\delta.$$

For any random events A, B, we have $\mathsf{P}(A \cap B) \leq \mathsf{P}(A)$. Therefore, the preceding inequality implies that

$$\Phi\left(\left\|\widehat{\boldsymbol{\mu}}_n^{\mathsf{hs}}\right\|\right) - \Phi(0) < \frac{5}{4}\varepsilon + 80\sqrt{\frac{3\pi e}{1-e^{-1}}}\sqrt{\frac{d}{n}} + \frac{7}{2}\sqrt{\frac{\log(1/\delta)}{n}} \quad (3.34)$$

holds with probability at least $1 - 2\delta$.

Part 4: Conclusion

We assumed that n is chosen such that (3.28) holds. Therefore, using this and the assumption $\varepsilon < 1/5$, from (3.34) we can deduce

$$\Phi\left(\left\|\widehat{\boldsymbol{\mu}}_n^{\mathsf{hs}}\right\|\right) - \Phi(0) < \frac{5}{4}\varepsilon + \frac{1}{5} < \frac{1}{4} + \frac{1}{5} = \frac{9}{20}.$$

Now we can use the fact that $\Phi(0) = 1/2$ and the quantile function of the standard Gaussian distribution is Lipschitz continuous on the interval $[1/2, 19/20]$ with the Lipschitz constant

$$C = \left[\varphi\left(\Phi^{-1}(19/20)\right)\right]^{-1},$$

where φ is the probability density function of $\mathsf{N}_1(0,1)$. Therefore, from (3.34) we can further deduce

$$\begin{aligned}\left\|\widehat{\boldsymbol{\mu}}_n^{\mathsf{hs}}\right\| &< C\left(\frac{5}{4}\varepsilon + 80\sqrt{\frac{3\pi e}{1-e^{-1}}}\sqrt{\frac{d}{n}} + \frac{7}{2}\sqrt{\frac{\log(1/\delta)}{n}}\right) \\ &\leq C\left(\left(\frac{5}{4} + 80\sqrt{\frac{3\pi e}{1-e^{-1}}}\right)\left(\sqrt{\frac{d}{n}} \quad \varepsilon\right) + \frac{7}{2}\sqrt{\frac{\log(1/\delta)}{n}}\right),\end{aligned}$$

hence, using 2-triangle inequality (see (3.3)), we have

$$\left\|\widehat{\boldsymbol{\mu}}_n^{\mathsf{hs}}\right\|^2 \leq C\left(\left(\frac{d}{n} \quad \varepsilon^2\right) + \frac{\log(1/\delta)}{n}\right),$$

where C is defined as

$$2\,(C\,)^2\left(\left(\frac{5}{4} + 80\sqrt{\frac{3\pi e}{1-e^{-1}}}\right)^2 + \frac{49}{4}\right).$$

All in all, we have proven that

$$\mathsf{P}\left[\left\|\widehat{\boldsymbol{\mu}}_n^{\mathsf{hs}}\right\|^2 \leq C\left(\left(\frac{d}{n} \quad \varepsilon^2\right) + \frac{\log(1/\delta)}{n}\right)\right] \geq 1 - 2\delta, \tag{3.35}$$

which concludes the proof.

Finally, the minimax optimality of Tukey's median follows.

Theorem 60 *For $\varepsilon < 1/5$, the location halfspace median $\widehat{\boldsymbol{\mu}}_n^{\mathsf{hs}}$ is the minimax optimal estimator of $\boldsymbol{\mu}$ in terms of the squared Euclidean loss in Huber's contamination model $H_\varepsilon^{\mathsf{loc}}$. The minimax optimal convergence rate is $d/n \quad \varepsilon^2$.*

Proof. From Theorem 58, we know that the minimax optimal convergence rate is bounded from below by $(d/n) \quad \varepsilon^2$. The statement follows after we prove that this rate is indeed attained for the location halfspace median. However, this is an immediate consequence of Theorem 59 after we recognize the subsequent two observations:

1. The following inequality holds

$$\begin{aligned}\left(\left(\frac{d}{n} \quad \varepsilon^2\right) + \frac{\log(1/\delta)}{n}\right) &\leq \left(\left(\frac{d}{n} \quad \varepsilon^2\right) + \frac{d\,\log(1/\delta)}{n}\right) \\ &\leq (1 + \log(1/\delta))\left(\frac{d}{n} \quad \varepsilon^2\right) = \widetilde{C}\left(\frac{d}{n} \quad \varepsilon^2\right)\end{aligned}$$

for an absolute constant $\widetilde{C} = (1 + \log(1/\delta)) > 0$ independent of n, d and ε. Therefore, by Theorem 59

$$\begin{aligned}
&\inf_{P_{(\mu,\mathbf{\Sigma},Q)}} \inf_{H_\varepsilon^{\mathsf{loc}}} P_{(\mu,\mathbf{\Sigma},Q)} \left(\left\| \widehat{\mu}_n^{\mathsf{hs}} - \mu \right\|^2 < C\, \widetilde{C} \left(\frac{d}{n} \quad \varepsilon^2 \right) \right) \\
&\qquad \geq \inf_{P_{(\mu,\mathbf{\Sigma},Q)}} \inf_{H_\varepsilon^{\mathsf{loc}}} P_{(\mu,\mathbf{\Sigma},Q)} \left(\left\| \widehat{\mu}_n^{\mathsf{hs}} - \mu \right\|^2 < C \left(\left(\frac{d}{n} \quad \varepsilon^2 \right) + \frac{\log(1/\delta)}{n} \right) \right) \\
&\qquad \geq 1 - 2\delta
\end{aligned}$$

holds.

2. By (3.23), Theorem 59 holds only for $n > (d - \log(\delta))/D$. However, this is not a problem. There are only finitely many $n \leq (d - \log(\delta))/D$. Therefore, when we look closely at the proof of Theorem 59, specifically at equation (3.35), we can always choose a larger constant $C > 0$ such that this inequality holds even for $n \leq (d - \log(\delta))/D$.

Note *As noted in [Chen et al., 2018, Remark 2.1], Theorem 59 remains valid for any $\varepsilon < 1/3 - c_0$, where c_0 can be an arbitrarily small constant. The threshold $1/3$ represents the maximum breakdown point for Tukey's median; see Theorem 26.*

To conclude, we have demonstrated that Tukey's median is a minimax optimal estimator in Huber's contamination model M^{loc}. This highlights another of its robustness properties, in addition to those introduced in Chapter 1. Specifically, the minimax optimality underlines Tukey's median's effectiveness in handling outliers and contaminated data, making it a highly reliable choice for statistical estimation in various real-world scenarios.

3.5 Optimality of scatter halfspace median

In this section, we consider the following parametric model

$$M^{\mathsf{cov}} = \{ N_d(\mathbf{0}, \mathbf{\Sigma}) \mid \mathbf{\Sigma} \quad \mathsf{PD}_d, \sigma_{max}(\mathbf{\Sigma}) \leq M \},$$

which consists of all centered Gaussian distributions with a covariance matrix whose spectrum is bounded by some fixed constant $M > 0$. The parameter of interest is the covariance matrix $\theta = \mathbf{\Sigma}$, hence the superscript "cov" in M^{cov}. We have

$$\Theta = \{ \mathbf{\Sigma} \quad \mathsf{PD}_d \mid \sigma_{max}(\mathbf{\Sigma}) \leq M \}.$$

The error of an estimator is measured by the squared spectral norm

$$L(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) = \left\| \mathbf{\Sigma}_1 - \mathbf{\Sigma}_2 \right\|_{\mathsf{op}}^2.$$

Same as the squared Euclidean norm, this loss function also satisfies the τ-triangle inequality for $\tau = 2$ (see inequality (3.3)). The following theorem presents various methods to determine L.

Theorem 61 *For any* $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2 \quad \mathsf{PD}_d$ *it holds that*

$$\mathbf{\Gamma}_1 - \mathbf{\Gamma}_2 \ _{op} = \max\{|\lambda_1|, \ldots, |\lambda_d|\} = \max_{u \ \mathsf{S}^{d-1}} \left| u^\mathsf{T} (\mathbf{\Gamma}_1 - \mathbf{\Gamma}_2) u \right|,$$

where $\lambda_1, \ldots, \lambda_d$ *are the eigenvalues of* $\mathbf{\Gamma}_1 - \mathbf{\Gamma}_2$.

Proof. We know that the spectral norm of a matrix is equal to its largest singular value. The matrix $\mathbf{\Gamma}_1 - \mathbf{\Gamma}_2$ is the difference between two symmetric matrices. Hence, it is symmetric. For a symmetric matrix, the singular values are the absolute values of its eigenvalues, which gives us the first equality.

To prove the second equality, we decompose

$$\mathbf{\Gamma}_1 - \mathbf{\Gamma}_2 = \mathbf{Q}^\mathsf{T} \mathbf{\Lambda} \mathbf{Q},$$

where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{\Lambda}$ is a diagonal matrix with the eigenvalues of $\mathbf{\Gamma}_1 - \mathbf{\Gamma}_2$ on the diagonal. Using this decomposition, we can calculate

$$\begin{aligned}
\max_{u \ \mathsf{S}^{d-1}} \left| u^\mathsf{T} (\mathbf{\Gamma}_1 - \mathbf{\Gamma}_2) u \right| &= \max_{u \ \mathsf{S}^{d-1}} \left| u^\mathsf{T} \mathbf{Q}^\mathsf{T} \mathbf{\Lambda} \mathbf{Q} u \right| \\
&= \max_{v \ \mathsf{S}^{d-1}} \left| v^\mathsf{T} \mathbf{\Lambda} v \right| \\
&= \max_{v \ \mathsf{S}^{d-1}} \left| \sum_{i=1}^{d} \lambda_i v_i^2 \right| \\
&\leq \max\{|\lambda_1|, \ldots, |\lambda_d|\} \max_{v \ \mathsf{S}^{d-1}} \left| \sum_{i=1}^{d} v_i^2 \right| \\
&= \max\{|\lambda_1|, \ldots, |\lambda_d|\}.
\end{aligned}$$

The second equality above holds because the mapping $u \to Qu$ maps S^{d-1} onto S^{d-1}. Note that equality is achieved when v is the canonical vector corresponding to the largest eigenvalue in absolute value. This proves the theorem.

The corresponding Huber's contamination model for M^{cov} is

$$H_\varepsilon^{\mathsf{cov}} = \left\{ (1-\varepsilon) N_d(\mathbf{0}, \mathbf{\Sigma}) + \varepsilon Q \mid \mathbf{\Sigma} \quad \mathsf{PD}_d, \sigma_{max}(\mathbf{\Sigma}) \leq M, Q \quad P(\mathsf{R}^d) \right\}.$$

In contrast to the previous section, here we use $P_{\mathbf{\Sigma}}$ to represent the probability distribution $N_d(\mathbf{0}, \mathbf{\Sigma})$, and $P_{(\mathbf{\Sigma},Q)}$ to denote the contaminated probability distribution $(1-\varepsilon)N_d(\mathbf{0}, \mathbf{\Sigma}) + \varepsilon Q$. The contamination level ε is not explicitly stated, as it will be clear from the context, just like before. Recall that we denote by $hD^{\mathsf{sc}}(\cdot;\cdot)$ the scatter halfspace depth with the fixed location $\mathbf{0}$, that is

$$hD^{\mathsf{sc}}(\mathbf{\Gamma}; P) = \inf_{u \ \mathsf{S}^{d-1}} \min\left\{ P(H_{u,\mathbf{\Gamma}}^{\mathsf{in}}), P(H_{u,\mathbf{\Gamma}}^{\mathsf{out}}) \right\}$$

for any $\mathbf{\Gamma} \quad \mathsf{PD}_d$ and $P \quad P(\mathsf{R}^d)$, where $H_{u,\mathbf{\Gamma}}^{\mathsf{in}}$ and $H_{u,\mathbf{\Gamma}}^{\mathsf{out}}$ are defined in (2.3). Estimators of $\mathbf{\Sigma}$ are denoted by $\widehat{\mathbf{\Sigma}}_n$. The deepest matrix is denoted by $\widehat{\mathbf{\Sigma}}_n^{\mathsf{hs}}$, i.e.

$$\widehat{\mathbf{\Sigma}}_n^{\mathsf{hs}} = \arg\max_{\mathbf{\Sigma} \ \mathsf{PD}_d} hD^{\mathsf{sc}}(\mathbf{\Gamma}; \widehat{P}_n)$$

and by $\widehat{\mathbf{\Sigma}}_n^{\mathsf{cov}}$ we denote the estimator of covariance matrix

$$\widehat{\mathbf{\Sigma}}_n^{\mathsf{cov}} = \frac{1}{\beta}\widehat{\mathbf{\Sigma}}_n^{\mathsf{sc}},$$

where $\beta > 0$ solves $\Phi\left(\sqrt{\beta}\right) = 3/4$, see Example 6.

First, we focus on establishing a lower bound for the minimax optimal convergence rate in the model M^{cov}.

Theorem 62 *In model M^{cov}, there exists an absolute constant $\delta > 0$ such that*

$$\inf_n \inf_{\widehat{\mathbf{\Sigma}}_n} \sup_{P_{\mathbf{\Sigma}} \in M^{\mathsf{cov}}} P_{\mathbf{\Sigma}}\left(\left\|\widehat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\right\|_{op}^2 \geq R_n^0\right) \geq \delta$$

holds for some $\{R_n^0\}$ d/n. The infimum is taken over all estimators $\widehat{\mathbf{\Sigma}}_n$ of $\mathbf{\Sigma}$ based on a random sample $X_1, \ldots, X_n$ $P_{\mathbf{\Sigma}}$.

Proof. This claim follows from the proof of Theorem 6 and Remark 2 in [Ma and Wu, 2015]. Note that the spectral norm $\|\cdot\|_{\mathsf{op}}$ is a unitarily invariant norm in terms of this paper, which is assumed by that theorem.

As in the section on the optimality of the location halfspace median, we can use this to determine a lower bound for the minimax optimal convergence rate in model $H_\varepsilon^{\mathsf{cov}}$.

Theorem 63 *There exists an absolute constant $\delta > 0$ such that for any ε $[0, 1)$*

$$\inf_n \inf_{\widehat{\mathbf{\Sigma}}_n} \sup_{P_{(\mathbf{\Sigma},Q)} \in H_\varepsilon^{\mathsf{cov}}} P_{(\mathbf{\Sigma},Q)}\left(\left\|\widehat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\right\|_{op}^2 \geq R_n^\varepsilon\right) \geq \delta$$

holds for some $\{R_n^\varepsilon\}$ (d/n) ε^2. The infimum is taken over all estimators $\widehat{\mathbf{\Sigma}}_n$ of $\mathbf{\Sigma}$ based on a random sample $X_1, \ldots, X_n$ $P_{(\mathbf{\Sigma},Q)}$.

Proof. Same as in the proof of Theorem 58, it suffices to bound the modulus of continuity $\omega(\varepsilon, \Theta)$ from below by ε^2 in order to use Theorem 55 and Theorem 62.

Without loss of generality, assume $M \geq 1 + \varepsilon$. As in the proof of [Chen et al., 2018, Theorem 3.2], consider two matrices $\mathbf{\Sigma}_a = \mathbf{I}_d$ and $\mathbf{\Sigma}_b = \mathbf{I}_d + \varepsilon\,\mathbf{M}$, where $\mathbf{M}$ is a matrix with the only non-zero entry $M_{(1,1)} = 1$. Then, obviously

$$\|\mathbf{\Sigma}_a - \mathbf{\Sigma}_b\|_{\mathsf{op}}^2 = \|\varepsilon\,\mathbf{M}\|_{\mathsf{op}}^2 = \varepsilon^2.$$

Using Pinsker's inequality given in Theorem 51 and the Kullback-Leibler di-

vergence of two Gaussian distributions (see Theorem 52), we can bound

$$\begin{aligned}
\mathsf{TV}\left(P_{\mathbf{\Sigma}_a}, P_{\mathbf{\Sigma}_b}\right)^2 &\leq \frac{1}{2} D(P_{\mathbf{\Sigma}_a} || P_{\mathbf{\Sigma}_b}) \\
&= \frac{1}{4}\left(\log\left(\frac{|\mathbf{\Sigma}_b|}{|\mathbf{\Sigma}_a|}\right) + \mathsf{tr}\left(\mathbf{\Sigma}_b^{-1}\mathbf{\Sigma}_a\right) - d\right) \\
&= \frac{1}{4}\left(\log(1+\varepsilon) + \frac{1}{1+\varepsilon} + (d-1) - d\right) \\
&= \frac{1}{4}\left(\log(1+\varepsilon) - \frac{\varepsilon}{1+\varepsilon}\right) \\
&\leq \frac{1}{4}\left(\varepsilon - \frac{\varepsilon}{1+\varepsilon}\right) \\
&= \frac{\varepsilon^2}{4(1+\varepsilon)} \leq \left(\frac{\varepsilon}{1-\varepsilon}\right)^2.
\end{aligned}$$

In the second inequality, we used the fact that $\log(1+x) \leq x$ for $x > 0$, and in the last inequality, we used the assumption $\varepsilon \in [0,1)$. Therefore

$$\mathsf{TV}\left(P_{\mathbf{\Sigma}_a}, P_{\mathbf{\Sigma}_b}\right) \leq \varepsilon/(1-\varepsilon).$$

As a consequence, the modulus of continuity can be bounded from below as follows

$$\begin{aligned}
\omega(\varepsilon, \Theta) &= \sup\left\{ \left\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\right\|_{\mathsf{op}}^2 \mid \mathsf{TV}\left(P_{\mathbf{\Sigma}_1}, P_{\mathbf{\Sigma}_2}\right) \leq \frac{\varepsilon}{1-\varepsilon}; \mathbf{\Sigma}_1, \mathbf{\Sigma}_2 \in \Theta \right\} \\
&\geq \left\|\mathbf{\Sigma}_a - \mathbf{\Sigma}_b\right\|_{\mathsf{op}} = \varepsilon^2.
\end{aligned}$$

This concludes the proof.

Now, we show that the convergence rate $(d/n) \wedge \varepsilon^2$ in model $H_\varepsilon^{\mathsf{cov}}$ is indeed achieved for the scatter halfspace median.

Theorem 64 *Assume that $\varepsilon \in (0, 1/5)$. Then, for any $\delta \in (0, 1/2)$ there exist absolute constants $C, D > 0$ such that for all $n \in \mathsf{N}$ satisfying*

$$\left(\frac{d}{n} + \frac{\log(1/\delta)}{n}\right) < D \tag{3.36}$$

we have

$$\inf_{P_{(\mathbf{\Sigma},Q)}} \inf_{H_\varepsilon^{\mathsf{loc}}} P_{(\mathbf{\Sigma},Q)}\left(\left\|\widehat{\mathbf{\Sigma}}_n^{\mathsf{cov}} - \mathbf{\Sigma}\right\|_{op}^2 < C\left(\left(\frac{d}{n} \wedge \varepsilon^2\right) + \frac{\log(1/\delta)}{n}\right)\right) \geq 1 - 2\delta,$$

hence

$$\sup_{P_{(\mathbf{\Sigma},Q)}} \inf_{H_\varepsilon^{\mathsf{loc}}} P_{(\mathbf{\Sigma},Q)}\left(\left\|\widehat{\mathbf{\Sigma}}_n^{\mathsf{cov}} - \mathbf{\Sigma}\right\|_{op}^2 \geq C\left(\left(\frac{d}{n} \wedge \varepsilon^2\right) + \frac{\log(1/\delta)}{n}\right)\right) < 2\delta.$$

Proof. This proof is largely inspired by [Chen et al., 2018, Theorem 3.1]. It closely resembles the proof of Theorem 59, offering more precise details and clarifying some ambiguities present in [Chen et al., 2018]. As before, the proof is structured into four parts.

Part 1: Auxiliary observations

First, we make the following observations, which will be useful in deriving the intended bound.

(S1) By Example 6, for any $\mathbf{\Sigma}, \mathbf{\Gamma} \in \mathsf{PD}_d$ and $u \in \mathsf{S}^{d-1}$ we have

$$P_{\mathbf{\Sigma}}(H^{\mathsf{in}}_{u,\mathbf{\Gamma}}) = 1 - P_{\mathbf{\Sigma}}(H^{\mathsf{out}}_{u,\mathbf{\Gamma}}) = 2\left(\Phi\left(\sqrt{\frac{u^\top \mathbf{\Gamma} u}{u^\top \mathbf{\Sigma} u}}\right) - \frac{1}{2}\right)$$

and

$$\max_{\mathbf{\Gamma} \in \mathsf{PD}_d} hD^{\mathsf{sc}}(\mathbf{\Gamma}; P_{\mathbf{\Sigma}}) = hD^{\mathsf{sc}}(\beta\mathbf{\Sigma}; P_{\mathbf{\Sigma}}) = \frac{1}{2}.$$

(S2) Given a random sample $\{X_i\}_{i=1}^n$ from a contaminated distribution $P_{(\mathbf{\Sigma},Q)}$, we can decompose $\{X_i\}_{i=1}^n = \{Y_i\}_{i=1}^{n_1} \cup \{Z_i\}_{i=1}^{n_2}$. Here, marginally, it holds that $n_2 \sim \mathsf{Binomial}(n, \varepsilon)$ and $n_1 = n - n_2$. Conditionally on n_1 and n_2, $\{Y_i\}_{i=1}^{n_1}$ is a random sample from $P_{\mathbf{\Sigma}}$, while $\{Z_i\}_{i=1}^{n_2}$ is a random sample from Q.

(S3) By Theorem A3 in the Appendix, we have with probability at least $1 - \delta$ that

$$\begin{aligned}
&\sup_{\mathbf{\Gamma} \in \mathsf{PD}_d} \left| hD^{\mathsf{sc}}(\mathbf{\Gamma}; P_{\mathbf{\Sigma}}) - hD^{\mathsf{sc}}(\mathbf{\Gamma}; \{Y_i\}_{i=1}^{n_1}) \right| \\
&\le \sup_{\mathbf{\Gamma} \in \mathsf{PD}_d, u \in \mathsf{S}^{d-1}} \max\left\{ \left| P_{\mathbf{\Sigma}}(H^{\mathsf{in}}_{u,\mathbf{\Gamma}}) - \widehat{P}_{n_1}(H^{\mathsf{in}}_{u,\mathbf{\Gamma}}) \right|, \left| P_{\mathbf{\Sigma}}(H^{\mathsf{out}}_{u,\mathbf{\Gamma}}) - \widehat{P}_{n_1}(H^{\mathsf{out}}_{u,\mathbf{\Gamma}}) \right| \right\} \\
&\le \sup_{u \in \mathsf{S}^{d-1}, t \ge 0} \max\left\{ \left| P_{\mathbf{\Sigma}}(H^{\mathsf{in}}_{u,t}) - \widehat{P}_{n_1}(H^{\mathsf{in}}_{u,t}) \right|, \left| P_{\mathbf{\Sigma}}(H^{\mathsf{out}}_{u,t}) - \widehat{P}_{n_1}(H^{\mathsf{out}}_{u,t}) \right| \right\} \\
&\le \sqrt{\frac{1440\pi e}{1 - e^{-1}}} \sqrt{\frac{2d+3}{n_1}} + \sqrt{\frac{\log(1/\delta)}{2n_1}},
\end{aligned}$$

where $\widehat{P}_{n_1}$ is the empirical distribution of $\{Y_i\}_{i=1}^{n_1}$.

(S4) By the definition of the sample scatter halfspace depth, it follows that for any $\mathbf{\Gamma} \in \mathsf{PD}_d$, we have

$$n_1\, hD^{\mathsf{sc}}(\mathbf{\Gamma}; \{Y_i\}_{i=1}^{n_1}) \ge n\, hD^{\mathsf{sc}}(\mathbf{\Gamma}; \{X_i\}_{i=1}^{n}) - n_2 \ge n_1\, hD^{\mathsf{sc}}(\mathbf{\Gamma}; \{Y_i\}_{i=1}^{n_1}) - n_2.$$

This can be proven exactly as (3.25) in the proof of Theorem 59.

(S5) By Theorem A4 in the Appendix, if

$$\sqrt{\frac{\log(1/\delta)}{2\,n}} < 1/5 \tag{3.37}$$

holds for $\delta > 0$, then

$$\frac{n_2}{n_1} \le \frac{\varepsilon}{1 - \varepsilon} + \frac{25}{12}\sqrt{\frac{\log(1/\delta)}{2\,n}} < \frac{2}{3}$$

holds with probability at least $1 - \delta$.

Part 2: Choice of constant D

To ensure the proof of the statement is valid, we need to define a constant D. Specifically, it needs to satisfy both (3.37) and the following condition

$$20\sqrt{\frac{30\pi e}{1 - e^{-1}}}\sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(1/\delta)}{n}} < \frac{1}{16}, \tag{3.38}$$

which will be useful later. For the remainder of the proof, fix δ $(0, 1/2)$ and consider n $\mathbb{N}$ such that the assumption (3.36) holds with

$$D = \left(\frac{\frac{1}{16}}{20\sqrt{\frac{30\pi e}{1-e^{-1}}} + 2}\right)^2 > 0.$$

Since $a + b < D$ implies a $b < D$ for all $a, b > 0$, it follows that

$$\sqrt{\frac{d}{n}} \quad \sqrt{\frac{\log(1/\delta)}{n}} < \sqrt{D}. \tag{3.39}$$

Therefore, by our choice of D and using (3.39), we can bound

$$20\sqrt{\frac{30\pi e}{1 - e^{-1}}}\sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(1/\delta)}{n}} < \left(20\sqrt{\frac{30\pi e}{1 - e^{-1}}} + 2\right)\sqrt{D} = \frac{1}{16}.$$

Consequently, inequality (3.38) holds. Additionally, (3.38) implies that

$$\sqrt{\frac{\log(1/\delta)}{2n}} < 2\sqrt{\frac{\log(1/\delta)}{n}} < \frac{1}{16} < \frac{1}{5},$$

thus, under the assumption $\varepsilon < 1/5$, (S5) gives that

$$\frac{n_2}{n_1} \leq \frac{\varepsilon}{1 - \varepsilon} + \frac{25}{12}\sqrt{\frac{\log(1/\delta)}{2n}} < \frac{2}{3} \tag{3.40}$$

holds with probability at least $1 - \delta$. Furthermore, note that

$$\frac{n_2}{n_1} < \frac{2}{3} \qquad 3n_2 < 2n_1 \qquad 3(n - n_1) < 2n_1 \qquad n_1 > 3n/5 \tag{3.41}$$

holds, much like before in the proof of Theorem 59. To summarize, for our fixed $\delta > 0$ and a sufficiently large n $\mathbb{N}$, at least $3/5$ of all observations are non-contaminating with probability at least $1 - \delta$. Moreover, inequality (3.38) holds. Both of these facts will be used later in the proof.

Part 3: Bound for the depth of $\widehat{\Sigma}_n^{\mathsf{hs}}$

Based on the earlier observations and our selection of D, we establish the following sequence of inequalities. These are valid with a probability of at least $1 - \delta$ given

the marginal counts n_1, n_2 for all $\mathbf{\Sigma} \in \Theta$. We have

$$\begin{aligned}
hD^{\mathsf{sc}}(\widehat{\mathbf{\Sigma}}_n^{\mathsf{hs}}; P_{\mathbf{\Sigma}}) &\overset{(S3)}{\geq} hD^{\mathsf{sc}}(\widehat{\mathbf{\Sigma}}_n^{\mathsf{hs}}; \{\mathbf{Y}_i\}_{i=1}^{n_1}) - \sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{2d+3}{n_1}} - \sqrt{\frac{\log(1/\delta)}{2n_1}} \\
&\overset{(S4)}{\geq} \frac{n}{n_1} hD^{\mathsf{sc}}(\widehat{\mathbf{\Sigma}}_n^{\mathsf{hs}}; \{\mathbf{X}_i\}_{i=1}^{n}) - \frac{n_2}{n_1} - \sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{2d+3}{n_1}} - \sqrt{\frac{\log(1/\delta)}{2n_1}} \\
&\geq \frac{n}{n_1} hD^{\mathsf{sc}}(\beta\mathbf{\Sigma}; \{\mathbf{X}_i\}_{i=1}^{n}) - \frac{n_2}{n_1} - \sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{2d+3}{n_1}} - \sqrt{\frac{\log(1/\delta)}{2n_1}} \\
&\overset{(S4)}{\geq} hD^{\mathsf{sc}}(\beta\mathbf{\Sigma}; \{\mathbf{Y}_i\}_{i=1}^{n_1}) - \frac{n_2}{n_1} - \sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{2d+3}{n_1}} - \sqrt{\frac{\log(1/\delta)}{2n_1}} \\
&\overset{(S3)}{\geq} hD^{\mathsf{sc}}(\beta\mathbf{\Sigma}; P_{\mathbf{\Sigma}}) - \frac{n_2}{n_1} - 2\sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{2d+3}{n_1}} - \sqrt{\frac{2\log(1/\delta)}{n_1}} \\
&\overset{(S1)}{=} \frac{1}{2} - \frac{n_2}{n_1} - 2\sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{2d+3}{n_1}} - \sqrt{\frac{2\log(1/\delta)}{n_1}}. \qquad (3.42)
\end{aligned}$$

The third inequality follows because $\widehat{\mathbf{\Sigma}}_n^{\mathsf{hs}}$ is the maximizer of $hD^{\mathsf{sc}}(\cdot; \{\mathbf{X}_i\}_{i=1}^n)$. Rewriting (3.42), we have that

$$\mathsf{P}\left[\frac{1}{2} - hD^{\mathsf{sc}}(\widehat{\mathbf{\Sigma}}_n^{\mathsf{hs}}; P_{\mathbf{\Sigma}}) \leq \frac{n_2}{n_1} + 2\sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{2d+3}{n_1}} + \sqrt{\frac{2\log(1/\delta)}{n_1}} \,\middle|\, n_1, n_2\right] \geq 1-\delta.$$

However, by taking the expectation on both sides (and considering its monotonicity), we obtain

$$\mathsf{P}\left[\frac{1}{2} - hD^{\mathsf{sc}}(\widehat{\mathbf{\Sigma}}_n^{\mathsf{hs}}; P_{\mathbf{\Sigma}}) \leq \frac{n_2}{n_1} + 2\sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{2d+3}{n_1}} + \sqrt{\frac{2\log(1/\delta)}{n_1}}\right] \geq 1-\delta.$$

Using (S1), this means that uniformly over all $\boldsymbol{u} \in \mathsf{S}^{d-1}$ the inequality

$$\begin{aligned}
&\frac{1}{2} - \min\left\{P_{\mathbf{\Sigma}}\left(H^{\mathsf{in}}_{\boldsymbol{u},\widehat{\mathbf{\Sigma}}_n^{\mathsf{hs}}}\right), 1 - P_{\mathbf{\Sigma}}\left(H^{\mathsf{in}}_{\boldsymbol{u},\widehat{\mathbf{\Sigma}}_n^{\mathsf{hs}}}\right)\right\} \\
&\qquad\qquad \leq \frac{n_2}{n_1} + 2\sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{2d+3}{n_1}} + \sqrt{\frac{2\log(1/\delta)}{n_1}} \qquad (3.43)
\end{aligned}$$

holds with probability at least $1-\delta$. Note that for all $x \in \mathsf{R}$ the equality $1/2 - \min\{x, 1-x\} = |1/2 - x|$ holds. Therefore, for all $\boldsymbol{u} \in \mathsf{S}^{d-1}$ we have

$$\begin{aligned}
\frac{1}{2} - \min\left\{P_{\mathbf{\Sigma}}\left(H^{\mathsf{in}}_{\boldsymbol{u},\widehat{\mathbf{\Sigma}}_n^{\mathsf{hs}}}\right), 1 - P_{\mathbf{\Sigma}}\left(H^{\mathsf{in}}_{\boldsymbol{u},\widehat{\mathbf{\Sigma}}_n^{\mathsf{hs}}}\right)\right\} &= \left|\frac{1}{2} - P_{\mathbf{\Sigma}}\left(H^{\mathsf{in}}_{\boldsymbol{u},\widehat{\mathbf{\Sigma}}_n^{\mathsf{hs}}}\right)\right| \\
&\overset{(S1)}{=} \left|P_{\mathbf{\Sigma}}(H^{\mathsf{in}}_{\boldsymbol{u},\beta\mathbf{\Sigma}}) - P_{\mathbf{\Sigma}}\left(H^{\mathsf{in}}_{\boldsymbol{u},\widehat{\mathbf{\Sigma}}_n^{\mathsf{hs}}}\right)\right| \\
&\overset{(S1)}{=} 2\left|\Phi\left(\sqrt{\beta}\right) - \Phi\left(\sqrt{\frac{\boldsymbol{u}^\mathsf{T}\widehat{\mathbf{\Sigma}}_n^{\mathsf{hs}}\boldsymbol{u}}{\boldsymbol{u}^\mathsf{T}\mathbf{\Sigma}\boldsymbol{u}}}\right)\right|. \qquad (3.44)
\end{aligned}$$

Combining (3.43) with (3.44), we have that

$$\sup_{\boldsymbol{u} \in \mathsf{S}^{d-1}}\left|\Phi\left(\sqrt{\beta}\right) - \Phi\left(\sqrt{\frac{\boldsymbol{u}^\mathsf{T}\widehat{\mathbf{\Sigma}}_n^{\mathsf{hs}}\boldsymbol{u}}{\boldsymbol{u}^\mathsf{T}\mathbf{\Sigma}\boldsymbol{u}}}\right)\right| \leq \frac{n_2}{2n_1} + \sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{2d+3}{n_1}} + \sqrt{\frac{\log(1/\delta)}{2n_1}}$$

holds with probability at least $1 - \delta$. Now, we combine this result with inequality (3.40). Same as in the proof of Theorem 59, we obtain

$$\mathsf{P}\left[\sup_{u \in \mathbb{S}^{d-1}} \left|\Phi\left(\sqrt{\beta}\right) - \Phi\left(\sqrt{\frac{u^\top \widehat{\Sigma}_n^{\mathsf{hs}} u}{u^\top \Sigma u}}\right)\right| \le \frac{n_2}{2\,n_1} + \sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{2d+3}{n_1}} + \sqrt{\frac{\log(1/\delta)}{2\,n_1}}, \right.$$
$$\left. \frac{n_2}{n_1} \le \frac{\varepsilon}{1-\varepsilon} + \frac{25}{12}\sqrt{\frac{\log(1/\delta)}{2n}} \le \frac{2}{3}\right] \ge 1 - 2\delta. \quad (3.45)$$

Under the condition of the second random event in (3.45), we can further upper bound

$$\begin{aligned}
\sup_{u \in \mathbb{S}^{d-1}} \left|\Phi\left(\sqrt{\beta}\right) - \Phi\left(\sqrt{\frac{u^\top \widehat{\Sigma}_n^{\mathsf{hs}} u}{u^\top \Sigma u}}\right)\right| &\overset{(3.45)}{\le} \frac{n_2}{2\,n_1} + \sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{2d+3}{n_1}} + \sqrt{\frac{\log(1/\delta)}{2\,n_1}} \\
&\overset{(3.40)}{\le} \frac{\varepsilon}{2(1-\varepsilon)} + \frac{25}{24}\sqrt{\frac{\log(1/\delta)}{2n}} + \sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{2d+3}{n_1}} + \sqrt{\frac{\log(1/\delta)}{2\,n_1}} \\
&\overset{(3.41)}{\le} \frac{\varepsilon}{2(1-\varepsilon)} + \frac{25}{24}\sqrt{\frac{\log(1/\delta)}{2n}} + \sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{5(2d+3)}{3n}} + \sqrt{\frac{5\log(1/\delta)}{6n}} \\
&\overset{d\ge 1}{\le} \frac{\varepsilon}{2(1-\varepsilon)} + \frac{25}{24}\sqrt{\frac{\log(1/\delta)}{2n}} + \sqrt{\frac{1440\pi e}{1-e^{-1}}}\sqrt{\frac{25d}{3n}} + \sqrt{\frac{5\log(1/\delta)}{6n}} \\
&= \frac{\varepsilon}{2(1-\varepsilon)} + 20\sqrt{\frac{30\pi e}{1-e^{-1}}}\sqrt{\frac{d}{n}} + \left(\frac{25}{24}\sqrt{\frac{1}{2}} + \sqrt{\frac{5}{6}}\right)\sqrt{\frac{\log(1/\delta)}{n}} \\
&< \frac{\varepsilon}{2(1-\varepsilon)} + 20\sqrt{\frac{30\pi e}{1-e^{-1}}}\sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(1/\delta)}{n}} \\
&\overset{\varepsilon<1/5}{<} \frac{5}{8}\varepsilon + 20\sqrt{\frac{30\pi e}{1-e^{-1}}}\sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(1/\delta)}{n}}.
\end{aligned}$$

In the first strict inequality, we used the fact that

$$\left(\frac{25}{24}\sqrt{\frac{1}{2}} + \sqrt{\frac{5}{6}}\right) < 2.$$

Ultimately, we have

$$\mathsf{P}\left[\sup_{u \in \mathbb{S}^{d-1}} \left|\Phi\left(\sqrt{\beta}\right) - \Phi\left(\sqrt{\frac{u^\top \widehat{\Sigma}_n^{\mathsf{hs}} u}{u^\top \Sigma u}}\right)\right| \le \frac{5}{8}\varepsilon + 20\sqrt{\frac{30\pi e}{1-e^{-1}}}\sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(1/\delta)}{n}}, \right.$$
$$\left. \frac{n_2}{n_1} \le \frac{\varepsilon}{1-\varepsilon} + \frac{25}{12}\sqrt{\frac{\log(1/\delta)}{2n}} \le \frac{2}{3}\right] \ge 1 - 2\delta.$$

This further implies (by the same argument as in the proof of Theorem 59) that

$$\sup_{u \in \mathbb{S}^{d-1}} \left|\Phi\left(\sqrt{\beta}\right) - \Phi\left(\sqrt{\frac{u^\top \widehat{\Sigma}_n^{\mathsf{hs}} u}{u^\top \Sigma u}}\right)\right| \le \frac{5}{8}\varepsilon + 20\sqrt{\frac{30\pi e}{1-e^{-1}}}\sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(1/\delta)}{n}} \quad (3.46)$$

holds with probability at least $1 - 2\delta$.

Part 4: Conclusion

We assumed that n is chosen such that (3.38) holds. Therefore, using this and the assumption $\varepsilon < 1/5$, from (3.46) we can deduce that for all $\boldsymbol{\Sigma} \quad \Theta$ it holds that

$$\sup_{\boldsymbol{u} \ \mathbb{S}^{d-1}} \left| \Phi\left(\sqrt{\beta}\right) - \Phi\left(\sqrt{\frac{\boldsymbol{u}^\mathsf{T} \widehat{\boldsymbol{\Sigma}}_n^{\mathsf{hs}} \boldsymbol{u}}{\boldsymbol{u}^\mathsf{T} \boldsymbol{\Sigma} \boldsymbol{u}}}\right) \right| < \frac{5}{8}\varepsilon + \frac{1}{16} < \frac{1}{8} + \frac{1}{16} = \frac{3}{16}. \tag{3.47}$$

Now we can use the fact that $\Phi(\sqrt{\beta}) = 3/4$ and that the quantile function of the standard Gaussian distribution is Lipschitz continuous on the interval

$$\left[\frac{3}{4} - \frac{3}{16}, \frac{3}{4} + \frac{3}{16}\right] = \left[\frac{9}{16}, \frac{15}{16}\right]$$

with the Lipschitz constant

$$C = \left[\varphi\left(\Phi^{-1}(15/16)\right)\right]^{-1} > 0.$$

Therefore, from (3.46) we can further deduce

$$\sup_{\boldsymbol{u} \ \mathbb{S}^{d-1}} \left| \sqrt{\beta} - \sqrt{\frac{\boldsymbol{u}^\mathsf{T} \widehat{\boldsymbol{\Sigma}}_n^{\mathsf{hs}} \boldsymbol{u}}{\boldsymbol{u}^\mathsf{T} \boldsymbol{\Sigma} \boldsymbol{u}}} \right| < C \left(\frac{5}{8}\varepsilon + 20\sqrt{\frac{30\pi e}{1 - e^{-1}}}\sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{3.48}$$

In contrast to the proof of Theorem 59, we need to apply the trick involving Lipschitz continuity once again. Specifically, we observe that the upper bound of the right side of (3.48) can be constrained similarly to (3.47) by $3C/16 > 0$. The function $x \to x^2$ is Lipschitz continuous on the interval

$$\left[\sqrt{\beta} - 3C/16, \sqrt{\beta} + 3C/16\right]$$

with the Lipschitz constant

$$C = 2\left(\sqrt{\beta} + 3C/16\right) > 0.$$

As a consequence, with probability at least $1 - 2\delta$, we have that

$$\sup_{\boldsymbol{u} \ \mathbb{S}^{d-1}} \left| \beta - \frac{\boldsymbol{u}^\mathsf{T} \widehat{\boldsymbol{\Sigma}}_n^{\mathsf{hs}} \boldsymbol{u}}{\boldsymbol{u}^\mathsf{T} \boldsymbol{\Sigma} \boldsymbol{u}} \right| < C \ C \left(\frac{5}{8}\varepsilon + 20\sqrt{\frac{30\pi e}{1 - e^{-1}}}\sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{3.49}$$

We can finally establish the required bound by summarizing all the arguments

above. With probability at least $1 - \delta$, for all $\Sigma \in \Theta$ it holds that

$$\begin{aligned}
\left\|\widehat{\Sigma}_n^{\text{cov}} - \Sigma\right\|_{\text{op}}^2 &\overset{\text{Thm. 61}}{=} \sup_{u \in S^{d-1}} \left|u^T(\Sigma - \widehat{\Sigma}_n^{\text{cov}})u\right|^2 \\
&= \sup_{u \in S^{d-1}} \left|u^T\Sigma u - u^T\widehat{\Sigma}_n^{\text{hs}}/\beta u\right|^2 = \sup_{u \in S^{d-1}} \frac{M^2}{\beta^2}\left|\frac{u^T\beta\Sigma u - u^T\widehat{\Sigma}_n^{\text{hs}}u}{M}\right|^2 \\
&\overset{\Sigma \in \Theta}{\leq} \sup_{u \in S^{d-1}} \frac{M^2}{\beta^2}\left|\frac{u^T\beta\Sigma u - u^T\widehat{\Sigma}_n^{\text{hs}}/u}{u^T\Sigma u}\right|^2 = \sup_{u \in S^{d-1}} \frac{M^2}{\beta^2}\left|\beta - \frac{u^T\widehat{\Sigma}_n^{\text{hs}}u}{u^T\Sigma u}\right|^2 \\
&\overset{(3.49)}{\leq} \left(\frac{M\,C\;C}{\beta}\right)^2\left(\frac{5}{8}\varepsilon + 20\sqrt{\frac{30\pi e}{1 - e^{-1}}}\sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(1/\delta)}{n}}\right)^2 \\
&\leq \left(\frac{M\,C\;C}{\beta}\right)^2\left(\left(\frac{5}{8} + 20\sqrt{\frac{30\pi e}{1 - e^{-1}}}\right)\left(\sqrt{\frac{d}{n}} \quad \varepsilon\right) + 2\sqrt{\frac{\log(1/\delta)}{n}}\right)^2 \\
&\overset{(3.3)}{\leq} 2\left(\frac{M\,C\;C}{\beta}\right)^2\left(\left(\frac{5}{8} + 20\sqrt{\frac{30\pi e}{1 - e^{-1}}}\right)^2\left(\frac{d}{n} \quad \varepsilon^2\right) + \frac{4\log(1/\delta)}{n}\right) \\
&\leq C\left(\left(\frac{d}{n} \quad \varepsilon^2\right) + \frac{\log(1/\delta)}{n}\right),
\end{aligned}$$

where

$$C = 2\left(\frac{M\,C\;C}{\beta}\right)^2\left(\left(\frac{5}{8} + 20\sqrt{\frac{30\pi e}{1 - e^{-1}}}\right)^2 + 4\right) > 0.$$

In the first inequality, we used the Cauchy-Schwarz inequality and the assumption that $\sigma_{max}(\Sigma) \leq M$. This concludes the proof.

Minimax optimality of the scatter halfspace median follows immediately.

Theorem 65 *For $\varepsilon < 1/5$, the scatter halfspace median $\widehat{\Sigma}_n^{\text{cov}}$ is the minimax optimal estimator of Σ in terms of the squared operator loss in Huber's contamination model $H_\varepsilon^{\text{cov}}$. The minimax optimal convergence rate is $d/n \quad \varepsilon^2$.*

Proof. This claim follows from Theorem 64 utilizing precisely the same arguments used in the proof of Theorem 60.

We have established that the scatter halfspace median is the minimax optimal estimator of the covariance matrix when the observations $X_1, \ldots, X_n$ are generated from a centered Gaussian distribution. Chen et al. [2018] provide two extensions. First, Theorem 65 can be extended to elliptically symmetric distributions, as shown in [Chen et al., 2018, Theorem 4.1]. Secondly, the assumption of a centered generating distribution can also be relaxed. The key idea is that if $X_1, \ldots, X_n \quad N_d(\mu, \Sigma)$, then for all $i, j \quad \{1, \ldots, n\}, i = j$,

$$Y_{ij} = \frac{1}{\sqrt{2}}(X_i - X_j) \quad N_d(0, \Sigma),$$

as follows from the basic properties of the Gaussian distribution. This motivates the use of the U-version of scatter depth. That is, we compute

$$\widehat{\mathbf{\Sigma}}_n^{\mathsf{hs},\mathsf{U}} = \underset{\mathbf{\Gamma}\ \mathsf{PD}_d}{\arg\max}\, hD^{\mathsf{sc}}\left(\mathbf{\Gamma}; \{Y_{ij}\}_{i<j}\right)$$

and define the estimator of $\mathbf{\Sigma}$ as

$$\widehat{\mathbf{\Sigma}}_n^{\mathsf{cov},\mathsf{U}} = \frac{1}{\beta}\widehat{\mathbf{\Sigma}}_n^{\mathsf{hs},\mathsf{U}}.$$

It turns out that the optimality results also hold for these U-estimators. For details, refer to Section A of the supplementary material in [Chen et al., 2018].

Conclusion

In this thesis, we examined the fundamental properties of location halfspace depth (Chapter 1) and scatter halfspace depth (Chapter 2), with an emphasis on the robustness of the corresponding halfspace medians. In Chapter 3, we introduced the minimax optimality of location and scatter halfspace medians. This allowed us to meet the objectives of the thesis.

We utilized tools from mathematical statistics, probability theory, measure theory, empirical processes, geometry, and combinatorics. The primary contribution of this thesis is the establishment of a unified framework for analyzing the convergence rates of estimators and the minimax optimality of estimators. We also identified and corrected several deficiencies in the literature, such as Theorem 23 (see Example 4), and Theorem 57. The proofs presented in this thesis include numerous additional details and clarify various ambiguities in the proofs from the original papers.

Future work could build on this thesis in several directions. For example, it would be beneficial to determine other measures of robustness for the scatter halfspace median, such as the breakdown point. Additionally, it remains an open question whether location and scatter halfspace medians are minimax optimal in Huber's contamination model when the contaminated distribution is not Gaussian or elliptically symmetric. Another interesting question is whether the scatter halfspace median would still be minimax optimal if we also had to non-trivially estimate the location parameter.

Bibliography

Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, third edition, 1995. A Wiley-Interscience Publication.

Tomáš Brabenec. Hloubka variančních matic. Master's thesis, Charles University, Faculty of Mathematics and Physics, Prague, 2021.

Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.*, 46(5):1932–1960, 2018.

Zhiqiang Chen and David E. Tyler. The influence function and maximum bias of Tukey's median. *Ann. Statist.*, 30(6):1737–1759, 2002.

Zhiqiang Chen and David E. Tyler. On the behavior of Tukey's depth and median under symmetric stable distributions. *J. Statist. Plann. Inference*, 122 (1-2):111–124, 2004.

David L. Donoho and Miriam Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.*, 20(4): 1803–1827, 1992.

David L. Donoho and Richard C. Liu. Geometrizing rates of convergence. II, III. *Ann. Statist.*, 19(2):633–667, 668–701, 1991.

Thomas P. Hettmansperger and Hannu Oja. Affine invariant multivariate multisample sign tests. *J. Roy. Statist. Soc. Ser. B*, 56(1):235–249, 1994.

Peter J. Huber and Elvezio M. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2009.

Rebecka Jörnsten. Clustering and classification based on the L_1 data depth. *J. Multivariate Anal.*, 90(1):67–89, 2004.

Petra Laketa and Stanislav Nagy. Halfspace depth for general measures: the ray basis theorem and its consequences. *Statist. Papers*, 63(3):849–883, 2022.

Gaëtan Louvet and Germain Van Bever. The influence function of scatter halfspace depth. In *Recent Advances in Econometrics and Statistics*. Springer Nature Switzerland, 2024.

Zongming Ma and Yihong Wu. Volume ratio, sparsity, and minimaxity under unitarily invariant norms. *IEEE Trans. Inform. Theory*, 61(12):6939–6956, 2015.

Karl Mosler and Pavlo Mozharovskyi. Choosing among notions of multivariate depth statistics. *Statist. Sci.*, 37(3):348–368, 2022.

Stanislav Nagy, Carsten Schütt, and Elisabeth M. Werner. Halfspace depth and floating body. *Stat. Surv.*, 13:52–118, 2019.

Davy Paindaveine and Germain Van Bever. Halfspace depths for scatter, concentration and shape matrices. *Ann. Statist.*, 46(6B):3276–3307, 2018.

Peter J. Rousseeuw and Ida Ruts. The depth function of a population distribution. *Metrika*, 49(3):213–244, 1999.

Jonathan Scarlett and Volkan Cevher. *An Introductory Guide to Fano's Inequality with Applications in Statistical Estimation*, pages 487—528. Cambridge University Press, Cambridge, 2021.

Robert Serfling. Depth functions in nonparametric multivariate inference. In *Data depth: robust multivariate analysis, computational geometry and applications*, volume 72 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 1–16. Amer. Math. Soc., Providence, RI, 2006.

Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

John W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians (Vancouver, B.C., 1974), Vol. 2*, pages 523–531. Canad. Math. Congr., Montreal, QC, 1975.

Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, New York, 1997.

Yijun Zuo and Robert Serfling. General notions of statistical depth function. *Ann. Statist.*, 28(2):461–482, 2000a.

Yijun Zuo and Robert Serfling. On the performance of some robust nonparametric location measures relative to a general notion of multivariate symmetry. *J. Statist. Plann. Inference*, 84(1-2):55–79, 2000b.

Auxiliary theorems

Theorem A1 (Lebesgue's dominated convergence theorem) *Let $(X, \mathcal{A}, \mu)$ be a space with measure. Consider measurable functions $\{f_n\}_{n=1}^{\infty}, f, g\colon X \to \mathbb{R}$. Assume that for all $n \in \mathbb{N}$ we have $|f_n| \le g$ μ-a.e., where g is integrable. If $f_n \to f$ μ-a.e. as $n \to \infty$, then $\{f_n\}_{n=1}^{\infty}, f$ are integrable and*

$$\int_X f_n(x)\,\mathrm{d}\mu(x) \to \int_X f(x)\,\mathrm{d}\mu(x) \quad \text{as } n \to \infty.$$

Proof. Refer to [Billingsley, 1995, Theorem 16.4].

Theorem A2 (Rate of convergence for halfspaces) *Let $P \in \mathcal{P}(\mathbb{R}^d)$ and consider the empirical distribution $\widehat{P}_n$ based on a random sample of size n from P. Then for all $\delta \in (0, 1/2)$ the inequality*

$$\sup_{H \in \mathcal{H}_d} \left|P(H) - \widehat{P}_n(H)\right| \le \sqrt{\frac{1440\pi e}{1 - e^{-1}}}\sqrt{\frac{d+1}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

holds with probability at least $1 - \delta$, where by $\mathcal{H}_d$ we denote the system of all closed halfspaces in $\mathbb{R}^d$, see (3.24).

Proof. This can be proven analogously to [Chen et al., 2018, Lemma 7.3] using that the VC dimension of $\mathcal{H}_d$ is $d + 1$.

Theorem A3 (Rate of convergence for slabs) *Let $P \in \mathcal{P}(\mathbb{R}^d)$ and consider the empirical distribution $\widehat{P}_n$ based on a random sample of size n from P. Then for all $\delta \in (0, 1/2)$ the inequality*

$$\sup_{u \in S^{d-1}, t \ge 0} \max\left\{\left|P(H_{u,t}^{\mathsf{in}}) - \widehat{P}_n(H_{u,t}^{\mathsf{in}})\right|, \left|P(H_{u,t}^{\mathsf{out}}) - \widehat{P}_n(H_{u,t}^{\mathsf{out}})\right|\right\} \le \sqrt{\frac{1440\pi e}{1 - e^{-1}}}\sqrt{\frac{2d+3}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

holds with probability at least $1 - \delta$. The sets $H_{u,t}^{\mathsf{in}}$ and $H_{u,t}^{\mathsf{out}}$ are defined in (2.3).

Proof. Refer to [Chen et al., 2018, Lemma 7.3].

Theorem A4 *Let $N \sim \mathrm{Binomial}(n, p)$ and assume $p < 1/5$. Then, for every $\delta > 0$ satisfying $\sqrt{\frac{\log(1/\delta)}{2n}} < 1/5$, we have*

$$\frac{N}{n - N} \le \frac{p}{1 - p} + \frac{25}{12}\sqrt{\frac{\log(1/\delta)}{2n}} < \frac{2}{3}$$

with probability at least $1 - \delta$.

Proof. Refer to [Chen et al., 2018, Lemma 7.1].

Theorem A5 *Let* d N, $r > 0$ *and* $s > 1$. *Consider* $\{x_1, \ldots, x_M\}$ $B_d(\mathbf{0}, r)$ *a set of the maximum cardinality* M *such that*

$$\|x_i - x_j\| > r/s$$

holds for all $i = j$. *Then,* $M \geq s^d$.

Proof. Refer to [Vershynin, 2018, Proposition 4.2.12].