# Contributions to robust estimation : minimax optimality vs. computational tractability

Amir-Hossein Bateni

INSTITUT POLYTECHNIQUE DE PARIS

ENSAE

IP PARIS

NNT : 2022IPPAG004

Thèse de doctorat

# Contributions to robust estimation: minimax optimality *vs.* computational tractability

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École nationale de la statistique et de l'administration économique

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 6 juillet 2022, par

**AMIR-HOSSEIN BATENI**

Composition du Jury :

| | |
|---|---|
| Alexandre Tsybakov<br>CREST, ENSAE, IP Paris | Président |
| Pierre Alquier<br>RIKEN AIP Tokyo | Rapporteur |
| Mahdi Soltanolkotabi<br>University of Southern California | Rapporteur |
| Guillaume Lecué<br>CREST, ENSAE, IP Paris | Examinateur |
| Nicolas Verzelen<br>INRAE, Université de Montpellier | Examinateur |
| Arnak Dalalyan<br>CREST, ENSAE, IP Paris | Directeur de thèse |

# Remerciements

Mes premières remerciements vont certainement à Arnak. Je ne saurais comment te remercier pour tout ce que tu as fait pour moi. Tu m'as fait confiance alors que j'avais le parcours d'un outlier dans le monde statistique. Tu as supporté mes bêtises et mes défauts, tu t'es engagé sérieusement pour que j'apprenne et que j'avance, et en même temps tu m'as laissé explorer mes propres idées en croyant en moi. Je te remercie vivement pour ta patience, ton engagement, ton esprit ouvert, et ta bienveillance. Tu m'as montré un éminent exemple de savoir-être et savoir-vivre.

Je tiens à exprimer ma reconnaissance aux membres du jury, dont la présence est une véritable honneur pour moi. Thank you Mahdi for taking your time to review thoughtfully this manuscript. Merci Pierre pour ton rapport détaillé. Tu a été un généreux et chaleureux collègue pour moi. Merci Sacha de m'avoir accueilli au sein du pôle statistique au CREST et d'avoir accepté d'être le président de mon jury. Tes travaux ont eu une influence majeure sur ma recherche. Merci Nicolas d'avoir eu la gentillesse d'accepter de faire partie de mon jury. Et merci Guillaume pour ton soutien constant et pour nos discussions que j'ai toujours appréciées.

Je tiens aussi à exprimer mes remerciements à mes collègues chercheurs à l'ENSAE. Merci Victor de tes efforts pour le bon déroulement de mon doctorat. Je regrette de ne pas avoir pu continué le projet intéressant qu'on avait ensemble. Merci Katia de tes grands soutiens pendant la bonne année que j'ai passée à Dauphine. Merci à Cristina, Nicolas, Mathieu, Vianney, Anna, et Jaouad. J'ai beaucoup appris de vous humainement et professionnellement lors de nos différents échanges. Un immense merci à Arshak. Tu es vraiment un bon ami et en plus un bon collaborateur. Je suis content d'avoir travaillé avec toi et j'espère qu'on pourra encore travailler ensemble.

Je remercie également les doctorants et les anciens doctorants de l'ENSAE et alentours. Merci Arya pour toutes nos conversations interminables. Tu es une source inépuisable de motivation, d'ambition, et d'inspiration. Notre amitié est une longue histoire et il n'y a pas de place ici pour tout évoquer. Merci Avetik ! C'était un sacré bonheur de t'avoir à côté de moi pendant plusieurs années. Tu te disposes d'une forte joie de vivre que tu transmets aussi aux autres. Merci Étienne, Clara, et Hugo de l'ambiance vivante que vous avez créée dans le bureau. J'ai passé de très bons moments avec vous et surtout vous m'avez appris plein d'expressions françaises. Un grand merci à Jules, Christophe et Lucie pour leur im-

portant travail de l'assistance d'enseignement. Merci aux différentes générations, à Lionel, Simo, Gautier, Alexis, Mamadou, Jérémy, Yannick, Evgenii, Léna, Geoffrey, Badr, Boris, Nicolas, Gabriel, Solenne, François-Pierre, Meyer, Julien, Suzanne, Dang, Flore, Nayel, Younes, Sasila, Yannis, Alexendre, Corentin, Maria, Théo, et tous ceux que j'ai éventuellement oubliés. Et pareillement merci aux post-doctorants, à Yann, Thomas, Martin, et Jiaying.

Je remercie grandement mes chers amis. Merci Ali et Sahar de votre gentillesse et générosité ! Vous êtes vraiment magnifiques ! Merci MohammadSaeid et Ebrahim de votre soutien et présence ! Sans vous, ces années auraient vraiment été difficiles.

Last but not least, I want to thank my family. Indeed, I can never thank you enough my lovely parents as you have made so many sacrifices for me. I also thank warmly my dear sisters, my brothers-in-law, and my sweet nephews. Your presence and your support have always been important for me.

# Chapter 1

# Introduction



## 1.1 Problem definition

In statistics and learning theory, it is common to assume that samples are independently and identically distributed according to a reference probability distribution. A more realistic approach could be to relax this assumption by allowing a fraction of samples to not necessarily follow the reference distribution. These disobeying samples, called *outliers*, may drastically skew the classical estimators.

Consider the following setting. Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ be $n$ i.i.d. samples drawn from $\mathcal{N}(\boldsymbol{\mu}, \sigma^2)$. Given $\varepsilon \in (0, 1/2)$, we observe $\varepsilon$-*contaminated* Gaussian random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ which

means that we have $\boldsymbol{X}_i = \boldsymbol{Y}_i$ for $i \in \mathcal{I}$ while $\boldsymbol{X}_i$ takes arbitrary values when $i \in \mathcal{O}$, where $|\mathcal{I}| = n(1-\varepsilon)$ and $|\mathcal{O}| = n\varepsilon$. Let $\mathcal{S}$ be $\mathcal{I} \cup \mathcal{O}$ which is indeed the set $\{1, \dots, n\}$. Our goal is to estimate $\boldsymbol{\mu}$ and we measure the estimation error by absolute value. When there is no outlier ($\mathcal{O} = \varnothing$), the maximum likelihood estimator (MLE) which is the sample mean, estimates $\boldsymbol{\mu}$ with optimal non-asymptotic error rate, while the presence of a single outlier with an extreme large value (far enough from $\boldsymbol{\mu}$) can cause an arbitrary large deviation of MLE from $\boldsymbol{\mu}$. So, we see that MLE is not robust to outliers.

In this work, we aim to construct estimators robust to outliers, and we focus on the fundamental task of mean estimation, the task to which many problems in statistics reduce. We are interested in the non-asymptotic behavior of estimators.

## 1.2 Robust approaches for one dimension

Let us start by exploring different ideas for estimating robustly $\boldsymbol{\mu}$ in the setting of $\varepsilon$-corrupted Gaussian samples.

### 1.2.1 Filtering

As we have seen the problem with MLE is its high sensitivity to extreme far outliers from $\boldsymbol{\mu}$. To get around this problem, we can simply remove the largest and smallest points and then apply MLE. Indeed, we sort the observations $\boldsymbol{X}_{(1)} \leq \cdots \leq \boldsymbol{X}_{(n)}$ and return $\widehat{\boldsymbol{\mu}}_F := \sum_{i \in F} \boldsymbol{X}_i / |F|$ where $F = \{i \in \mathcal{I} \cup \mathcal{O} | \boldsymbol{X}_{(2n\varepsilon)} < \boldsymbol{X}_i < \boldsymbol{X}_{(n-2n\varepsilon)}\}$[1]. After filtering even if there are still outliers among the samples, we are sure that they are at most as far as $\boldsymbol{Y}_{(n\varepsilon)}$ and $\boldsymbol{Y}_{(n-n\varepsilon)}$ (where the initial samples are sorted: $\boldsymbol{Y}_{(1)} \leq \cdots \leq \boldsymbol{Y}_{(n)}$). Hence, this estimator, called *Trimmed mean*, is not so much influenced by extreme outliers. See Figure 1.1.

We try to analyze informally the estimation error of $\widehat{\boldsymbol{\mu}}_F$ without entering into the details. Let $\boldsymbol{\xi}_i$ denote $\boldsymbol{Y}_i - \boldsymbol{\mu}$ for every $i \in \mathcal{S}$. By the triangle inequality, we have

$$\begin{aligned} |\widehat{\boldsymbol{\mu}}_F - \boldsymbol{\mu}| &\leq \frac{1}{|F|} \left| \sum_{i \in \mathcal{I} \cap F} (\boldsymbol{X}_i - \boldsymbol{\mu}) \right| + \frac{1}{|F|} \left| \sum_{i \in \mathcal{O} \cap F} (\boldsymbol{X}_i - \boldsymbol{\mu}) \right| \\ &\leq \frac{1}{|F|} \left| \sum_{i \in \mathcal{I} \cap F} \boldsymbol{\xi}_i \right| + \frac{|\mathcal{O} \cap F|}{|F|} \max\left( |\boldsymbol{\xi}_{(n\varepsilon)}|, |\boldsymbol{\xi}_{(n-n\varepsilon)}| \right). \end{aligned} \tag{1.1}$$

We aim to upper bound the two terms in (1.1).

**Remark 1.** *While the random variables $(\boldsymbol{Y}_i)_{i \in \mathcal{S}}$ are independent, the random variables $(\boldsymbol{Y}_i \mathbb{1}(i \in \mathcal{I}))_{i \in \mathcal{S}}$ are generally not independent. Indeed, we do not have any particular assumption on outliers and this implies that worst scenarios are possible. For instance, we may suppose that there is an adversary who observes the initial samples $(\boldsymbol{Y}_i)_{i \in \mathcal{S}}$, and knowing our estimator, replaces $\varepsilon n$ of them by $\varepsilon n$ arbitrary values in a malicious way. At the end, we receive*

[1] Without loss of generality we suppose that $n\varepsilon$ is an integer.

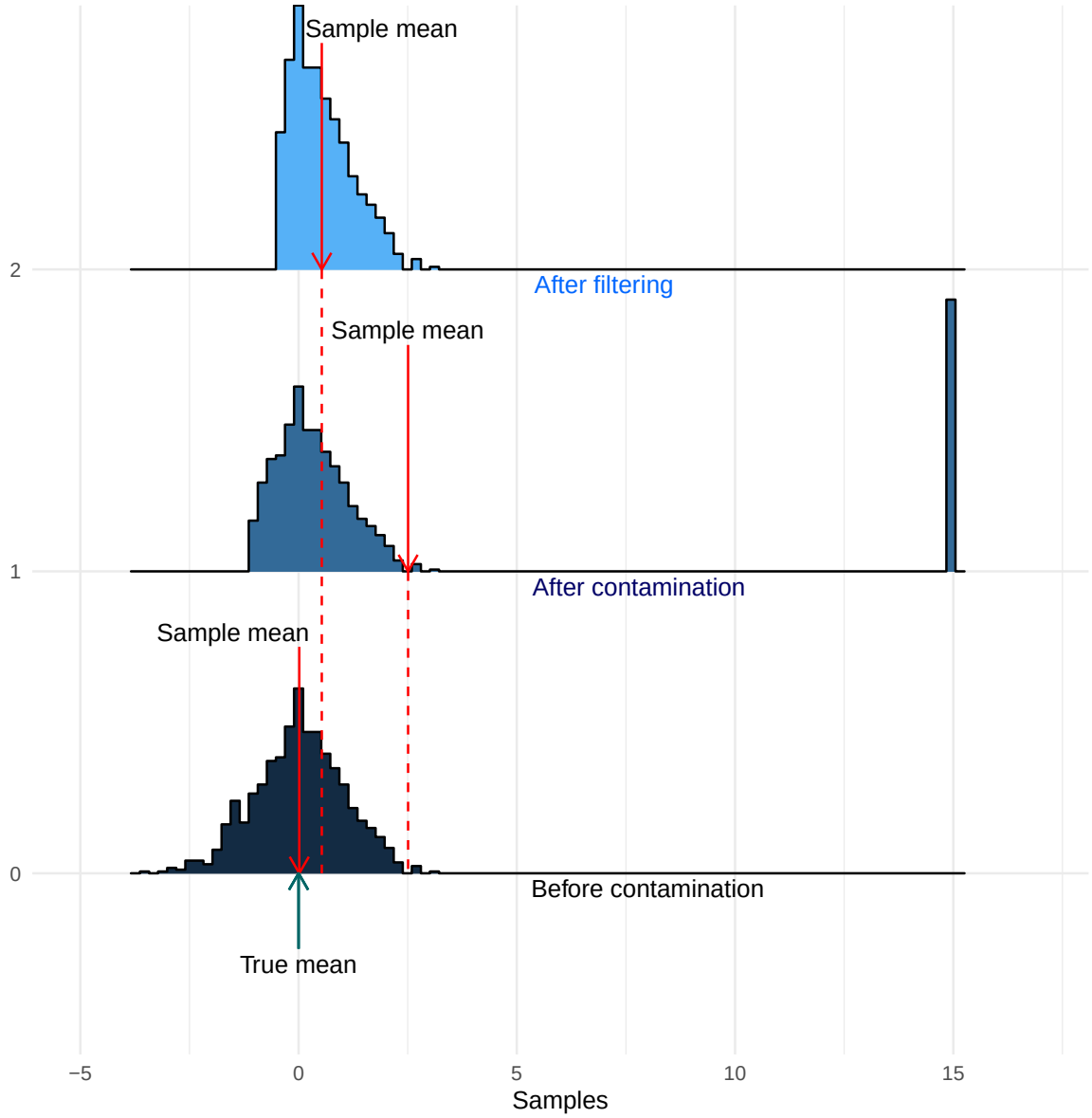


Figure 1.1: At the first floor the histogram of the initial samples $(\boldsymbol{Y}_i)_{i\in\mathcal{S}}$ is depicted. At the second floor, we have the histogram of the contaminated samples $(\boldsymbol{X}_i)_{i\in\mathcal{S}}$ where the smallest $n\varepsilon$ samples among $(\boldsymbol{Y}_i)_{i\in\mathcal{S}}$ are replaced by $n\varepsilon$ samples equal to 15. At the third floor, we see the histogram of samples after the filtering procedure where the $2n\varepsilon$ greatest and $2n\varepsilon$ smallest samples among $(\boldsymbol{X}_i)_{i\in\mathcal{S}}$ are removed. The sample means for each set are marked by red lines. The effect of the filtering procedure on the performance of the sample mean is apparent.

*only $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ as described at the beginning of the chapter. This is called the* adversarial contamination model*. In this case, the nature of the subset $\mathcal{I}$ might depend on many factors included all the initial samples $(\boldsymbol{Y}_i)_{i \in S}$.*

To bypass this problem of lack of independence in the first term of (1.1), we can state by the triangle inequality that

$$\begin{aligned} \frac{1}{|F|}\left|\sum_{i \in \mathcal{I} \cap F} \boldsymbol{\xi}_i\right| &\leq \frac{1}{|F|}\left|\sum_{i \in S} \boldsymbol{\xi}_i\right| + \frac{1}{|F|}\left|\sum_{i \in S \setminus (\mathcal{I} \cap F)} \boldsymbol{\xi}_i\right| \\ &\leq \frac{1}{|F|}\left|\sum_{i=1}^{n} \boldsymbol{\xi}_i\right| + \max_{|J| \leq 5n\varepsilon} \frac{1}{|F|}\left|\sum_{j \in J} \boldsymbol{\xi}_j\right|, \end{aligned} \tag{1.2}$$

where in the second inequality we used $n - 5n\varepsilon \leq |\mathcal{I} \cap F|$. With high probability, the first term in (1.2) is of order $(1-4\varepsilon)^{-1}(\sigma/\sqrt{n})$ (using the Gaussian tail bounds and the fact that $|F| = n - 4n\varepsilon$), and the second term is bounded by a term of order $(1-4\varepsilon)^{-1}\sigma\varepsilon\sqrt{\log(e/\varepsilon)}$ (using the Gaussian properties and the union bound). For more details on these calculations, we refer the reader to the proof of Lemma 6.

To upper bound the second term in (1.1), we can use Hoeffding's inequality to prove that $\boldsymbol{\xi}_{(n\varepsilon)}$ and $\boldsymbol{\xi}_{(n-n\varepsilon)}$ respectively are concentrated around $\sigma\Phi^{-1}(\varepsilon)$ and $\sigma\Phi^{-1}(1-\varepsilon)$ where $\Phi$ is the cumulative distribution function of $\mathcal{N}(0,1)$. The absolute value of these two last terms is bounded by $\sigma\sqrt{\log(1/\varepsilon)}$ since we deal with Gaussian variables. This entails that the second term in (1.1) is bounded by a term of order $(1-4\varepsilon)^{-1}\sigma\varepsilon\sqrt{\log(1/\varepsilon)}$ with high probability. Therefore, we informally proved that $\widehat{\boldsymbol{\mu}}_F$ estimates $\boldsymbol{\mu}$ with an error rate of order at most

$$\frac{(\sigma/\sqrt{n}) + \sigma\varepsilon\sqrt{\log(1/\varepsilon)}}{1-4\varepsilon}$$

with high probability.

The error rate of the trimmed mean is composed of the optimal error rate $\sigma/\sqrt{n}$ in the case of non contamination (which is achieved by MLE), and $\sigma\varepsilon\sqrt{\log(1/\varepsilon)}$ which shows the impact of outliers in our estimation in worst case. If there is no outlier ($\varepsilon = 0$), the trimmed mean does not remove any point and returns the sample mean. Nevertheless, this estimator is tolerant to $\varepsilon$ not larger than 1/4.

By removing other quantities of points than $4n\varepsilon$ in the filtering step, one can change the performance of the trimmed mean. A special case is the sample median where $\lceil n/2 \rceil - 1$ points are filtered out from each side. We analyze the sample median as an estimator of the mean in the next section.

### 1.2.2 Mean as center of symmetry

The Gaussian distribution is a symmetric distribution, and the mean is not only the barycentre of the distribution, but also the center of symmetry. The idea is to estimate the mean, as a

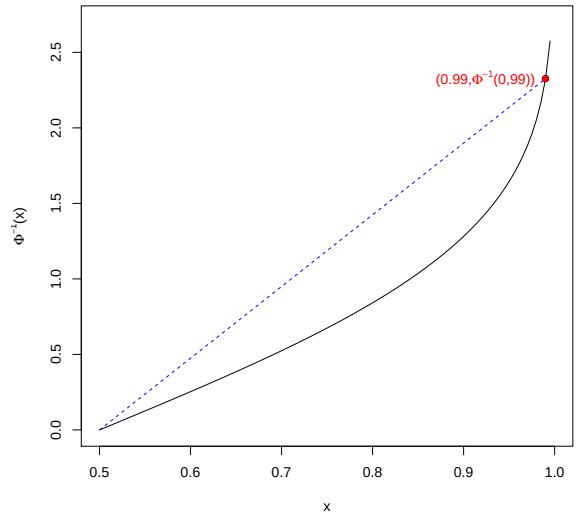


Figure 1.2: The behavior of the probit function: for $x \in (0, 0.49)$, $\Phi^{-1}(0.5 + x) < 5x$.

center of symmetry. There are various notions of symmetry. The one that we consider here is the symmetry induced by mass where there are same masses in the both sides of the center. More precisely, in this symmetry, the center is the median. Thus, we may estimate the center, *i.e.* the mean, by the sample median which is in fact the center of symmetry for the empirical mass.

To analyze the sample median, we exceptionally provide more details since it makes our presentation simpler. The sample median of the contaminated data[2], $\boldsymbol{X}_{(\lceil n/2 \rceil)}$, is located between $\boldsymbol{Y}_{(\lceil n/2 \rceil - n\varepsilon)}$ and $\boldsymbol{Y}_{(\lceil n/2 \rceil + n\varepsilon)}$, and by Hoeffding's inequality we can prove that $|\boldsymbol{Y}_{(\lceil n/2 \rceil - n\varepsilon)} - \boldsymbol{\mu}|$ and $|\boldsymbol{Y}_{(\lceil n/2 \rceil + n\varepsilon)} - \boldsymbol{\mu}|$ are bounded by

$$\sigma \Phi^{-1}\Big(\frac{1}{2} + \varepsilon + \sqrt{\frac{\log(2/\delta)}{2n}}\Big)$$

with probability at least $1 - \delta$ for any $\delta \in (0, 1)$ satisfying $\frac{1}{2} + \varepsilon + \sqrt{\frac{\log(2/\delta)}{2n}} < 1$. Observing the behavior of the function $\Phi^{-1}$ (called *probit*) around 1/2 in Figure 1.2, we note that if $\varepsilon + \sqrt{\frac{\log(2/\delta)}{2n}} \leq 0.49$, we have

$$\Phi^{-1}\Big(\frac{1}{2} + \varepsilon + \sqrt{\frac{\log(2/\delta)}{2n}}\Big) \leq 5\Big(\varepsilon + \sqrt{\frac{\log(2/\delta)}{2n}}\Big).$$

We can conclude that if $\varepsilon + \sqrt{\frac{\log(2/\delta)}{2n}} \leq 0.49$, the inequality $|\boldsymbol{X}_{(n/2)} - \boldsymbol{\mu}| \leq 5\sigma\Big(\varepsilon + \sqrt{\frac{\log(2/\delta)}{2n}}\Big)$ holds with probability at least $1 - \delta$.

**Remark 2.** *Let fix $\delta \in (0, 1)$. We have seen that for $\varepsilon < \frac{1}{2} - \sqrt{\frac{\log(2/\delta)}{2n}}$ the estimation error of*

[2]For sake of simplicity, here we suppose that $n$ is odd.

*the sample median is bounded with probability at least* $1-\delta$*. What if* $\varepsilon \geq \frac{1}{2} - \sqrt{\frac{\log(2/\delta)}{2n}}$*? For* $\varepsilon \in \left[\frac{1}{2} - \sqrt{\frac{\log(2/\delta)}{2n}}, \frac{1}{2} - \frac{1}{n}\right]$*, we can claim that the sample median* $\boldsymbol{X}_{(\lceil n/2 \rceil)}$ *is located between* $\boldsymbol{Y}_{(1)}$ *and* $\boldsymbol{Y}_{(n)}$ *which means that* $|\boldsymbol{X}_{(\lceil n/2 \rceil)} - \boldsymbol{\mu}|$ *is bounded by a term of order* $\sigma\sqrt{\log(n)}$ *with high probability. For* $\varepsilon > \frac{1}{2} - \frac{1}{n}$*, the sample median could be an outlier (since the majority of data is outlier) and therefore, the estimation error is not bounded in worst case. So, we can state that for fixed* $n$ *the* breakdown point *of the sample median is* $\frac{1}{2} - \frac{1}{n}$*. The* breakdown point *is the maximum proportion of outliers for which the estimation error is bounded in worst case. For instance, the breakdown point of the sample mean is* $1/n$ *and that of the trimmed mean is* $\frac{1}{4} - \frac{1}{n}$*. Given our non-asymptotic study framework, we are interested here in the notion of finite sample breakdown point. However, in Section 1.5.2 and Chapter 3, we will consider the asymptotic notion of the breakdown point.*

As we emphasized, the sample median is characterized by the empirical mass, and thereby its robustness could be interpreted by the robustness of the empirical mass since the empirical mass is not sensitive to extreme far outliers (it can be changed at most by $2\varepsilon$ when data become contaminated). Another robust method could be to leverage the empirical mass to estimate the probability density function of the reference distribution. We will expose this method in the coming section.

### 1.2.3 Minimum distance estimator

We know the explicit expression of the probability density function (p.d.f.) for our reference distribution in terms of the mean parameter. Another approach for our problem might be to determine a value as mean in a way that the corresponding p.d.f. fits best the observations. To do so, we can select a set of possible candidates for the mean, and choose the one with the best fitting of associated p.d.f. to data.

First, we introduce, as it is presented in (Devroye and Lugosi, 2000), the *Scheffé*'s estimate which chooses the best between two p.d.f. $f_1$ and $f_2$. Let $f$ be a p.d.f. and $\widehat{\nu}$ be a probability distribution. The Scheffé's estimate $\widehat{f}$ is defined as

$$\widehat{f} := \begin{cases} f_1 & \text{if } \left|\int_A f_1 - \widehat{\nu}(A)\right| < \left|\int_A f_2 - \widehat{\nu}(A)\right|, \\ f_2 & \text{otherwise,} \end{cases}$$

where $A = \{x : f_1(x) > f_2(x)\}$. (Devroye and Lugosi, 2000, Theorem 6.1) states that

$$\int |f - \widehat{f}| \leq 3 \min\left(\int |f_1 - f|, \int |f_2 - f|\right) + 4 \max_{A \in \mathcal{A}} \left|\int_A f - \widehat{\nu}(A)\right|, \tag{1.3}$$

where $\mathcal{A} = \{\{f_1 > f_2\}, \{f_2 > f_1\}\}$. Now, assume that $f_{\boldsymbol{\mu}}$ is the p.d.f. of the reference distribution $\mathcal{N}(\boldsymbol{\mu}, \sigma)$ and $\nu$ is the empirical measure of the non-contaminated random variables $(\boldsymbol{Y}_i)_{i \in \mathcal{S}}$, *i.e.*, for every set $A$, $\nu(A) = \sum_{i=1}^{n} \mathbb{1}(\boldsymbol{Y}_i \in A)/n$. However, for choosing the best p.d.f.

we have only access to the empirical measure $\widehat{\nu}$ of the contaminated data which associates to every set $A$ the mass $\widehat{\nu}(A) = \sum_{i=1}^{n} \mathbb{1}(\boldsymbol{X}_i \in A)/n$. Using (1.3) and the triangle inequality, we obtain

$$\int |f_{\boldsymbol{\mu}} - \widehat{f}| \leq 3 \min\left( \int |f_1 - f_{\boldsymbol{\mu}}|, \int |f_2 - f_{\boldsymbol{\mu}}| \right) + 4 \max_{A \in \mathcal{A}} \left| \int_A f_{\boldsymbol{\mu}} - \nu(A) \right| + 4 \max_{A \in \mathcal{A}} |\nu(A) - \widehat{\nu}(A)|.$$

The term $\max_{A \in \mathcal{A}} \left| \int_A f_{\boldsymbol{\mu}} - \nu(A) \right|$ is the difference between the empirical and theoretical measures. It is bounded by a term of order $\sqrt{1/n}$ with high probability via Hoeffding's inequality and the union bound. The term $\max_{A \in \mathcal{A}} |\nu(A) - \widehat{\nu}(A)|$ is at most two times the maximum mass that the adversary is allowed to transport in order to contaminate data, namely $2\varepsilon$. If $f_1$ and $f_2$ respectively are the p.d.f. of $\mathcal{N}(\boldsymbol{\mu}_1, \sigma^2)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \sigma^2)$, for $i \in \{1, 2\}$ we have $\int |f_i - f_{\boldsymbol{\mu}}| = 2\mathrm{TV}\big(\mathcal{N}(\boldsymbol{\mu}_i, \sigma^2), \mathcal{N}(\boldsymbol{\mu}, \sigma^2)\big)$ (Scheffé's theorem) where TV denotes the total variation distance. Furthermore, by Pinsker's inequality we know that $\mathrm{TV}\big(\mathcal{N}(\boldsymbol{\mu}_i, \sigma^2), \mathcal{N}(\boldsymbol{\mu}, \sigma^2)\big) \leq |\boldsymbol{\mu}_i - \boldsymbol{\mu}|/(2\sigma)$. So, finally we deduce that there exists $C > 0$ such that the inequality

$$\int |f_{\boldsymbol{\mu}} - \widehat{f}| \leq \frac{3}{\sigma} \min\left( |\boldsymbol{\mu}_1 - \boldsymbol{\mu}|, |\boldsymbol{\mu}_2 - \boldsymbol{\mu}| \right) + C \frac{1}{\sqrt{n}} + 8\varepsilon,$$

is satisfied with high probability. As $\boldsymbol{\mu}$ is located between $\boldsymbol{X}_{((\varepsilon+0.05)n)}$ and $\boldsymbol{X}_{((1-\varepsilon-0.05)n)}$ with high probability, given $\gamma > 0$ we can construct a covering set $M_\gamma = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{N_\gamma}\}$ over $[\boldsymbol{X}_{((\varepsilon+0.05)n)}, \boldsymbol{X}_{((1-\varepsilon-0.05)n)}]$ where for all $\boldsymbol{m}$ in $[\boldsymbol{X}_{((\varepsilon+0.05)n)}, \boldsymbol{X}_{((1-\varepsilon-0.05)n)}]$ there exists $i \in \{1, \ldots, N_\gamma\}$ such that $|\boldsymbol{\mu}_i - \boldsymbol{m}| \leq \gamma$. Now, we consider a finite family of p.d.f. candidates $(f_i)_{i \in \{1, \ldots, N_\gamma\}}$ where $f_i$ is the p.d.f. of $\mathcal{N}(\boldsymbol{\mu}_i, \sigma^2)$ and extend the Scheffé's estimate in order to choose the best p.d.f. among $(f_i)_{i \in \{1, \ldots, N_\gamma\}}$ by running a tournament between these candidates. We declare $f_{\widehat{k}}$ as the winner of the tournament with

$$\widehat{k} = \underset{k \in \{1, \ldots, N_\gamma\}}{\arg\min} \max_{A \in \mathcal{A}_\gamma} \left| \int_A f_k - \widehat{\nu}(A) \right|, \tag{1.4}$$

where $\mathcal{A}_\gamma = \big\{ \{x : f_i(x) > f_j(x)\} : i, j \in \{1, \ldots, N_\gamma\} \big\}$. When there are multiple minima for (1.4), $\widehat{k}$ is defined to be the smallest index. Using (Devroye and Lugosi, 2000, Theorem 6.3), Hoeffding's inequality and the union bound, we can show that there exists $C > 0$ such that the inequality

$$\int |f_{\boldsymbol{\mu}} - f_{\widehat{k}}| \leq \frac{3}{\sigma} \gamma + C \sqrt{\frac{\log(N_\gamma^2)}{n}} + 8\varepsilon,$$

holds with high probability. We set $\gamma = \frac{\sigma}{\sqrt{n}}$, and thereby $N_\gamma^2$ is of order $4n$ as the length of $[\boldsymbol{X}_{((\varepsilon+0.05)n)}, \boldsymbol{X}_{((1-\varepsilon-0.05)n)}]$ is less than $4\sigma$ with high probability. This yields the existence of $C > 0$ such that

$$\int |f_{\boldsymbol{\mu}} - f_{\widehat{k}}| \leq C \sqrt{\frac{\log(n)}{n}} + 8\varepsilon,$$

is true with high probability. Now, we argue that $\boldsymbol{\mu}_{\widehat{k}}$ is a robust estimator of $\boldsymbol{\mu}$. Assume without loss of generality $\boldsymbol{\mu} < \boldsymbol{\mu}_{\widehat{k}}$. Then we have

$$\begin{aligned}\int |f_{\boldsymbol{\mu}} - f_{\widehat{k}}| = 2\int_{-\infty}^{(\boldsymbol{\mu}+\boldsymbol{\mu}_{\widehat{k}})/2} (f_{\boldsymbol{\mu}} - f_{\widehat{k}}) &= \frac{2}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{(\boldsymbol{\mu}+\boldsymbol{\mu}_{\widehat{k}})/2} e^{-(x-\boldsymbol{\mu})^2/(2\sigma^2)}dx \\ &\quad - \frac{2}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{(\boldsymbol{\mu}+\boldsymbol{\mu}_{\widehat{k}})/2} e^{-(x-\boldsymbol{\mu}_{\widehat{k}})^2/(2\sigma^2)}dx \\ &= 2\Phi\left(\frac{\boldsymbol{\mu}_{\widehat{k}} - \boldsymbol{\mu}}{2\sigma}\right) - 2\Phi\left(\frac{\boldsymbol{\mu} - \boldsymbol{\mu}_{\widehat{k}}}{2\sigma}\right) \\ &= 4\Phi\left(\frac{\boldsymbol{\mu}_{\widehat{k}} - \boldsymbol{\mu}}{2\sigma}\right) - 2. \end{aligned} \tag{1.5}$$

This leads to

$$|\boldsymbol{\mu} - \boldsymbol{\mu}_{\widehat{k}}| \leq 2\sigma\Phi^{-1}\left(\frac{C}{4}\sqrt{\frac{\log(n)}{n}} + 2\varepsilon + \frac{1}{2}\right),$$

with high probability if $\frac{C}{4}\sqrt{\frac{\log(n)}{n}} + 2\varepsilon < \frac{1}{2}$. Again, given the behavior of the probit function around 1/2 (cf. Figure 1.2), we deduce that

$$|\boldsymbol{\mu} - \boldsymbol{\mu}_{\widehat{k}}| \leq 10\sigma\left(\frac{C}{4}\sqrt{\frac{\log(n)}{n}} + 2\varepsilon\right),$$

is satisfied with high probability if $\frac{C}{4}\sqrt{\frac{\log(n)}{n}} + 2\varepsilon \leq 0.49$. This estimator $\boldsymbol{\mu}_{\widehat{k}}$, called *skeleton estimate*, has a breakdown point of at least $\frac{1}{4} - \frac{C}{8}\sqrt{\frac{\log(n)}{n}}$.

While the skeleton estimate appears to be based on the estimation of p.d.f., we should stress that its key tool is the empirical mass, and this provides robustness for this estimator.

### 1.2.4 Lower bound

We have obtained so far three different error rates for estimating robustly the Gaussian mean in one dimension, namely

- with the trimmed mean: $\frac{\sigma}{\sqrt{n}} + \sigma\varepsilon\log(1/\varepsilon)$,
- with the sample median: $\frac{\sigma}{\sqrt{n}} + \sigma\varepsilon$,
- and with the skeleton estimate: $\sigma\sqrt{\frac{\log(n)}{n}} + \sigma\varepsilon$.

The best rate is achieved by the sample median. At this stage, the question is whether it is possible to do better than the sample median. Here, we attempt to show a lower bound for *Huber's contamination model*, a contamination model weaker than the adversarial model. We assume that the observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are drawn from the mixture distribution defined by

$$(1-\varepsilon)\mathcal{N}(\boldsymbol{\mu}, \sigma^2) + \varepsilon Q,$$

where $Q$ is an unknown distribution representing the corruption in our problem. In Chapter 2 (Proposition 3), we show that Huber's model is nearly contained in the adversarial model, and this implies that our lower bound for Huber's model holds for the adversarial model, too. This lower bound is originally proposed in (Chen et al., 2018).

Consider two distributions $\mathcal{N}(\boldsymbol{\mu}_1, \sigma^2)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \sigma^2)$ with their respective p.d.f. $f_1$ and $f_2$. We propose two distributions $Q_1$ and $Q_2$ defined respectively by the p.d.f. $\frac{1-\varepsilon}{\varepsilon}(f_1 - f_2)\mathbb{1}(f_1 \geq f_2)$ and $\frac{1-\varepsilon}{\varepsilon}(f_2 - f_1)\mathbb{1}(f_2 \geq f_1)$. It is now easy to verify that $(1-\varepsilon)\mathcal{N}(\boldsymbol{\mu}_1, \sigma^2) + \varepsilon Q_1$ and $(1-\varepsilon)\mathcal{N}(\boldsymbol{\mu}_2, \sigma^2) + \varepsilon Q_2$ are the same distribution while we have $\mathrm{TV}\big(\mathcal{N}(\boldsymbol{\mu}_1, \sigma^2), \mathcal{N}(\boldsymbol{\mu}_2, \sigma^2)\big) = \frac{\varepsilon}{1-\varepsilon}$. That is to say that the two different distributions $\mathcal{N}(\boldsymbol{\mu}_1, \sigma^2)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \sigma^2)$ might be contaminated in a way that they are not distinguishable one from another. In this situation, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ have the same chance to be the mean of the reference distribution and our best estimation of the mean would be the middle point $(\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)/2$. Thus, our estimation error is at least $|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1|/2$. Let $\delta$ denote $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, if $\boldsymbol{\mu}_1 < \boldsymbol{\mu}_2$ by similar calculations as in (1.5) we have

$$\mathrm{TV}\big(\mathcal{N}(\boldsymbol{\mu}_1, \sigma^2), \mathcal{N}(\boldsymbol{\mu}_2, \sigma^2)\big) = \int_{-\infty}^{\boldsymbol{\mu}_1 + \delta/2} (f_1 - f_2) = 2\Phi\Big(\frac{\delta}{2\sigma}\Big) - 1,$$

and this entails that $|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2| = 2\sigma\Phi^{-1}\big(\frac{1}{2} + \frac{1}{2}\frac{\varepsilon}{1-\varepsilon}\big)$. Adding to this, the lower bound for estimating the Gaussian mean with non-contaminated data, namely $\sigma/\sqrt{n}$, our error rate under contamination cannot be better than

$$\sigma\Phi^{-1}\Big(\frac{1}{2} + \frac{\varepsilon}{2(1-\varepsilon)}\Big) + \frac{\sigma}{\sqrt{n}}.$$

This lower bound matches our upper bound for the sample median $\sigma\Phi^{-1}\Big(\frac{1}{2} + \varepsilon + \frac{1}{\sqrt{n}}\Big)$ and we conclude that the sample median is a minimax estimator of the mean when data are corrupted. Note that when $\varepsilon$ tends to 1/2, the minimax risk and the risk of the sample median explode to infinity.

**Remark 3.** *The three studied estimators in this section are computable in polynomial time in terms of sample size. The trimmed mean requires the knoweledge of $\varepsilon$, the skeleton estimate requires the knowledge of $\sigma$ whereas the sample median does not require any of them. Moreover, the sample median is tolerant to larger values for $\varepsilon$. Consequently, in addition to its optimal error rate, the sample median seems to be the most interesting estimator for our problem.*

## 1.3 Robust approaches for high dimension

Now, we assume that data are in $\mathbb{R}^p$ and the initial random variables $(\boldsymbol{Y}_i)_{i\in\{1,\dots,n\}}$ follow the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_p)$. As in the previous setting, we observe $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$, a $\varepsilon$-contaminated version of data under the adversarial model. We quantify the estimation error by the Euclidean distance $\|.\|_2$.

A naive solution for the multidimensional setting would be to estimate each coordinate of $\boldsymbol{\mu}$ separately via a robust method for one dimension. By the union bound, we may establish that applying for example the sample median to each coordinate gives an estimator $\widehat{\boldsymbol{\mu}}_{\mathrm{CM}}$ satisfying

$$\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{\mathrm{CM}}\|_2 \leq 5\sigma\Big(\sqrt{\frac{p\log(2p/\delta)}{2n}} + \varepsilon\sqrt{p}\Big),$$

with probability at least $1-\delta$ if $\varepsilon + \sqrt{\log(2/\delta)/(2n)} \leq 0.49$. This estimator is known as the *componentwise median*. The problem with such estimators is that the term $\varepsilon\sqrt{p}$ in the error could become problematic when $p$ is large. Chen et al. (2018) prove that under Huber's contamination model, the error rate of the componentwise median cannot be better than $\sigma(\sqrt{p/n} + \varepsilon\sqrt{p})$ when the contamination distribution $Q$ is defined as a Dirac on the point $\boldsymbol{\mu} + \sigma(1,\dots,1)^\top$. The same contamination distribution can be deployed to show that the other robust methods for one dimension will not have a better result in high dimension. Now, the question is whether we can have a better dependency on $p$.

**Remark 4.** *One might improve slightly the dependency on $p$ for the componentwise median by applying it not in the canonical basis but in the basis formed by the eigenvectors of the sample covariance matrix. Against this method, the adversary may choose outliers $\boldsymbol{Z}_1,\dots,\boldsymbol{Z}_{n\varepsilon}$ where for $\boldsymbol{Z}_i$ the $i$ first coordinates are $\sigma$ and the rest is zero. In such situation, one can show that the error rate of the componentwise median is at least $\sigma(\sqrt{p/n} + \varepsilon\sqrt{\min(n\varepsilon, p)})$.*

### 1.3.1 Filtering (multidimensional case)

In the multidimensional setting, there is no notion of order. Therefore, we cannot apply the filtering method proposed previously for one dimension. However, we know that multivariate Gaussian samples are concentrated on a sphere of radius $\sigma\sqrt{p}$ with the mean as center (note the contrast with univariate Gaussian samples which are concentrated close to the mean). More precisely, there exists $c > 0$ such that for all $t > 0$

$$\mathbf{P}\big(\big|\|\boldsymbol{\xi}_i\|_2 - \sigma\sqrt{p}\big| > t\big) \leq 2\exp(-ct^2/\sigma^2), \tag{1.6}$$

(see *e.g.*, (Vershynin, 2018, Theorem 3.1.1)). See Figure 1.3.

We can now design a filter based on the property described by (1.6). In fact, this property states that given $\tau \in (0,1)$, at least $1-\tau$ fraction of $(\boldsymbol{\xi}_i)_{i\in\{1,\dots,n\}}$ satisfies with high probability

$$\|\boldsymbol{\xi}_i\|_2 \leq C\sigma\big(\sqrt{p} + \sqrt{\log(1/\tau)}\big),$$

where $C$ is a positive constant. For a formal statement of this property, see Lemma 3. On this event, given $\varepsilon$-contaminated data, we can claim that at least $1-\tau-\varepsilon$ fraction of $(\boldsymbol{X}_i)_{i\in\{1,\dots,n\}}$

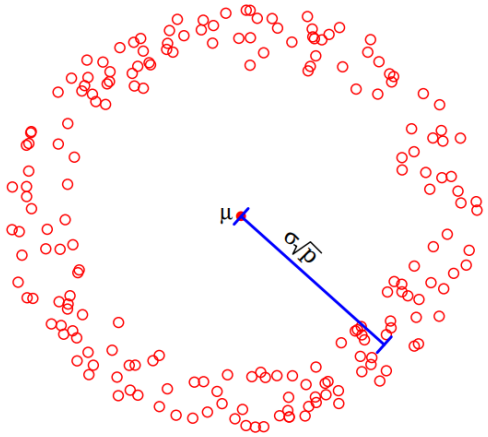


Figure 1.3: A schematic representation of the Gaussian point cloud in high dimension. Points are concentrated around a sphere with center $\boldsymbol{\mu}$ and radius $\sigma\sqrt{p}$.

satisfies

$$\|\boldsymbol{X}_i - \boldsymbol{\mu}\|_2 \leq C\sigma\big(\sqrt{p} + \sqrt{\log(1/\tau)}\big). \tag{1.7}$$

For each point $\boldsymbol{X}_i$, we consider the set $A_i := \{\boldsymbol{X}_j : \|\boldsymbol{X}_j - \boldsymbol{X}_i\|_2 \leq 2C\sigma\big(\sqrt{p} + \sqrt{\log(1/\tau)}\big)\}$. If the cardinality of this set is more than $n(\tau+\varepsilon)$, there is a point $\boldsymbol{X}_j$ in $A_i$ satisfying (1.7). By the triangle inequality, this implies that $\boldsymbol{X}_i$ has a distance at most $3C\sigma\big(\sqrt{p} + \sqrt{\log(1/\tau)}\big)$ from $\boldsymbol{\mu}$. Hence, all the points exhibited by the set $F := \{i : |A_i| \geq n(\varepsilon+\tau)\}$ are in a distance of order $\sigma\big(\sqrt{p} + \sqrt{\log(1/\tau)}\big)$ from $\boldsymbol{\mu}$. Note that $F$ contains all $\boldsymbol{X}_i$ satisfying (1.7) with $i \in \mathcal{I}$ if $n(1-\tau-\varepsilon) > n(\tau+\varepsilon)$, *i.e.*, if $\tau + \varepsilon < 1/2$. Otherwise, $F$ could be of cardinality zero in worst case. So, $F$ is of cardinality at least $n(1-\tau-\varepsilon)$ if $\tau+\varepsilon < 1/2$. This procedure of filtering is called *Naive pruning* (Diakonikolas et al., 2016a).

The empirical mean $\widehat{\boldsymbol{\mu}}_F$ of the points in $F$ satisfies

$$\begin{aligned}\|\widehat{\boldsymbol{\mu}}_F - \boldsymbol{\mu}\|_2 &\leq \frac{1}{|F|}\Big\|\sum_{i\in\mathcal{I}\cap F}(\boldsymbol{X}_i - \boldsymbol{\mu})\Big\|_2 + \frac{1}{|F|}\Big\|\sum_{i\in\mathcal{O}\cap F}(\boldsymbol{X}_i - \boldsymbol{\mu})\Big\|_2 \\ &\leq \frac{1}{|F|}\Big\|\sum_{i\in\mathcal{I}\cap F}\boldsymbol{\xi}_i\Big\|_2 + \frac{|\mathcal{O}\cap F|}{|F|}\max_{i\in\mathcal{O}\cap F}\|\boldsymbol{X}_i - \boldsymbol{\mu}\|_2. \end{aligned}\tag{1.8}$$

The fist term of (1.8) is bounded by a term of order $(1-\tau-\varepsilon)^{-1}\big(\sigma\sqrt{p/n} + \sigma(\varepsilon+\tau)\sqrt{\log(1/(\varepsilon+\tau))}\big)$ via the same argument as in one dimension (*cf.* Lemma 6), and the second term is of order at most $(1-\tau-\varepsilon)^{-1}\varepsilon\sigma\big(\sqrt{p} + \sqrt{\log(1/\tau)}\big)$. Therefore, choosing $\tau = \varepsilon$, we conclude that if $\varepsilon < 1/4$ the error rate of the sample mean $\widehat{\boldsymbol{\mu}}_F$ after naive pruning is

$$\frac{\sigma}{1-2\varepsilon}\big(\sqrt{p/n} + \varepsilon\sqrt{\log(1/\varepsilon)} + \varepsilon\sqrt{p}\big).$$

At the end, $\sqrt{p}$ unfortunately is still present behind $\varepsilon$ in our estimation error.

### 1.3.2 Mean as center of symmetry (multidimensional case)

We may study three notions of symmetry for multivariate Gaussian distribution (borrowed from Zuo and Serfling (2000)). Each one proposes a different estimator for the center, a role played by the Gaussian mean under the three symmetries.

#### Central symmetry

A random vector $\boldsymbol{Y}$ is *centrally symmetric* if $\boldsymbol{Y}-\boldsymbol{\theta}$ and $\boldsymbol{\theta}-\boldsymbol{Y}$ have the same distribution where $\boldsymbol{\theta}$ is the center of symmetry. This implies that $\boldsymbol{\theta} = \mathbf{E}[\boldsymbol{Y}]$ and suggests to estimate $\boldsymbol{\theta}$ by the sample mean, however, the sample mean is not robust as we have already outlined.

#### Angular symmetry

A random vector $\boldsymbol{Y}$ is *angularly symmetric* about $\boldsymbol{\theta}$ if $(\boldsymbol{Y}-\boldsymbol{\theta})/\|\boldsymbol{Y}-\boldsymbol{\theta}\|_2$ is centrally symmetric. In our case, to estimate $\boldsymbol{\mu}$, the center of angular symmetry, we may compute a point $\widehat{\boldsymbol{\mu}}_{\mathrm{GM}}$ at which the function $\phi : \boldsymbol{x} \mapsto \sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{x})/\|\boldsymbol{X}_i - \boldsymbol{x}\|_2$ becomes null vector. It turns out that $\phi$ is the gradient of the function $\boldsymbol{x} \mapsto \sum_{i=1}^{n}\|\boldsymbol{X}_i - \boldsymbol{x}\|_2$. The last function is convex and its minimum is attained somewhere in the convex hull of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Consequently $\phi$ is equal to null vector at this point. That is to say

$$\widehat{\boldsymbol{\mu}}_{\mathrm{GM}} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^p} \sum_{i=1}^{n} \|\boldsymbol{X}_i - \boldsymbol{x}\|_2, \tag{1.9}$$

and this formulation of $\widehat{\boldsymbol{\mu}}_{\mathrm{GM}}$ is known as the *geometric median* of the samples. There are various optimization methods for approximating the minimum in (1.9). In particular, via these methods we may reach a point $\widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}$ satisfying

$$\frac{1}{n}\big\|\phi(\widehat{\boldsymbol{\mu}}'_{\mathrm{GM}})\big\|_2 \leq \frac{1}{\sqrt{n}}. \tag{1.10}$$

This condition certifies the existence of a positive constant $C$ such that with high probability we have

$$\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}\|_2 \leq C\big(\sigma\sqrt{p/n} + \sigma\varepsilon\sqrt{p}\big). \tag{1.11}$$

Let us show this informally. Indeed, the random vector $(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}'_{\mathrm{GM}})/\|\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}\|_2$ enjoys the sub-Gaussian convergence rate, and since $\mathrm{Tr}\left(\mathrm{Var}\big[\frac{\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}}{\|\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}\|_2}\big]\right)$ is less that one, there exists $C' > 0$ such that

$$\left\|\mathbf{E}\Big[\frac{\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}}{\|\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}\|_2}\Big]\right\|_2 \leq \frac{1}{n}\big\|\phi(\widehat{\boldsymbol{\mu}}'_{\mathrm{GM}})\big\|_2 + C'\frac{1}{\sqrt{n}} + 2\varepsilon, \tag{1.12}$$

is valid with high probability. For large values of $p$ the random variable $\|\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}\|_2$ concentrates around its mean. This leads to

$$\begin{aligned}\left\|\mathbf{E}\Big[\frac{\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}}{\|\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}\|_2}\Big]\right\|_2 &\approx \left\|\mathbf{E}\Big[\frac{\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}}{\mathbf{E}\|\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}\|_2}\Big]\right\|_2 \\ &= \frac{\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}\|_2}{\mathbf{E}\|\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}\|_2} \\ &\geq \frac{\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}\|_2}{\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}\|_2 + \sqrt{\mathbf{E}\|\boldsymbol{Y}_i - \boldsymbol{\mu}\|_2^2}} = \frac{\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}\|_2}{\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}\|_2 + \sigma\sqrt{p}},\end{aligned} \tag{1.13}$$

where we used the triangle and Jensen's inequalities. Combining (1.10), (1.12) and (1.13), one obtains (1.11) with high probability if for some positive $C''$ we have $\frac{C''}{\sqrt{n}} + 2\varepsilon < 1$. Therefore, we obtain an estimator $\widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}$ of $\boldsymbol{\mu}$ with the convergence rate $\sigma\sqrt{p/n} + \sigma\varepsilon\sqrt{p}$ in high dimension.

This analysis requires that $\varepsilon < \frac{1}{2} - \frac{C''}{2\sqrt{n}}$ which mean that the breakdown point of $\widehat{\boldsymbol{\mu}}'_{\mathrm{GM}}$ is at least $\frac{1}{2} - \frac{C''}{2\sqrt{n}}$. Lopuhaa and Rousseeuw (1991) prove that the geometric median has the benefit of the maximum breakdown point value, namely $\frac{1}{2} - \frac{1}{n}$.

On an related note, under a more general setting, (Lai et al., 2016, Proposition 2.1) construct an oblivious contamination (a contamination model weaker than the adversarial one, *cf.* 2.2.3) for which the error rate of the geometric median cannot be better than $\sigma\varepsilon\sqrt{p}$.

### Halfspace symmetry

A random vector $\boldsymbol{Y}$ is *halfspace symmetric* about $\boldsymbol{\theta}$ if $\mathbf{P}(\boldsymbol{Y} \in H) \geq 1/2$ for every closed halfspace $H$ containing $\boldsymbol{\theta}$. In other words, $\boldsymbol{\theta}$ is the center of the symmetry induced by mass in every direction. Indeed, this generalizes the notion of the median for higher dimensions. This means that for all unitary vector $\boldsymbol{v}$, $\boldsymbol{v}^\top\boldsymbol{\theta}$ is the median of the random variable $\boldsymbol{v}^\top\boldsymbol{Y}$. To translate this notion of center for our sample distribution, we consider the projection of the samples in all the directions, and we look for the point minimizing the difference of empirical masses in both sides of the point for every direction. So, we may define the center as the point minimizing the maximum value of this difference over all the directions. This gives a point $\widehat{\boldsymbol{\mu}}_{\mathrm{TM}}$ belonging to the set

$$\arg\min_{\boldsymbol{x}\in\mathbb{R}^p} \max_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \left| \sum_{i=1}^n \mathbb{1}(\boldsymbol{v}^\top\boldsymbol{X}_i \leq \boldsymbol{v}^\top\boldsymbol{x}) - \sum_{i=1}^n \mathbb{1}(\boldsymbol{v}^\top\boldsymbol{X}_i > \boldsymbol{v}^\top\boldsymbol{x}) \right|,$$

where $\left| \sum_{i=1}^n \mathbb{1}(\boldsymbol{v}^\top\boldsymbol{X}_i \leq \boldsymbol{v}^\top\boldsymbol{x}) - \sum_{i=1}^n \mathbb{1}(\boldsymbol{v}^\top\boldsymbol{X}_i > \boldsymbol{v}^\top\boldsymbol{x}) \right|$ expresses the difference of empirical masses at the two sides of $\boldsymbol{v}^\top\boldsymbol{x}$ when the samples are projected on direction $\boldsymbol{v}$. It can easily be shown that

$$\widehat{\mu}_{\mathrm{TM}} \in \arg\max_{\boldsymbol{x}\in\mathbb{R}^p} \min_{\boldsymbol{v}\in\mathbb{R}^{p-1}} \sum_{i=1}^n \mathbb{1}(\boldsymbol{v}^\top\boldsymbol{X}_i \leq \boldsymbol{v}^\top\boldsymbol{x}).$$

The last set is a convex polytope, and thus we can define $\widehat{\boldsymbol{\mu}}_{\mathrm{TM}}$ in a unique way as the barycentre of this polytope. The point $\widehat{\boldsymbol{\mu}}_{\mathrm{TM}}$ is known as *Tukey's median* ((Tukey, 1975)) of the samples, and represents a generalization of the sample median for higher dimensions. Given a point $\boldsymbol{x}$, the value

$$\min_{\boldsymbol{v}\in\mathbb{R}^{p-1}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\boldsymbol{v}^\top \boldsymbol{X}_i \leq \boldsymbol{v}^\top \boldsymbol{x})$$

is called the *Tukey's halfspace depth* of $\boldsymbol{x}$ with respect to the samples, and so, Tukey's median is determined as the barycentre of the points of maximum depth with respect to the samples.

Chen et al. (2018) prove that under some conditions the convergence rate of Tukey's median is of order $\sigma\sqrt{p/n}+\sigma\varepsilon$. More precisely, under Huber's contamination, they establish that for some positive constants $C_1$ and $C_2$, given $\delta \in (0, 1/2)$, $\widehat{\boldsymbol{\mu}}_{\mathrm{TM}}$ satisfies

$$\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{\mathrm{TM}}\|_2 \leq \sigma\Phi^{-1}\left(\frac{1}{2}+\frac{\varepsilon}{1-\varepsilon}+C_1\sqrt{\frac{p}{n}}+C_2\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

with probability at least $1-2\delta$. The above formula is meaningful only if

$$\frac{1}{2}+\frac{\varepsilon}{1-\varepsilon}+C_1\sqrt{\frac{p}{n}}+C_2\sqrt{\frac{\log(1/\delta)}{n}} < 1.$$

This implies that $\varepsilon < \frac{1}{3} - C\sqrt{\frac{p}{n}}$ for some positive constant $C$, and therefore the breakdown point of Tukey's median is at least $\frac{1}{3} - C\sqrt{\frac{p}{n}}$. Furthermore, results from (Liu et al., 2017) show that in general position the breakdown point of the Tukey's median is at most $\frac{1}{3} - \frac{2}{9}\frac{p-3}{n}$.

**Remark 5.** *We finally succeeded in avoiding the factor $\sqrt{p}$ behind $\varepsilon$ in the error rate. However, computation of Tukey's median demands a number of operations of order $n^p$. Indeed, computation of the depth of any point in the space with respect to the samples costs already an exponential in $p$ number of operations.*

**Remark 6.** *Central symmetry implies angular symmetry and angular symmetry implies halfspace symmetry. The foregoing material suggests that broader is the adopted notion of symmetry, more robust is the method estimating the mean as center of symmetry, and harder is the computation of the corresponding estimator.*

**Remark 7.** *The symmetry based estimators have two benefits: first, they do not require the knowledge of the covariance matrix and the contamination rate; second, they are translation, uniform scaling and orthogonal transformation equivariant.*

### 1.3.3 Minimum distance estimator (multidimensional case)

The skeleton estimate can be built for higher dimensions in the same way as for one dimension. We need just to specify the construction of the candidates' set. Given a naive estimator

of $\boldsymbol{\mu}$, for instance the estimator $\widehat{\boldsymbol{\mu}}_F$ proposed by the filtering method, we know that $\boldsymbol{\mu}$ belongs with high probability to the ball of center $\widehat{\boldsymbol{\mu}}_F$ and radius $C\sqrt{p}$ where $C$ is a positive constant. A covering set $M_\gamma = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{N_\gamma}\}$ can be constructed on this ball with $N_\gamma \leq (2C\sqrt{p}/\gamma)^p$. We consider the family $(f_i)_{i\in\{1,\dots,N_\gamma\}}$ where $f_i$ is the p.d.f. of $\mathcal{N}(\boldsymbol{\mu}_i, \sigma\mathbf{I}_p)$. Let $\widehat{\nu}$ be the empirical measure defined by the samples $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$. We define

$$\widehat{k} = \underset{k\in\{1,\dots,N_\gamma\}}{\arg\min} \max_{A\in\mathcal{A}_\gamma} \left| \int_A f_k - \widehat{\nu}(A) \right|,$$

where $\mathcal{A}_\gamma = \big\{\{x : f_i(x) > f_j(x)\} : i, j \in \{1, ..., N_\gamma\}\big\}$ with $\gamma = \sigma\sqrt{p/n}$. Using arguments similar to those of the one dimensional setting, the skeleton estimate $\boldsymbol{\mu}_{\widehat{k}}$ satisfies with high probability

$$\|\boldsymbol{\mu} - \boldsymbol{\mu}_{\widehat{k}}\|_2 \leq 10\sigma\left(C\sqrt{\frac{p\log(n)}{n}} + 2\varepsilon\right),$$

where $C$ is a positive constant and $C\sqrt{p\log(n)/n} + 2\varepsilon \leq 0.49$.

**Remark 8.** *The set $\mathcal{A}_\gamma$ is constructed in time exponential in $p$. For this reason, the skeleton estimate as Tukey's median is not computationally tractable. In fact, both estimators are determined by a minimax optimization problem and this suggests that exponential complexity is indispensable to achieve a error rate of order $\sigma\sqrt{p/n} + \sigma\varepsilon$ (without the factor $\sqrt{p}$ multiplying $\varepsilon$). This raises the trade off challenge between the computational efficiency and statistical accuracy.*

### 1.3.4 Leveraging covariance matrix

Applying the same technique introduced in Section 1.2.4, it is easily established that

$$\sigma\Phi^{-1}\Big(\frac{1}{2} + \frac{\varepsilon}{2(1-\varepsilon)}\Big) + \frac{\sigma\sqrt{p}}{\sqrt{n}}$$

is a lower bound for estimating the mean under Huber's contamination model when the non-contaminated samples follow $\mathcal{N}(\boldsymbol{\mu}, \sigma\mathbf{I}_p)$. This shows that Tukey's median is a minimax estimator of the multivariate Gaussian mean when the samples are adversarially corrupted. As we have outlined, there is a significant gap between the convergence rate of the computational tractable methods presented here and that of the optimal method, *i.e.*, Tukey's median, which is not tractable. In recent years, one of the main challenges in robust estimation has been to fill this gap and various methods have been invented. The common point of almost all of these methods is to use the knowledge of the covariance matrix of the reference Gaussian distribution.

We try to give the high level scheme of these methods. Consider a more general setting where the initial samples $(\boldsymbol{Y}_i)_{i\in\{1,\dots,n\}}$ are drawn from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The key idea is the following.

If we have (contaminated) samples with a sample covariance matrix close to $\boldsymbol{\Sigma}$, then we can certify that the empirical mean of these samples is close to $\boldsymbol{\mu}$. These methods use some or all of the following ingredients:

1. Majority of the inliers ($\boldsymbol{X}_i$ with $i \in \mathcal{I}$) lie in a distance of order $\sqrt{p}$ from $\boldsymbol{\mu}$. Via a filtering method, we can obtain a set containing only samples with such property (see *e.g.*, Lemma 4).

2. Let $\overline{\boldsymbol{\mu}}_S$ be the sample mean of the set $\{\boldsymbol{Y}_i : i \in S\}$. For all subsets $S \subset \{1, \dots, n\}$ of size at least $(1-\varepsilon)n$, $\|\overline{\boldsymbol{\mu}}_S - \boldsymbol{\mu}\|_2$ is bounded by a term of order $\sqrt{p/n} + \varepsilon\sqrt{\log(1/\varepsilon)}$ (see *e.g.*, Lemma 6).

3. Let $\overline{\boldsymbol{\Sigma}}_S$ be the sample covariance matrix of the set $\{\boldsymbol{Y}_i : i \in S\}$. For all subsets $S \subset \{1, \dots, n\}$ of size at least $(1-\varepsilon)n$, $\|\overline{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}\|_{\mathrm{op}}$ is bounded by a term of order $\max\left(\sqrt{p/n}, p/n\right) + \varepsilon\log(1/\varepsilon)$ where $\|.\|_{\mathrm{op}}$ denotes the operator norm for matrices (see *e.g.*, Lemma 7).

4. Let $\widehat{\boldsymbol{\mu}}_S$ and $\widehat{\boldsymbol{\Sigma}}_S$ be the sample mean and sample covariance matrix of the set $\{\boldsymbol{X}_i : i \in S\}$. We can show that $\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_S\|_2^2$ is controlled by a term depending on $\varepsilon\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_S\|_{\mathrm{op}}$ (see *e.g.*, Proposition 6).

Hence, now the goal would be to find a set $S$ with a enough small $\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_S\|_{\mathrm{op}}$. We can recapitulate various existing algorithms for realizing this goal in two categories:

- Filtering algorithms: they filter samples in order to control $\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_S\|_{\mathrm{op}}$. For filtering they deploy spectral methods on the matrix $\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_S$ to detect in which directions there is more anomaly and then they remove the samples responsible for these anomalies. In this category, we can cite methods introduced in (Lai et al., 2016), (Diakonikolas et al., 2017), (Diakonikolas et al., 2018a) and our method presented in Chapter 3.

- Weighting algorithms: they attribute to each sample $\boldsymbol{X}_i$ a weight $\omega_i$ (such that $\sum_{i=1}^n \omega_i = 1$) and try to find weights in order to minimize $\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_\omega\|_{\mathrm{op}}$ where $\widehat{\boldsymbol{\Sigma}}_\omega$ is the weighted sample covariance matrix. Finally, the weighted sample mean $\widehat{\boldsymbol{\mu}}_\omega$ determined by such weights appears to be a robust estimator of $\boldsymbol{\mu}$. For this problem, different optimization methods are proposed in (Diakonikolas et al., 2016a), (Cheng et al., 2019a), (Dalalyan and Minasyan, 2020), and (Cheng et al., 2020).

All of these algorithms succeed to achieve in polynomial time a error rate with dependency $\varepsilon\sqrt{\log(1/\varepsilon)}$ instead of $\varepsilon\sqrt{p}$. Diakonikolas et al. (2016b) provide a strong evidence that the dependency $\varepsilon\sqrt{\log(1/\varepsilon)}$ in error rate is necessary for any robust polynomial time algorithm under the adversarial model. Indeed, they demonstrate that any *Statistical Query* algorithm estimating $\boldsymbol{\mu}$ under the adversarial model with dependency $\varepsilon$ (instead of $\varepsilon\sqrt{\log(1/\varepsilon)}$) in error rate needs to have access to sample moments of degree at least $\log(1/\varepsilon)^{1/4}$, and this costs at least $p^{\log(1/\varepsilon)^{1/4}}$ operations, namely, an exponential number of operations in $p$. The *Statistical*

*Query* algorithms, first introduced in (Kearns, 1998), are a class of algorithms that are allowed to query expectations of functions of the reference distribution modulo an error rather than directly access the samples. This can model a wide range of algorithms in statistics and learning theory.

The four ingredients mentioned above can be generalized for samples drawn from other classes of distributions (such as the sub-Gaussian distributions or distributions with bounded moments), albeit with possibly different bounds corresponding to their respective concentration properties. For instance, Steinhardt et al. (2017) call a general form of the second ingredient *resilience*, Diakonikolas et al. (2020) call a general form of the second and third ingredients together *stability*, or the three first ingredients together with some other conditions are called *goodness* in (Diakonikolas et al., 2017). These generalizations make the algorithms with this approach extensible for robust mean estimation of a more general class of reference distributions.

## 1.4 Prior work

Robust estimation is an area of active research in Statistics since at least five decades (Donoho and Gasko, 1992; Donoho and Huber, 1983; Huber and Ronchetti, 2011; Huber, 1964; Maronna et al., 2006; Rousseeuw et al., 2011; Rousseeuw and Hubert, 1999; Tukey, 1975). Until very recently, theoretical guarantees were almost exclusively formulated in terms of the notions of breakdown point, sensitivity curve, influence function, etc. In 2015, a new line of research in this area started by the work of Chen et al. (2018) who considered the problem of estimating the mean and the covariance matrix of a Gaussian distribution in the high-dimensional setting under Huber's contamination model by studying the rate of convergence of the minimax risk as a function of the sample size $n$, dimension $p$ and the rate of contamination $\varepsilon$. The authors showed that the minimax optimal rate is attained by Tukey's median which is not computationally tractable whereas classical tractable robust methods such as the componentwise median do not achieve this rate.

This phenomenon stimulated researches for computational tractable methods with optimal rate. First algorithms were proposed by computer science community. In this community, the error rates are expressed differently: it contains only the term depending on $\varepsilon$ while the sample size necessary for obtaining this error rate is also considered. For example the error rates for the methods presented in the last section reduce to their term depending on $\varepsilon$, namely, $\varepsilon\sqrt{p}$, $\varepsilon\sqrt{\log(1/\varepsilon)}$ or $\varepsilon$ and the other term which depends on $p$ and $n$ determines how many samples the methods need to achieve this error rate. Thus, there are two kinds of complexity in this approach: the computational complexity and the sample complexity.

For the Gaussian mean, Lai et al. (2016) proposed a recursive algorithm halving the dimension at each step based on a spectral analysis of the sample covariance matrix. Their algorithm achieves the error rate $\varepsilon\sqrt{\log(p)}$ under the oblivious contamination model (see the

definition in Section 2.2.3). Diakonikolas et al. (2016a) designed two algorithms with error rate $\varepsilon\sqrt{\log(1/\varepsilon)}$: one based on an iterative spectral filtering method and one based on a convex programming defined by the weighted samples. Cheng et al. (2019a) transformed this convex programming to a semi-definite programming (SDP) and proposed a faster iterative method with same error rate but with optimal sample complexity, *i.e.*, sample size of order $d/\varepsilon^2$. Dalalyan and Minasyan (2020) introduce a different iterative method based on SDP by giving statistical guarantees for the error rate $\sqrt{\mathrm{Tr}(\boldsymbol{\Sigma})/n} + \varepsilon\sqrt{\|\boldsymbol{\Sigma}\|_{\mathrm{op}}\log(1/\varepsilon)}$ in expectation and in probability. Finally, Cheng et al. (2021) prove that a suitable vesrion of the gradient descent method on minimization of $\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_\omega\|$ converges to an estimator $\widehat{\boldsymbol{\mu}}_\omega$ of $\boldsymbol{\mu}$ with same risk $\varepsilon\sqrt{\log(1/\varepsilon)}$. The only estimator giving a error rate of order $\varepsilon$ in polynomial time (under oblivious contamination model) is proposed in (Diakonikolas et al., 2017) based on a spectral filtering method, however, it has a high sample complexity. Some other methods can be found in (Balakrishnan et al., 2017; Diakonikolas and Kane, 2019; Diakonikolas et al., 2020; Dong et al., 2019). All of these methods are extensible for sub-Gaussian distributions. In Chapter 3, we design a new estimator inspired by spectral filtering ideas in Lai et al. (2016) and Diakonikolas et al. (2017).

Robust mean estimation is studied also for the distributions with bounded moments. Many of the algorithms for Gaussian mean can be generalized for the distributions with bounded moments, see (Cheng et al., 2019a; Diakonikolas et al., 2017; Diakonikolas and Kane, 2019; Diakonikolas et al., 2020; Dong et al., 2019; Lai et al., 2016). For this problem, other algorithms with the idea of weighting the sample are proposed, see *e.g.*, (Depersin and Lecué, 2022; Liu et al., 2020a; Prasad et al., 2019; Steinhardt et al., 2017). In particular, Depersin and Lecué (2022) give a computational tractable estimator under bounded second moment assumption with the optimal error rate $\sqrt{\mathrm{Tr}(\boldsymbol{\Sigma})/n} + \sqrt{\|\boldsymbol{\Sigma}\|_{\mathrm{op}}\varepsilon}$, while the other mentioned methods are at best nearly optimal. All we have seen so far belong to the category of the approaches leveraging covariance matrix presented in Section 1.3.4. Minsker (2015) analyses the geometric median under bounded moments assumptions and proves that it has sub-optimal error rate. Lugosi and Mendelson (2021) extend the trimmed mean to high-dimensional setting and show that their extension enjoys optimal error rate, though, it cannot be computed in polynomial time. There are various other approaches, see *e.g.*, (Bahmani, 2021; Depersin and Lecué, 2021; Minsker, 2018b; Minsker and Ndaoud, 2021; Prasad et al., 2020).

A closely related problem is that of the mean estimation with sub-Gaussian rates for heavy tailed distributions. For one dimension, the so-called *median-of-means* estimator, introduced independently in (Nemirovsky and Yudin, 1983), (Jerrum et al., 1986) and (Alon et al., 1996), has the optimal performance. Other methods exist for one dimension, see (Catoni, 2012; Devroye et al., 2016; Minsker and Ndaoud, 2021) and for a more complete overview, we refer the reader to the survey (Lugosi and Mendelson, 2019a). For high-dimension, various estimators are proposed such as a median-of-means based estimator (Joly et al., 2017), geometric median (Hsu and Sabato, 2016; Minsker, 2015) or (Catoni and Giulini, 2017, 2018). The first estimator with optimal performance is inroduced in (Lugosi and Mendelson, 2019b) which is

not computationally efficient. Hopkins (2020) formulates a SDP relaxation of the estimator of Lugosi and Mendelson (2019b) using *sum-of-squares* methods and designs a polynomial time algorithm with the optimal sub-Gaussian rates. Cherapanamjeri et al. (2019) combine this SDP formulation by a non-convex gradient descent procedure and improve the computational complexity of the last algorithm. Depersin and Lecué (2022) combine the idea of SDP and median-of-means and construct another tractable optimal estimator. Lei et al. (2019) propose a different optimal estimator not using SDP (SDP causes a high runtime). Furthermore, unified robust approaches against outliers and heavy tailed distributions are considered in (Bahmani, 2021; Depersin and Lecué, 2021; Depersin and Lecué, 2022; Minsker, 2015; Minsker and Ndaoud, 2021; Prasad et al., 2019).

Concerning the robust mean estimation of discrete distributions, Chen et al. (2020); Jain and Orlitsky (2019); Qiao and Valiant (2018); Steinhardt et al. (2017) studied the case of group-contamination where the observations are batches of samples with $\varepsilon$-fraction of batches contaminated by an adversary. In Chapter 2, we study this problem under the normal adversarial model.

In addition, high-dimensional robust estimation is studied under a very different framework, Generative Adversarial Networks (GAN), and robust estimators based on GANs are constructed for various statistical tasks, see (Gao et al., 2018, 2020; Wang and Tan, 2021; Zhu et al., 2022). This line of work is close in spirit to the minimum distance estimator discussed in 1.3.3.

Beyond Huber and adversarial contamination models, other contamination models are considered in the literature such as parameter contamination models (Bhatia et al., 2017; Carpentier et al., 2018; Collier and Dalalyan, 2019) (cf. Section 2.2.4), oblivious contamination (Diakonikolas et al., 2017; Feng et al., 2014; Lai et al., 2016) (cf. Section 2.2.3) , Wasserstein and total variation contaminations (Diakonikolas et al., 2016b; Zhu et al., 2019, 2020).

Robust estimation is investigated for other statistical tasks as covariance matrix estimation (many paper studying mean estimation treat also covariance matrix estimation, see also (Cheng et al., 2019b)), linear regression (Alquier and Gerber, 2020; Audibert and Catoni, 2011; Bakshi and Prasad, 2021; Cherapanamjeri et al., 2020; Chinot, 2020; Dalalyan and Thompson, 2019; Depersin, 2020; Diakonikolas et al., 2019b; Gao, 2020; Liu et al., 2018; Pensia et al., 2020; Sasai and Fujisawa, 2021), risk minimization (Chinot et al., 2018; Chinot et al., 2020), stochastic optimization (Diakonikolas et al., 2019a), classification (Lecué et al., 2020), density estimation (Jain and Orlitsky, 2021; Liu and Gao, 2019), etc.

## 1.5 Contributions

We summarize our contributions in two sections. The first section is devoted to the problem of estimating the mean of a distribution supported by the $k$-dimensional probability simplex in the setting where an $\varepsilon$ fraction of observations are subject to adversarial corruption. A

simple particular example is the problem of estimating the distribution of a discrete random variable. Assuming that the discrete variable takes $k$ values, the unknown parameter $\boldsymbol{\theta}$ is a $k$-dimensional vector belonging to the probability simplex. There we describe also the various models of contamination. Chapter 2 exposes a complete version of the contribution. This work is presented in (Bateni and Dalalyan, 2020).

The second section summarizes our contributions to the problem of robust estimation of the mean vector of a Gaussian distribution. We introduce an estimator based on spectral dimension reduction (SDR) and establish a finite sample upper bound on its error that is minimax-optimal up to a logarithmic factor. Furthermore, we prove that the (asymptotic) breakdown point of the SDR estimator is equal to $1/2$, the highest possible value of the breakdown point. In addition, the SDR estimator is equivariant by similarity transforms and has low computational complexity. This work is developed in Chapter 3 and is presented in (Bateni et al., 2022).

### 1.5.1 Contamination models and robust estimation on the probability simplex

Assume $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are $n$ independent and identically distributed random variables taking their values in the $k$-dimensional probability simplex $\Delta^{k-1} = \{\boldsymbol{v} \in \mathbb{R}^k_+ : v_1 + \ldots + v_k = 1\}$. Our goal is to estimate the unknown vector $\boldsymbol{\theta} = \mathbf{E}[\boldsymbol{X}_i]$ in the case where the observations are contaminated by outliers. An important particular case is the estimation of the distribution of a discrete random variable $\boldsymbol{X}$ taking $k$ distinct values. In this particular case, $\boldsymbol{X}_i$'s take values in $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_k\}$, the set of the vectors of the canonical basis, which are also the extreme points of the simplex $\Delta^{k-1}$.

In this introduction, to convey the main messages, we limit ourselves to the Huber contamination model, although our results apply to the more general adversarial contamination. Huber's contamination model assumes that there are two probability measures $\boldsymbol{P}$, $\boldsymbol{Q}$ on $\Delta^{k-1}$ and a real $\varepsilon \in [0, 1/2)$ such that $\boldsymbol{X}_i$ is drawn from

$$\boldsymbol{P}_i = (1-\varepsilon)\boldsymbol{P} + \varepsilon \boldsymbol{Q}, \qquad \forall i \in \{1, \dots, n\}.$$

This amounts to assuming that $(1-\varepsilon)$-fraction of observations, called inliers, are drawn from a reference distribution $\boldsymbol{P}$, whereas $\varepsilon$-fraction of observations are outliers and are drawn from another distribution $\boldsymbol{Q}$. In general, all the three parameters $\boldsymbol{P}$, $\boldsymbol{Q}$ and $\varepsilon$ are unknown. The parameter of interest is some functional (such as the mean, the standard deviation, etc.) of the reference distribution $\boldsymbol{P}$, whereas $\boldsymbol{Q}$ and $\varepsilon$ play the role of nuisance parameters.

When the unknown parameter lives on the probability simplex, there are many appealing ways of defining the risk. We focus on the following three metrics: total-variation, Hellinger

and $\mathbb{L}^2$ distances[3]

$$
\begin{aligned}
d_{\mathsf{TV}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &:= {}^1\!/_2 \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1, \\
d_{\mathsf{H}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &:= {}^1\!/_{\sqrt{2}} \|\widehat{\boldsymbol{\theta}}^{1/2} - \boldsymbol{\theta}^{1/2}\|_2, \\
d_{\mathbb{L}^2}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &:= \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2.
\end{aligned}
$$

The Hellinger distance above is well defined when the estimator $\widehat{\boldsymbol{\theta}}$ is non-negative, which will be the case throughout this work. We will further assume that the dimension $k$ may be large, but the vector $\boldsymbol{\theta}$ is $s$-sparse, for some $s \leq k$, *i.e.* $\#\{j : \theta_j \neq 0\} \leq s$. Our main interest is in constructing confidence regions and evaluating the minimax risk

$$
\mathfrak{R}_{\square}(n, k, s, \varepsilon) := \inf_{\bar{\boldsymbol{\theta}}_n} \sup_{\boldsymbol{P}, \boldsymbol{Q}} \mathbf{E}[d_{\square}(\bar{\boldsymbol{\theta}}_n, \boldsymbol{\theta})],
$$

where the *inf* is over all estimators $\bar{\boldsymbol{\theta}}_n$ built upon the observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{iid}}{\sim} (1-\varepsilon)\boldsymbol{P} + \varepsilon \boldsymbol{Q}$ and the *sup* is over all distributions $\boldsymbol{P}$, $\boldsymbol{Q}$ on the probability simplex such that the mean $\boldsymbol{\theta}$ of $\boldsymbol{P}$ is $s$-sparse. The subscript $\square$ of $\mathfrak{R}$ above refers to the distance used in the risk, so that $\square$ is TV, H, or $\mathbb{L}^2$.

### Various models of contamination

Different mathematical frameworks have been used in the literature to model the outliers. We present here five of them, from the most restrictive one to the most general, and describe their relationship. We present these frameworks in the general setting when the goal is to estimate the parameter $\boldsymbol{\theta}^*$ of a reference distribution $\boldsymbol{P}_{\boldsymbol{\theta}^*}$ when $\varepsilon$ proportion of the observations are outliers.

**Huber's contamination** The most popular framework for studying robust estimation methods is perhaps the one of Huber's contamination. In this framework, there is a distribution $\boldsymbol{Q}$ defined on the same space as the reference distribution $\boldsymbol{P}_{\boldsymbol{\theta}^*}$ such that all the observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent and drawn from the mixture distribution $\boldsymbol{P}_{\varepsilon, \boldsymbol{\theta}^*, \boldsymbol{Q}} := (1-\varepsilon)\boldsymbol{P}_{\boldsymbol{\theta}^*} + \varepsilon \boldsymbol{Q}$.

This corresponds to the following mechanism: one decides with probabilities $(1-\varepsilon, \varepsilon)$ whether a given observation is an inlier or an outlier. If the decision is made in favor of being inlier, the observation is drawn from $\boldsymbol{P}_{\boldsymbol{\theta}^*}$, otherwise it is drawn from $\boldsymbol{Q}$. More formally, if we denote by $\widehat{O}$ the random set of outliers, then conditionally to $\widehat{O} = O$,

$$
\{\boldsymbol{X}_i : i \notin O\} \overset{\text{iid}}{\sim} \boldsymbol{P}_{\boldsymbol{\theta}^*}, \ \{\boldsymbol{X}_i : i \in O\} \overset{\text{iid}}{\sim} \boldsymbol{Q}, \ \{\boldsymbol{X}_i : i \in O\} \perp\!\!\!\perp \{\boldsymbol{X}_i : i \notin O\}, \tag{1.14}
$$

for every $O \subset \{1, \ldots, n\}$. Furthermore, for every subset $O$ of the observations, we have

[3] We write $\|\boldsymbol{u}\|_q = (\sum_{j=1}^k |u_j|^q)^{1/q}$ and $\boldsymbol{u}^q = (u_1^q, \ldots, u_k^q)$ for any $\boldsymbol{u} \in \mathbb{R}_+^k$ and $q > 0$.

$\boldsymbol{P}(\widehat{O} = O) = (1-\varepsilon)^{n-|O|}\varepsilon^{|O|}$. We denote by[4] $\mathcal{M}_n^{\mathsf{HC}}(\varepsilon, \boldsymbol{\theta}^*)$ the set of joint probability distributions $\boldsymbol{P}_n$ of the random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ satisfying the foregoing condition.

**Huber's deterministic contamination** The set of outliers as well as the number of outliers in Huber's model of contamination are random. This makes it difficult to compare this model to the others that will be described later in this section. To cope with this, we define here another model, termed Huber's deterministic contamination. As its name indicates, this new model has the advantage of containing a deterministic number of outliers, in the same time being equivalent to Huber's contamination in a sense that will be made precise below.

We say that the distribution $\boldsymbol{P}_n$ of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ belongs to the Huber's deterministic contamination model denoted by $\mathcal{M}_n^{\mathsf{HDC}}(\varepsilon, \boldsymbol{\theta}^*)$, if there are a set $O \subset \{1, \ldots, n\}$ of cardinality at most $n\varepsilon$ and a distribution $\boldsymbol{Q}$ such that (1.14) is true. The apparent similarity of models $\mathcal{M}_n^{\mathsf{HC}}(\varepsilon, \boldsymbol{\theta}^*)$ and $\mathcal{M}_n^{\mathsf{HDC}}(\varepsilon, \boldsymbol{\theta}^*)$ can also be formalized mathematically in terms of the orders of magnitude of minimax risks. To ease notation, we let $R_d^{\square}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}})$ to be the worst-case risk of an estimator $\widehat{\boldsymbol{\theta}}$, where $\square$ is either HC or HDC. More precisely, for $\mathcal{M}_n^{\square}(\varepsilon, \Theta) := \cup_{\boldsymbol{\theta}\in\Theta}\mathcal{M}_n^{\square}(\varepsilon, \boldsymbol{\theta})$, we set[5]

$$R_d^{\square}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}) := \sup_{P_n \in \mathcal{M}_n^{\square}(\varepsilon,\Theta)} \mathbf{E}[d(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)].$$

This definition assumes that the parameter space $\Theta$ is endowed with a pseudo-metric $d : \Theta\times\Theta \to \mathbb{R}_+$. When $\Theta = \{\boldsymbol{\theta}^*\}$ is a singleton, we write $R_{d,n}^{\square}(\varepsilon, \boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}})$ instead of $R_d^{\square}(n, \varepsilon, \{\boldsymbol{\theta}^*\}, \widehat{\boldsymbol{\theta}})$.

**Proposition 1.** *Let $\widehat{\boldsymbol{\theta}}_n$ be an arbitrary estimator of $\boldsymbol{\theta}^*$. For any $\varepsilon \in (0, 1/2)$,*

$$R_d^{HC}(n, \varepsilon, \boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}_n) \le R_{d,n}^{HDC}(2\varepsilon, \boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}_n) + e^{-n\varepsilon/3} R_{d,n}^{HDC}(1, \boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}_n),$$
$$\sup_{\boldsymbol{P}_n \in \mathcal{M}_n^{HC}(\varepsilon,\boldsymbol{\theta}^*)} r\mathbf{P}\big(d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}^*) > r\big) \le R_{d,n}^{HDC}(2\varepsilon, \boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}_n) + re^{-n\varepsilon/3}.$$

Denote by $\mathcal{D}_\Theta$ the diameter of $\Theta$, $\mathcal{D}_\Theta := \max_{\boldsymbol{\theta},\boldsymbol{\theta}'} d(\boldsymbol{\theta}, \boldsymbol{\theta}')$. Proposition 1 implies that

$$\inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathsf{HC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) \le \inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathsf{HDC}}(n, 2\varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) + e^{-n\varepsilon/3}\mathcal{D}_\Theta.$$

When $\Theta$ is bounded, the last term is typically of smaller order than the minimax risk over $\mathcal{M}_n^{\mathsf{HDC}}(2\varepsilon, \Theta)$. Therefore, the minimax rate of estimation in Huber's model is not slower than the minimax rate of estimation in Huber's deterministic contamination model. This entails that a lower bound on the minimax risk established in HC-model furnishes a lower bound in HDC-model.

**Oblivious contamination** A third model of contamination that can be of interest is the oblivious contamination. In this model, it is assumed that the set $O$ of cardinality $o$ and the joint

[4] The superscript HC refers to the Huber's contamination

[5] The subscript $d$ refers to the distance $d$ used in the definition of the risk.

distribution $\boldsymbol{Q}_O$ of outliers are determined in advance, possibly based on the knowledge of the reference distribution $\boldsymbol{P}_{\boldsymbol{\theta}^*}$. Then, the outliers $\{\boldsymbol{X}_i : i \in O\}$ are drawn randomly from $\boldsymbol{Q}_O$ independently of the inliers $\{\boldsymbol{X}_i : i \in O^c\}$. The set of all the joint distributions $\boldsymbol{P}_n$ of random variables $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ generated by such a mechanism will be denoted by $\mathcal{M}_n^{\mathsf{OC}}(\varepsilon, \boldsymbol{\theta}^*)$. The model of oblivious contamination is strictly more general than that of Huber's deterministic contamination, since it does not assume that the outliers are i.i.d. Therefore, the minimax risk over $\mathcal{M}_n^{\mathsf{OC}}(\varepsilon, \Theta)$ is larger than the minimax risk over $\mathcal{M}_n^{\mathsf{HDC}}(\varepsilon, \Theta)$:

$$\inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathsf{HDC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) \leq \inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathsf{OC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n).$$

The last inequality holds true for any set $\Theta$, any contamination level $\varepsilon \in (0, 1)$ and any sample size.

**Parameter contamination** In the three models considered above, the contamination acts on the observations. One can also consider the case where the parameters of the distributions of some observations are contaminated. More precisely, for some set $O \subset \{1, \dots, n\}$ selected in advance (but unobserved), the outliers $\{\boldsymbol{X}_i : i \in O\}$ are independent and independent of the inliers $\{\boldsymbol{X}_i : i \in O^c\}$. Furthermore, each outlier $\boldsymbol{X}_i$ is drawn from a distribution $\boldsymbol{Q}_i = \boldsymbol{P}_{\boldsymbol{\theta}_i}$ belonging to the same family as the reference distribution, but corresponding to a contaminated parameter $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}^*$. Thus, the joint distribution of the observations can be written as $(\bigotimes_{i \in O^c} \boldsymbol{P}_{\boldsymbol{\theta}^*}) \otimes (\bigotimes_{i \in O} \boldsymbol{P}_{\boldsymbol{\theta}_i})$. The set of all such distributions $\boldsymbol{P}_n$ will be denoted by $\mathcal{M}_n^{\mathsf{PC}}(\varepsilon, \boldsymbol{\theta}^*)$, where PC refers to "parameter contamination".

**Adversarial contamination** The last model of contamination we describe in this work, the adversarial contamination, is the most general one. It corresponds to the following two-stage data generation mechanism. In a first stage, iid random variables $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ are generated from a reference distribution $\boldsymbol{P}_{\boldsymbol{\theta}^*}$. In a second stage, an adversary having access to $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ chooses a (random) set $\widehat{O}$ of (deterministic) cardinality $s$ and arbitrarily modifies data points $\{\boldsymbol{Y}_i : i \in \widehat{O}\}$. The resulting sample, $\{\boldsymbol{X}_i : i = 1, \dots, n\}$, is revealed to the Statistician. In this model, we have $\boldsymbol{X}_i = \boldsymbol{Y}_i$ for $i \notin \widehat{O}$. However, since $\widehat{O}$ is random and potentially dependent of $\boldsymbol{Y}_{1:n}$, it is not true that conditionally to $\widehat{O} = O$, $\{\boldsymbol{X}_i : i \in O^c\}$ are i.i.d. drawn from $\boldsymbol{P}_{\boldsymbol{\theta}^*}$ (for any deterministic set $O$ of cardinality $o$).

We denote by $\mathcal{M}_n^{\mathsf{AC}}(\varepsilon, \boldsymbol{\theta}^*)$ the set of all the joint distributions $\boldsymbol{P}_n$ of all the sequences $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ generated by the aforementioned two-stage mechanism. This set $\mathcal{M}_n^{\mathsf{AC}}(\varepsilon, \boldsymbol{\theta}^*)$ is larger than all the four sets of contamination introduced in this section. Therefore, the following inequalities hold:

$$\inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathsf{PC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) \leq \inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathrm{OC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) \leq \inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathsf{AC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n),$$

for any $n$, $\varepsilon$, $\Theta$ and any distance $d$.

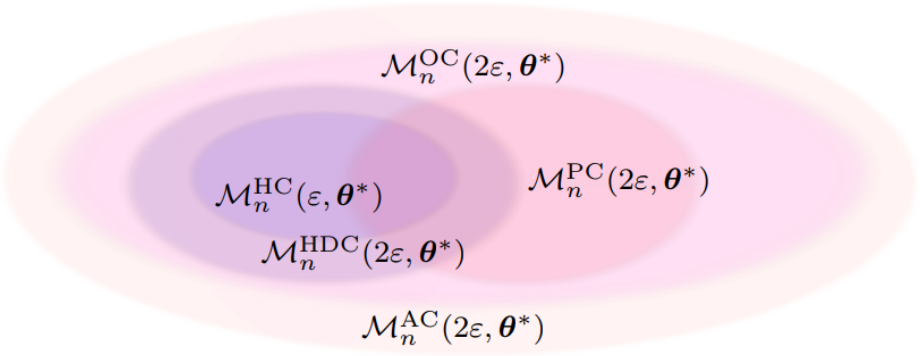


Figure 1.4: Visual representation of the hierarchy between various contamination models. Note that the inclusion of $\mathcal{M}_n^{\mathsf{HC}}(\varepsilon, \boldsymbol{\theta}^*)$ in $\mathcal{M}_n^{\mathsf{HDC}}(2\varepsilon, \boldsymbol{\theta}^*)$ is somewhat heuristic, based on the relation on the worst-case risks reported in Proposition 1

### Minimax rates on the "sparse" simplex and confidence regions

We now specialize the general setting of Section 1.5.1 to a reference distribution $\boldsymbol{P}$, with expectation $\boldsymbol{\theta}^*$, defined on the simplex $\Delta^{k-1}$. Along with this reference model describing the distribution of inliers, we will use different models of contamination. More precisely, we will establish upper bounds on worst-case risks of the sample mean in the most general, adversarial, contamination setting. Then, matching lower bounds will be provided for minimax risks under Huber's contamination.

**Upper bounds: worst-case risk of the sample mean** We denote by $\Delta_s^{k-1}$ the set of all $\boldsymbol{v} \in \Delta^{k-1}$ having at most $s$ non-zero entries.

**Theorem 1.** *For every triple of positive integers $(k, s, n)$ and for every $\varepsilon \in [0, 1]$, the sample mean $\bar{\boldsymbol{X}}_n := \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$ satisfies*

$$R_{\mathsf{TV}}^{\mathsf{AC}}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{X}}_n) \le (s/n)^{1/2} + 2\varepsilon,$$
$$R_{\mathsf{H}}^{\mathsf{AC}}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{X}}_n) \le (s/n)^{1/2} + \sqrt{2}\,\varepsilon^{1/2},$$
$$R_{\mathbb{L}^2}^{\mathsf{AC}}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{X}}_n) \le (1/n)^{1/2} + \sqrt{2}\,\varepsilon.$$

An unexpected and curious phenomenon unveiled by this theorem is that all the three rates are different. As a consequence, the largest possible number of outliers, $o_d^*(n, s)$, that does not impact the minimax rate of estimation of $\boldsymbol{\theta}^*$ crucially depends on the considered distance $d$. Taking into account the relation $\varepsilon = o/n$, we get

$$o_{\mathsf{TV}}^*(n, s) \asymp (ns)^{1/2}, \quad o_{\mathsf{H}}^*(n, s) \asymp s, \quad o_{\mathbb{L}^2}^*(n, s) \asymp n^{1/2}.$$

**Lower bounds on the minimax risk** A natural question, answered in the next theorem, is how tight are the upper bounds obtained in the last theorem. More importantly, one can wonder whether there is an estimator that has a worst-case risk of smaller order than that of the sample mean.

**Theorem 2.** *There are universal constants $c > 0$ and $n_0$, such that for any integers $k \geq 3$, $s \leq k \wedge n$, $n \geq n_0$ and for any $\varepsilon \in [0,1]$, we have*

$$\inf_{\bar{\boldsymbol{\theta}}_n} R^{\mathsf{HC}}_{\mathsf{TV}}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{\theta}}_n) \geq c\{(s/n)^{1/2} + \varepsilon\},$$
$$\inf_{\bar{\boldsymbol{\theta}}_n} R^{\mathsf{HC}}_{\mathsf{H}}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{\theta}}_n) \geq c\{(s/n)^{1/2} + \varepsilon^{1/2}\},$$
$$\inf_{\bar{\boldsymbol{\theta}}_n} R^{\mathsf{HC}}_{\mathbb{L}^2}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{\theta}}_n) \geq c\{(1/n)^{1/2} + \varepsilon\},$$

*where $\inf_{\bar{\boldsymbol{\theta}}_n}$ stands for the infimum over all measurable functions $\bar{\boldsymbol{\theta}}_n$ from $(\Delta^{k-1})^n$ to $\Delta^{k-1}$.*

The main consequence of this theorem is that whatever the contamination model is (among those described in Section 1.5.1), the rates obtained for the MLE in Theorem 1 are minimax optimal. Indeed, Theorem 2 yields this claim for Huber's contamination, and in Chapter 2 we will show that the lower bounds obtained for HC remain valid for all the other contamination models and are minimax optimal.

**Confidence regions** In previous parts, we established bounds for the expected value of estimation error. The aim of this part is to present bounds on estimation error of the sample mean holding with high probability. This also leads to constructing confidence regions for the parameter vector $\boldsymbol{\theta}^*$. To this end, the contamination rate $\varepsilon$ and the sparsity $s$ are assumed to be known. It is an interesting open question whether one can construct optimally shrinking confidence regions for unknown $\varepsilon$ and $s$.

**Theorem 3.** *Let $\delta \in (0,1)$ be the tolerance level. If $\boldsymbol{\theta}^* \in \Delta_s^{k-1}$, then under any contamination model, the regions of $\Delta^{k-1}$ defined by each of the following inequalities*

$$d_{\mathbb{L}^2}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}) \leq (1/n)^{1/2} + \sqrt{2}\,\varepsilon + \big(\log(1/\delta)/n\big)^{1/2},$$
$$d_{\mathsf{TV}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}) \leq (s/n)^{1/2} + 2\,\varepsilon + \big(2\log(1/\delta)/n\big)^{1/2},$$
$$d_{\mathsf{H}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}) \leq \sqrt{5}\big((s/n)\log(2s/\delta)\big)^{1/2} + \varepsilon^{1/2} + \big((1/2n)\log(2/\delta)\big)^{1/2},$$

*contain $\boldsymbol{\theta}^*$ with probability at least $1 - \delta$.*

To illustrate the shapes of these confidence regions, we depicted them in Figure 1.5 for a three dimensional example, projected onto the plane containing the probability simplex. The sample mean in this example is equal to $(1/3, 1/2, 1/6)$.

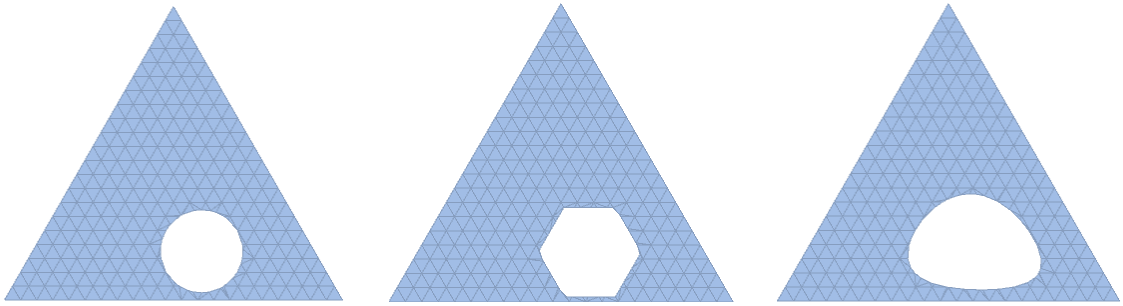

Figure 1.5: The shape of confidence sets (white regions) for the distances $\mathbb{L}^2$ (left), TV (center), and Hellinger (right) when the sample mean is $(1/3, 1/2, 1/6)$.

**Instance based bounds** When the dimension $k$ is not finite, we can provide bounds which depend on the reference distribution $\boldsymbol{\theta}^*$ or the sample mean $\bar{\boldsymbol{X}}_n$. We restrict $\boldsymbol{X}_i$'s to take value in $\{\boldsymbol{e}_1, \boldsymbol{e}_2 \dots\}$, and assume $\boldsymbol{e}_j$ occurs with probability $\boldsymbol{\theta}^*_j$. We define $\alpha_n(\boldsymbol{\theta}) := 2\sum_{\boldsymbol{\theta}_j < 1/n} \boldsymbol{\theta}_j$ and $\beta_n(\boldsymbol{\theta}) := \frac{1}{\sqrt{n}} \sum_{\boldsymbol{\theta} \geq 1/n} \sqrt{\boldsymbol{\theta}_j}$. Using the results of Berend and Kontorovich (2013) the following upper and lower bounds are obtained for the error of the sample mean under the TV distance and adversarial model with $\varepsilon$-contamination.

**Proposition 2.** *Suppose $\boldsymbol{X}_i$'s take value in $\{\boldsymbol{e}_1, \boldsymbol{e}_2 \dots\}$, and for $j \in \mathbb{N}$, $\boldsymbol{e}_j$ occurs with probability $\boldsymbol{\theta}^*_j$. For every $n$ and for every $\varepsilon \in [0, 1]$, the sample mean $\bar{\boldsymbol{X}}_n$ satisfies*

$$\frac{\alpha_n(\boldsymbol{\theta}^*_j) + \beta_n(\boldsymbol{\theta}^*_j)}{4} - \frac{1}{4\sqrt{n}} - 2\varepsilon \leq \mathbf{E} d_{\mathsf{TV}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*) \leq \alpha_n(\boldsymbol{\theta}^*_j) + \beta_n(\boldsymbol{\theta}^*_j) + 2\varepsilon.$$

These bounds need the knowledge of the reference distribution. The next theorem represent bounds based on the sample mean.

**Theorem 4.** *Suppose $\boldsymbol{X}_i$'s take value in $\{\boldsymbol{e}_1, \boldsymbol{e}_2 \dots\}$, and for $j \in \mathbb{N}$, $\boldsymbol{e}_j$ occurs with probability $\boldsymbol{\theta}^*_j$. For every $n$ and for every $\varepsilon \in [0, 1]$, the sample mean $\bar{\boldsymbol{X}}_n$ satisfies*

$$d_{\mathsf{TV}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*) \leq \frac{1}{\sqrt{n}} \big\| \bar{\boldsymbol{X}}_n^{1/2} \big\|_1 + 2\varepsilon + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

*with probability at least $1 - \delta$, where $\delta \in (0, 1)$. We also have*

$$\mathbf{E} d_{\mathsf{TV}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*) \leq \frac{1}{\sqrt{n}} \mathbf{E} \big\| \bar{\boldsymbol{X}}_n^{1/2} \big\|_1 + 2\varepsilon.$$

### 1.5.2 Robust Estimation of Gaussian Mean

The goal of this work is to make a step forward by designing an estimator which is not only nearly rate optimal and computationally tractable, but also has a (asymptotic) breakdown point equal to $1/2$, which is the highest possible value of the breakdown point. To construct the estimator, termed iterative spectral dimension reduction or SDR, we combine and suitably adapt

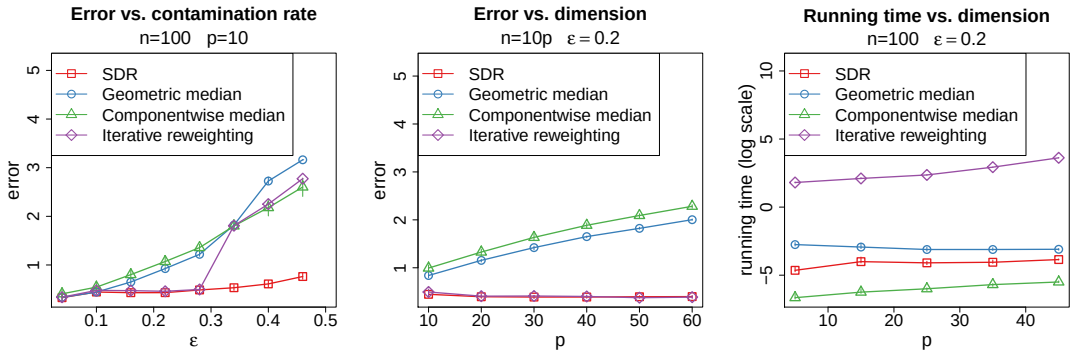


Figure 1.6: Plots that help to visually compare four robust estimators: SDR (our estimator), geometric median (GM) given by (3.1), componentwise median (CM), iteratively reweighted mean (IRM) of (Dalalyan and Minasyan, 2020). The first two plots show that SDR is as accurate as IRM for small $\varepsilon$, with SDR outperforming IRM for $\varepsilon$ close to $1/2$. IRM and SDR are naturally much more accurate than GM and CM. The last plot shows that the running time of SDR is comparable to that of GM and is much smaller than that of IRM. More details on these experiments are provided in Section 3.5.

ideas from (Lai et al., 2016) and (Diakonikolas et al., 2017). The main underlying observation is that if we remove some clear outliers and restrict our attention to the subspace spanned by the eigenvectors of the sample covariance matrix corresponding to small eigenvalues, then the sample mean of the projected data points is a rate-optimal estimator. This allows us to iteratively reduce the dimension and eventually to estimate the remaining low-dimensional component of the mean by a standard robust estimator such as the componentwise median or the trimmed mean, see Algorithm 1.

The papers that are the closest to the present work are (Lai et al., 2016), (Diakonikolas et al., 2017) and (Dalalyan and Minasyan, 2020). The spectral dimension reduction scheme was proposed by (Lai et al., 2016) along with an initial sample splitting step ensuring the independence of the estimators over different subspaces. In the case of spherical Gaussian distribution contaminated by non-adversarial outliers, the paper states that the proposed estimator has a squared error at most of order $p \log^2 p \log(p/\varepsilon)/n + \varepsilon^2 \log p$. Compared to this, our results are valid in the more general setting of sub-Gaussian distribution, with arbitrary covariance matrix and adversarial contamination. In addition, our estimator does not rely on sample splitting and, therefore, has a risk with a better dependence on $p$. As compared to the filtering method of (Diakonikolas et al., 2017), our estimator has the advantage of being independent of $\varepsilon$ and our error bound is valid for every covariance matrix and every confidence level. On the down side, our error bound has an extra factor $\log p$ in front of $\varepsilon^2$. We believe that this factor is an artifact of the proof, but we were unable to remove it. Finally, compared to the iteratively reweighted mean (Dalalyan and Minasyan, 2020), the SDR estimator studied in the present work has a higher breakdown point, does not require the knowledge of $\varepsilon$ and is much faster to compute. The advantages and shortcomings of these estimators are summarized in Table 1.1 and Figure 1.6.

*Notation.* For any pair of integers $k$ and $d$ such that $1 \le k \le d$, we denote by $\mathscr{V}_k^d$ the set of all $k$-dimensional linear subspaces $V$ of $\mathbb{R}^d$. For $V \in \mathscr{V}_k^d$, we write $k = \dim(V)$ and

| | Comput. tractable | Breakdown point | known $\varepsilon$ | Squared error rate | known $\mathbf{\Sigma}$ or $\mathbf{\Sigma} \propto \mathbf{I}$ |
|---|---|---|---|---|---|
| | Gaussian distribution | | | | |
| C./G. Median | yes | 0.5 | no | $\mathbf{r}_{\mathbf{\Sigma}}/n + \varepsilon^2 p$ | no |
| Tukey's Median | no | 0.33 | no | $\mathbf{r}_{\mathbf{\Sigma}}/n + \varepsilon^2$ | no |
| Agnostic Mean | yes | – | yes | $(p/n)\log^3 p + \varepsilon^2 \log p$ | yes |
| | Gaussian and sub-Gaussian distribution | | | | |
| Iter. Rew. Mean | yes | 0.28 | yes | $(\mathbf{r}_{\mathbf{\Sigma}}/n) + \varepsilon^2 \log(1/\varepsilon)$ | yes |
| Iterative Filtering | yes | – | yes | $(p/n)\log^a p + \varepsilon^2 \log(1/\varepsilon)$ | yes |
| SDR (this work) | yes | 0.5 | no | $(\mathbf{r}_{\mathbf{\Sigma}}/n + \varepsilon^2 \log(1/\varepsilon))\log p$ | yes |

Table 1.1: Properties of various robust estimators. Agnostic mean, iteratively reweighted mean and iterative filtering are the estimators studied in (Lai et al., 2016), (Dalalyan and Minasyan, 2020) and (Diakonikolas et al., 2017), respectively. The error rates reported for Tukey's median, componentwise median, geometric median and the agnostic mean have been proved for non-adversarial contamination. The squared error rate is provided in the case of a covariance matrix satisfying $\|\mathbf{\Sigma}\|_{\mathrm{op}} = 1$.

denote by $\mathrm{P}_V$ the orthogonal projection matrix onto $V$. $\mathbb{S}^{d-1}$ stands for the unit sphere in $\mathbb{R}^d$. For a $d \times d$ symmetric matrix $\mathbf{M}$, we denote by $\lambda_1(\mathbf{M}), \ldots, \lambda_d(\mathbf{M})$ its eigenvalues sorted in increasing order, and use the notation $\lambda_{\min}(\mathbf{M}) = \lambda_1(\mathbf{M})$, $\lambda_{\max}(\mathbf{M}) = \lambda_d(\mathbf{M})$, $\|\mathbf{M}\|_{\mathrm{op}} = \max(|\lambda_{\min}(\mathbf{M})|, |\lambda_{\max}(\mathbf{M})|)$, $\mathrm{Tr}(\mathbf{M}) = (\lambda_1 + \ldots + \lambda_d)(\mathbf{M})$ and $\mathbf{r}_{\mathbf{M}} = \mathrm{Tr}(\mathbf{M})/\|\mathbf{M}\|_{\mathrm{op}}$. For any integer $n > 0$, we set $[n] = \{1, \ldots, n\}$. We will denote by $\mathcal{O} \subset [n]$ the subscripts of the outliers and by $\mathcal{I} = [n] \setminus \mathcal{O}$ the subscripts of inliers. We also use notation $\log_+(x) = \max\{0, \log(x)\}$.

**Algorithm 1** SDR$(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; \mathbf{\Sigma}, t)$

**let** $p$ be the dimension of $\boldsymbol{X}_1$
**let** $\widehat{\boldsymbol{\mu}}^{\mathrm{GM}}$ be the geometric median of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$
**let** $\mathcal{S} \leftarrow \{i : \|\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}^{\mathrm{GM}}\| \le t\sqrt{p}\}$
**let** $\overline{\boldsymbol{X}}_{\mathcal{S}}$ be the sample mean of the filtered sample $\{\boldsymbol{X}_i : i \in \mathcal{S}\}$
**let** $\widehat{\mathbf{\Sigma}}_{\mathcal{S}}$ be the covariance matrix of the filtered sample $\{\boldsymbol{X}_i : i \in \mathcal{S}\}$
**if** $p > 1$ **then**
  **let** $V$ be the span of the top $\lceil p/e \rceil$ principal components of $\widehat{\mathbf{\Sigma}}_{\mathcal{S}} - \mathbf{\Sigma}$
  **let** $\mathrm{P}_V$ be the orth. projection onto $V$
  **let** $\mathrm{P}_{V^\perp}$ be the orth. projection onto the orth. complement of $V$
  **let** $\widehat{\boldsymbol{\mu}} \leftarrow \mathrm{P}_{V^\perp}\overline{\boldsymbol{X}}_{\mathcal{S}} + \mathrm{SDR}(\mathrm{P}_V\boldsymbol{X}_1, \ldots, \mathrm{P}_V\boldsymbol{X}_n; \mathrm{P}_V\mathbf{\Sigma}\,\mathrm{P}_V, t)$
**else**
  **let** $\widehat{\boldsymbol{\mu}} \leftarrow \widehat{\boldsymbol{\mu}}^{\mathrm{GM}}$
**end if**
**return** $\widehat{\boldsymbol{\mu}}$

**Adversarially corrupted sub-Gaussian model and spectral dimension reduction**

We assume that a set $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of $n$ data points drawn from a distribution $\mathsf{P}_n$ is given. This set is assumed to contain at least $n - [n\varepsilon]$ inliers, the remaining points being outliers. All the points lie in the $p$-dimensional Euclidean space and the inliers are independently drawn from a reference distribution, assumed to be sub-Gaussian with mean $\boldsymbol{\mu}^* \in \mathbb{R}^p$ and covariance matrix $\boldsymbol{\Sigma}$. To state the assumptions imposed on the observations in a more precise way, let us recall that the random vector $\boldsymbol{\zeta}$ is said to be sub-Gaussian with zero mean and identity covariance matrix, if $\mathbb{E}[\boldsymbol{\zeta}] = 0$, $\mathbb{E}[\boldsymbol{\zeta}\boldsymbol{\zeta}^\top] = \mathbf{I}_p$ and for some $\mathfrak{s} > 0$, we have

$$\mathbb{E}\big[e^{\boldsymbol{v}^\top \boldsymbol{\zeta}}\big] \leq \exp\big\{\mathfrak{s}\|\boldsymbol{v}\|^2/2\big\}, \quad \forall \boldsymbol{v} \in \mathbb{R}^p.$$

The parameter $\mathfrak{s}$ is commonly called the variance proxy and the writing $\boldsymbol{\zeta} \sim \mathsf{SG}_p(\mathfrak{s})$ is used.

**Definition 1.** *We say that the data generating distribution $\mathsf{P}_n$ is an adversarially corrupted sub-Gaussian distribution with mean $\boldsymbol{\mu}^*$, covariance matrix $\boldsymbol{\Sigma}$, variance proxy $\mathfrak{s}$ and contamination rate $\varepsilon$, if there is a probability space on which we can define a sequence of random vectors $(\boldsymbol{X}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{Y}_n)$ such that*

1. *$\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are independent and $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}^*) \sim \mathsf{SG}_p(\mathfrak{s})$ for every $i \in [n]$.*
2. *the cardinality of $\mathcal{O} = \{i \in [n] : \boldsymbol{Y}_i \neq \boldsymbol{X}_i\}$ is at most equal to $n\varepsilon$.*
3. *the distribution of $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is $\mathsf{P}_n$.*

*We write then*[6] $\mathsf{P}_n \in \mathsf{SGAC}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \mathfrak{s}, \varepsilon)$. *In the particular case where all $\boldsymbol{Y}_i$ are Gaussian, we will write $\mathsf{P}_n \in \mathsf{GAC}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon)$.*

The estimator we analyze in this work is termed iterative spectral dimension reduction and denoted by $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}$. It is closely related to the agnostic mean (Lai et al., 2016) and to iterative filtering (Diakonikolas et al., 2017) estimators. We will prove that SDR enjoys most of desired properties in the setting of robust estimation of the sub-Gaussian mean.

The parameters given as input to the iterative spectral dimension reduction algorithm are a strictly decreasing sequence of positive integers $p_0, \ldots, p_L$ such that $p_0 = p$ and a positive threshold $t > 0$. We recall that the geometric median is defined by

$$\widehat{\boldsymbol{\mu}}^{\mathsf{GM}} \in \mathsf{arg}\min_{\boldsymbol{\mu} \in \mathbb{R}^p} \sum_{i=1}^n \|\boldsymbol{X}_i - \boldsymbol{\mu}\|_2.$$

The algorithm for computing the SDR estimator reads as follows.

1. Start by setting $\mathbf{V}_0 = \mathbf{I}_p$.
2. For $\ell = 0, \ldots L-1$ do
   (a) Define $\bar{\boldsymbol{\mu}}^{(\ell)} \in \mathbb{R}^{p_\ell}$ as the geometric median of $\{\mathbf{V}_\ell^\top \boldsymbol{X}_i : i \in [n]\}$.

[6] SGAC stands for sub-Gaussian with adversarial contamination.

(b) Define the set $\mathcal{S}^{(\ell)} = \big\{i \in [n] : \|\mathbf{V}_\ell^\top \boldsymbol{X}_i - \bar{\boldsymbol{\mu}}^{(\ell)}\|_2 \le t\sqrt{p_\ell}\big\}$ of filtered data points.

(c) Let $\overline{\boldsymbol{X}}^{(\ell)}$ and $\widehat{\boldsymbol{\Sigma}}^{(\ell)}$ be the mean vector and the covariance matrix of the filtered sample $\{\boldsymbol{X}_i : i \in \mathcal{S}^{(\ell)}\}$, that is

$$\overline{\boldsymbol{X}}^{(\ell)} = \frac{1}{|\mathcal{S}^{(\ell)}|}\sum_{i\in\mathcal{S}^{(\ell)}} \boldsymbol{X}_i, \qquad \widehat{\boldsymbol{\Sigma}}^{(\ell)} = \frac{1}{|\mathcal{S}^{(\ell)}|}\sum_{i\in\mathcal{S}^{(\ell)}} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})^{\otimes 2}.$$

(d) Set $\widehat{\boldsymbol{\mu}}^{(\ell)} = \mathbf{V}_\ell\mathbf{U}_\ell^\top\mathbf{U}_\ell\mathbf{V}_\ell^\top\overline{\boldsymbol{X}}^{(\ell)}$, where $\mathbf{U}_\ell$ is a $(p_\ell - p_{\ell+1}) \times p_\ell$ orthogonal matrix the rows of which are the eigenvectors of $\mathbf{V}_\ell^\top(\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma})\mathbf{V}_\ell$ corresponding to its $(p_\ell - p_{\ell+1})$ smallest eigenvalues.

(e) Set $\mathbf{V}_{\ell+1} = \mathbf{V}_\ell(\mathbf{U}_\ell^\perp)^\top \in \mathbb{R}^{p\times p_{\ell+1}}$, where $\mathbf{U}_\ell^\perp$ is a $p_{\ell+1}\times p_\ell$ orthogonal matrix orthogonal to $\mathbf{U}_\ell$, that is $\mathbf{U}_\ell^\perp\mathbf{U}_\ell^\top = \mathbf{0}$.

3. Define $\bar{\boldsymbol{\mu}}^{(L)}$ as the geometric median of $\mathbf{V}_L^\top\boldsymbol{X}_i$ for $i = 1,\dots,n$ and set $\mathcal{S}^{(L)} = \big\{i \in [n] : \|\mathbf{V}_L^\top\boldsymbol{X}_i - \bar{\boldsymbol{\mu}}^{(L)}\|_2 \le t\sqrt{p_L}\big\}$.

4. Define $\widehat{\boldsymbol{\mu}}^{(L)} = \mathbf{V}_L\mathbf{V}_L^\top\,\overline{\boldsymbol{X}}^{(L)}$, the average of filtered and projected vectors.

5. Return $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}} = \widehat{\boldsymbol{\mu}}^{(0)} + \widehat{\boldsymbol{\mu}}^{(1)} + \ldots + \widehat{\boldsymbol{\mu}}^{(L)}$.

The steps described above can be summarized as follows. At each iteration $\ell < L$, we start by determining a filtered subsample $\mathcal{S}^{(\ell)}$ and a "nearly-outlier-orthogonal" subspace $\mathscr{U}_\ell = \mathrm{Im}(\mathbf{V}_\ell\mathbf{U}_\ell^\top)$ of $\mathbb{R}^p$ of dimension $p_\ell - p_{\ell+1}$. We define the projection of $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}$ onto $\mathscr{U}_\ell$ as the sample mean of the filtered and projected subsample, and we move to the next step for determining the projection of $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}$ onto the remaining part of the space. At the last iteration $L$, when the dimension is well reduced, the projection of $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}$ onto the subspace $\mathscr{U}_L$ is defined as the average of the filtered subsample projected onto $\mathscr{U}_L$. The subspaces $\mathscr{U}_\ell$ are two-by-two orthogonal and span the whole space $\mathbb{R}^p$. Each subspace is determined from the spectral decomposition of the covariance matrix of the data points projected onto $(\mathscr{U}_0 \oplus \ldots \oplus \mathscr{U}_{\ell-1})^\perp$, after removing the points lying at an abnormally large distance from the geometric median.

**Choice of the dimension reduction regime** The analysis of the error of the SDR estimator conducted in this work leads to an upper bound in which the sequence $(p_0,\dots,p_L)$ is involved only through the expression

$$F(p_0,\dots,p_L) = \sum_{\ell=1}^{L}\frac{p_{\ell-1}}{p_\ell}.$$

Therefore, an appealing way of choosing this sequence is to minimize the function $F$ under the constraint that the sequence is decreasing and $p_0 = p$ and $p_L = 1$. It follows from the inequality between the arithmetic and geometric means that $F(p_0,\dots,p_L) \ge Lp^{1/L}$. Furthermore, the

equality is achieved[7] in the case when all the terms in the definition of $F$ are equal, *i.e.*, when for some $c > 0$ we have $p_{\ell-1} = cp_\ell$ for every $\ell \in [L]$. Since $p_0 = p$ and $p_L = 1$, this yields $c = p^{1/L}$ or, equivalently, $L = \log p / \log c$. Using these relations, we find that the function $F$ is lower bounded by $Lc = (c/\log c)\log p$. The last step is to find the minimum of the function $c \mapsto c/\log c$ over the interval $(1, \infty)$. One easily checks that this function has a unique minimum at $c = e$. All these considerations advocate for using the dimension reduction regime defined by

$$p_0 = p, \quad p_\ell = \lfloor p_{\ell-1}/e \rfloor + 1, \ \ell \in [L], \quad p_L = 1, \tag{1.15}$$

where $\lfloor x \rfloor$ is the largest integer strictly smaller than $x$. Such a definition of $(p_\ell)$ ensures that $p_{\ell-1}/p_\ell \le e$ and that[8] $L \le 2\log p$. In the rest of the paper, we assume that the sequence $(p_\ell)$ is chosen as in (1.15).

**Choice of the threshold** The SDR procedure has one important tuning parameter: the threshold $t$ used to discard clearly outlying data points. Let us introduce the auxiliary notation

$$\bar{\mathsf{r}}_n = \frac{\sqrt{\mathsf{r}_{\mathbf{\Sigma}}} + \sqrt{2\log(2/\delta)}}{\sqrt{n}}, \qquad \text{and} \qquad \tau = \frac{1}{4} \bigwedge \frac{\bar{\mathsf{r}}_n}{\sqrt{\log_+(2/\bar{\mathsf{r}}_n)}}. \tag{1.16}$$

Note that $\bar{\mathsf{r}}_n$ is essentially the quantile, up to a universal constant factor, of order $1-\delta$ of the distribution of $\|\bar{\boldsymbol{Y}}_n - \boldsymbol{\mu}^*\|_2$ where $\boldsymbol{Y}_i$'s are independently drawn from $\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{\Sigma})$ with $\|\mathbf{\Sigma}\|_{\mathrm{op}} = 1$. Our theoretical results advocate for using the value $t = t_1 + t_2$, where

$$t_1 = \frac{2(1+\bar{\mathsf{r}}_n)}{1-2\varepsilon^*}, \qquad t_2 = 1 + \frac{\bar{\mathsf{r}}_n}{\sqrt{\tau}} + \sqrt{2 + \log(2/\tau)},$$

where $\varepsilon^* < 1/2$ is the largest value of the contamination rate that the algorithm may handle.

Let $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n$ be independent Gaussian with zero mean and covariance $\mathbf{\Sigma}$. The expression of $t_1$ is obtained as an upper bound on the quantile of order $1-\delta/2$ of the distribution of the random variable

$$T_1 = \sup_V \frac{2}{n(1-2\varepsilon)\dim(V)} \sum_{i=1}^n \|\mathrm{P}_V \boldsymbol{\xi}_i\|_2,$$

see Lemma 2 and its proof for further details. Similarly, $t_2$ is defined so that the event

$$\sup_V \sum_{i=1}^n \mathbb{1}\big(\|\mathrm{P}_V \boldsymbol{\xi}_i\|_2^2 > t_2^2 \dim(V)\big) \le n\tau$$

has a probability at lest $1-\delta/2$. Although we tried to get sharp values for these thresholds $t_1$

[7] We relax here the assumption that all the entries $p_\ell$ are integers.

[8] To check this inequality, one can use the fact that $3 \le p_{L-2} \le pe^{2-L} + e/(e-1)$. This implies $L \le 2\log p$ for $p \ge 6$. For smaller values of $p$, the inequality can be checked by direct computations.

and $t_2$, it is certainly possible to improve these values either by better mathematical arguments or by empirical considerations. Of course, smaller values of the thresholds $t_1$ and $t_2$ satisfying aforementioned conditions lead to an SDR estimator having smaller error.

### Assessing the error of the SDR estimator

The iterative spectral dimension reduction estimator defined in previous sections has some desirable properties of a robust estimator that are easy to check. In particular, it is clearly equivariant by translation, orthogonal linear transform and global scaling. Furthermore, the breakdown point of the estimator is equal to that of the geometric median, that is to $1/2$. This means that even if almost the half of data points are chosen to be infinitely large, the estimator will not "break down" in the sense of becoming infinitely large. However, the fact that the estimated value does not become infinitely large, it might be not very close to the true mean. The next theorem shows that this is not the case and that the error of the SDR estimator has a nearly rate-optimal behavior even when the contamination rate is close to $1/2$. The adverb "nearly" is used here to reflect the presence of the $\sqrt{\log p}$ factor in the error bound, which is not present in the minimax rate.

**Theorem 5.** *Let $\varepsilon^* \in (0, 1/2)$, and $\delta \in (0, 1/2)$. Define $\bar{\mathsf{r}}_n$ and $\tau$ as in (1.16). For every $\varepsilon \le \varepsilon^*$, let $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}$ be the estimator returned by Algorithm 1.5.2 with*

$$t = \frac{3 - 2\varepsilon^*}{1 - 2\varepsilon^*}\left(1 + \frac{\bar{\mathsf{r}}_n}{\sqrt{\tau}}\right) + \sqrt{2 + 2\log\left(1/\tau\right)}.$$

*There exists a universal constant $\mathsf{C}$ such that for every $\mathsf{P}_n \in \mathsf{GAC}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon)$ with $\varepsilon \le \varepsilon^*$ and[9] $\|\boldsymbol{\Sigma}\|_{\mathrm{op}} = 1$, the probability of the event*

$$\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}} - \boldsymbol{\mu}^*\right\|_2 \le \frac{\mathsf{C}\sqrt{\log p}}{1 - 2\varepsilon^*}\left(\sqrt{\frac{\mathbf{r}_{\boldsymbol{\Sigma}}}{n}} + \varepsilon\sqrt{\log(2/\varepsilon)} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$$

*is at least $1 - \delta$. Moreover, the constant $\mathsf{C}$ from the last display can be made explicit by replacing the effective rank $\mathbf{r}_{\boldsymbol{\Sigma}}$ by the dimension $p$ in the definition of $\bar{\mathsf{r}}_n$: That is, for every $\delta \in (0, 1/5)$ the inequality*

$$\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}} - \boldsymbol{\mu}^*\right\|_2 \le \frac{156\sqrt{2\log p}}{1 - 2\varepsilon^*}\left(\sqrt{\frac{2p}{n}} + \varepsilon\sqrt{\log(2/\varepsilon)} + \sqrt{\frac{3\log(2/\delta)}{n}}\right)$$

*holds with probability at least $1 - 5\delta$.*

If we compare this result with its counterpart established in (Dalalyan and Minasyan, 2020) for the iteratively reweighted mean, besides the extra $\log p$ factor, we see that the above error bound does not reduce to the error of the empirical mean when the contamination rate goes

[9] Since in this theorem $\boldsymbol{\Sigma}$ is assumed to be known, we can always divide all the data points $\boldsymbol{X}_i$ by $\|\boldsymbol{\Sigma}\|_{\mathrm{op}}^{1/2}$ to get a data set with a covariance matrix satisfying $\|\boldsymbol{\Sigma}\|_{\mathrm{op}} = 1$.

to zero. We do not know whether this is just a drawback of our proof, or it is an intrinsic property of the estimator. Our numerical experiments reported later on suggest that is might be a property of the estimator.

There is another logarithmic factor, $\sqrt{\log(2/\varepsilon)}$, present in the second term of the error bounds provided by the last theorem, which does not appear in the minimax rate. There are computationally intractable robust estimators of the Gaussian mean, such as the Tukey median, that have an error bound free of this factor. However, all the known error bounds provably valid for polynomial time algorithms has this extra $\sqrt{\log(2/\varepsilon)}$ factor. Furthermore, this factor is known to be unavoidable in the case of sub-Gaussian model with adversarial contamination[10], see (Lugosi and Mendelson, 2021, Section 2).

### The case of unknown covariance matrix

The SDR estimator, as defined in Algorithm 1, requires the knowledge of covariance matrix $\boldsymbol{\Sigma}$. In this section we consider the case where the matrix $\boldsymbol{\Sigma}$ is unknown, but an approximation of the latter is available. Namely, we assume that we have access to a matrix $\widetilde{\boldsymbol{\Sigma}}$ and to a real number $\gamma > 0$ such that $\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathsf{op}} \le \gamma \|\boldsymbol{\Sigma}\|_{\mathsf{op}}$. In such a situation, we can replace in the SDR estimator the true covariance matrix by its approximation $\widetilde{\boldsymbol{\Sigma}}$. This will necessarily require to adjust the threshold $t$ accordingly. The goal of the present section is to propose a suitable choice of $t$ and to show the impact of the approximation error $\gamma$ on the estimation accuracy.

As mentioned, the parameter $t$ used in Algorithm 1 needs to be properly tuned in order to account for the approximation error in the covariance matrix. To this end, we introduce the following auxiliary notation similar to those presented in (1.16):

$$\widetilde{\mathsf{r}}_n = \frac{\sqrt{\mathsf{C}_\gamma \mathbf{r}_{\widetilde{\boldsymbol{\Sigma}}}} + \sqrt{2\log(2/\delta)}}{\sqrt{n}} \qquad \text{and} \qquad \widetilde{\tau} = \frac{1}{4} \bigwedge \frac{\widetilde{\mathsf{r}}_n}{\sqrt{\log_+(2/\widetilde{\mathsf{r}}_n)}}, \tag{1.17}$$

where $\mathsf{C}_\gamma = (1+\gamma)/(1-\gamma)$. Compared to (1.16), the main difference here is the presence of the factor $\mathsf{C}_\gamma$ (which is equal to one if $\gamma = 0$) and the substitution of the effective rank of $\boldsymbol{\Sigma}$ by that of its approximation $\widetilde{\boldsymbol{\Sigma}}$. In the rest of this section, we assume that $\boldsymbol{\Sigma}$ is invertible.

**Theorem 6.** *Let $\varepsilon^* \in (0, 1/2)$, $\delta \in (0, 1/2)$ and define $\widetilde{\mathsf{r}}_n$ and $\tau$ as in (1.17). Assume that $\widetilde{\boldsymbol{\Sigma}}$ satisfies $\|\boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2} - \mathbf{I}_p\|_{\mathsf{op}} \le \gamma$ for some $\gamma \in (0, 1/2]$. Let $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}$ be the output of* $\mathsf{SDR}(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n; \widetilde{\boldsymbol{\Sigma}}, \widetilde{t}_\gamma)$*, see Algorithm 1, with*

$$\widetilde{t}_\gamma = \frac{\|\widetilde{\boldsymbol{\Sigma}}\|_{\mathsf{op}}}{1-\gamma} \left\{ \frac{3-2\varepsilon^*}{1-2\varepsilon^*} \left(1 + \frac{\widetilde{\mathsf{r}}_n}{\sqrt{\widetilde{\tau}}}\right) + \sqrt{2 + \log\left(2/\tau\right)} \right\}.$$

*Then, there exists a universal constant* $\mathsf{C}$ *such that for every data generating distribution* $\mathsf{P}_n \in$

[10] Both sub-Gaussianity of the reference distribution and the adversarial nature of the contamination are important for getting the extra $\sqrt{\log(2/\varepsilon)}$ factor in the minimax rate.

GAC$(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon)$ *with* $\varepsilon \leq \varepsilon^*$*, the probability of the event*

$$\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}} - \boldsymbol{\mu}^*\right\|_2 \leq \frac{\mathsf{C}\,\|\boldsymbol{\Sigma}\|_{\mathsf{op}}^{1/2}\sqrt{\log p}}{1-2\varepsilon^*}\left(\sqrt{\frac{\mathbf{r}_{\boldsymbol{\Sigma}}}{n}} + \varepsilon\sqrt{\log(2/\varepsilon)} + \sqrt{\varepsilon\gamma} + \sqrt{\frac{\log(1/\delta)}{n}}\right) \tag{1.18}$$

*is at least* $1-\delta$*.*

On the one hand, if the value of $\gamma$ is at most of order $\sqrt{(\mathbf{r}_{\boldsymbol{\Sigma}}/n)\log(1/\varepsilon)} + \varepsilon\log(1/\varepsilon)$ then Theorem 6 implies that the estimation error is of the same order as in the case of known covariance matrix $\boldsymbol{\Sigma}$ (Theorem 5). For instance, if the matrix $\boldsymbol{\Sigma}$ is assumed to be diagonal, one can defined $\widetilde{\boldsymbol{\Sigma}}$ as the diagonal matrix composed of robust estimators of the variances of univariate contaminated Gaussian samples; see, for instance, Section 2 in (Comminges et al., 2021). For recent advances on robust estimation of (non-diagonal) covariance matrices by computationally tractable algorithms we refer the reader to (Cheng et al., 2019b).

On the other hand, if the value of $\gamma$ for which the condition $\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathsf{op}} \leq \gamma\|\boldsymbol{\Sigma}\|_{\mathsf{op}}$ is known to be true is of larger order than $\sqrt{(\mathbf{r}_{\boldsymbol{\Sigma}}/n)\log(1/\varepsilon)} + \varepsilon\log(1/\varepsilon)$, then $\sqrt{\varepsilon\gamma}$ dominates the other terms appearing in the error bound (1.18). Moreover, if $\gamma$ is of constant order, then we get the error rate $\sqrt{\frac{\mathbf{r}_{\boldsymbol{\Sigma}}}{n}} + \sqrt{\varepsilon}$, which is in line with previously known bounds for computationally tractable estimators; see for example (Lai et al., 2016, Theorem 1.1), (Diakonikolas et al., 2017, Theorem 3.2), (Dalalyan and Minasyan, 2020, Theorem 4).

# Chapter 2

# Contamination models and robust estimation on the probability simplex

We consider the problem of estimating the mean of a distribution supported by the $k$-dimensional probability simplex in the setting where an $\varepsilon$ fraction of observations are subject to adversarial corruption. A simple particular example is the problem of estimating the distribution of a discrete random variable. Assuming that the discrete variable takes $k$ values, the unknown parameter $\boldsymbol{\theta}$ is a $k$-dimensional vector belonging to the probability simplex. We first describe various settings of contamination and discuss the relation between these settings. We then establish minimax rates when the quality of estimation is measured by the total-variation distance, the Hellinger distance, or the $\mathbb{L}^2$-distance between two probability measures. We also provide confidence regions for the unknown mean that shrink at the minimax rate. Our analysis reveals that the minimax rates associated to these three distances are all different, but they are all attained by the sample average. Furthermore, we show that the latter is adaptive to the possible sparsity of the unknown vector. Some numerical experiments illustrating our theoretical findings are reported.

---

## 2.1 Introduction

Assume $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are $n$ independent and identically distributed random variables taking their values in the $k$-dimensional probability simplex $\Delta^{k-1} = \{\boldsymbol{v} \in \mathbb{R}_+^k : v_1 + \ldots + v_k = 1\}$. Our goal is to estimate the unknown vector $\boldsymbol{\theta} = \mathbf{E}[\boldsymbol{X}_i]$ in the case where the observations are contaminated by outliers. An important particular case is the estimation of the distribution of a discrete random variable $\boldsymbol{X}$ taking $k$ distinct values. In this particular case, $\boldsymbol{X}_i$'s take values in $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k\}$, the set of the vectors of the canonical basis, which are also the extreme points of the simplex $\Delta^{k-1}$.

In this introduction, to convey the main messages, we limit ourselves to the Huber contamination model, although our results apply to the more general adversarial contamination. Huber's contamination model assumes that there are two probability measures $\boldsymbol{P}$, $\boldsymbol{Q}$ on $\Delta^{k-1}$ and a real $\varepsilon \in [0, 1/2)$ such that $\boldsymbol{X}_i$ is drawn from

$$\boldsymbol{P}_i = (1-\varepsilon)\boldsymbol{P} + \varepsilon \boldsymbol{Q}, \qquad \forall i \in \{1, \ldots, n\}.$$

This amounts to assuming that $(1-\varepsilon)$-fraction of observations, called inliers, are drawn from a reference distribution $\boldsymbol{P}$, whereas $\varepsilon$-fraction of observations are outliers and are drawn from another distribution $\boldsymbol{Q}$. In general, all the three parameters $\boldsymbol{P}$, $\boldsymbol{Q}$ and $\varepsilon$ are unknown. The parameter of interest is some functional (such as the mean, the standard deviation, etc.) of the reference distribution $\boldsymbol{P}$, whereas $\boldsymbol{Q}$ and $\varepsilon$ play the role of nuisance parameters.

When the unknown parameter lives on the probability simplex, there are many appealing ways of defining the risk. We focus on the following three metrics: total-variation, Hellinger and $\mathbb{L}^2$ distances[1]

$$\begin{aligned} d_{\mathsf{TV}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &:= {}^1\!/_2 \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1, \\ d_{\mathsf{H}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &:= {}^1\!/_{\sqrt{2}} \|\widehat{\boldsymbol{\theta}}^{1/2} - \boldsymbol{\theta}^{1/2}\|_2, \\ d_{\mathbb{L}^2}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &:= \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2. \end{aligned}$$

The Hellinger distance above is well defined when the estimator $\widehat{\boldsymbol{\theta}}$ is non-negative, which will be the case throughout this work. We will further assume that the dimension $k$ may be large,

[1] We write $\|\boldsymbol{u}\|_q = (\sum_{j=1}^k |u_j|^q)^{1/q}$ and $\boldsymbol{u}^q = (u_1^q, \ldots, u_k^q)$ for any $\boldsymbol{u} \in \mathbb{R}_+^k$ and $q > 0$.

but the vector $\boldsymbol{\theta}$ is $s$-sparse, for some $s \leq k$, *i.e.* $\#\{j : \theta_j \neq 0\} \leq s$. Our main interest is in constructing confidence regions and evaluating the minimax risk

$$\mathfrak{R}_{\square}(n, k, s, \varepsilon) := \inf_{\bar{\boldsymbol{\theta}}_n} \sup_{\boldsymbol{P}, \boldsymbol{Q}} \mathbf{E}[d_{\square}(\bar{\boldsymbol{\theta}}_n, \boldsymbol{\theta})], \tag{2.1}$$

where the *inf* is over all estimators $\bar{\boldsymbol{\theta}}_n$ built upon the observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{iid}}{\sim} (1-\varepsilon)\boldsymbol{P} + \varepsilon \boldsymbol{Q}$ and the *sup* is over all distributions $\boldsymbol{P}$, $\boldsymbol{Q}$ on the probability simplex such that the mean $\boldsymbol{\theta}$ of $\boldsymbol{P}$ is $s$-sparse. The subscript $\square$ of $\mathfrak{R}$ above refers to the distance used in the risk, so that $\square$ is TV, H, or $\mathbb{L}^2$.

The problem described above arises in many practical situations. One example is an election poll: each participant expresses his intention to vote for one of $k$ candidates. Thus, each $\theta_j$ is the true proportion of electors of candidate $j$. The results of the poll contain outliers, since some participants of the poll prefer to hide their true opinion. Another example related to elections, is the problem of counting votes across all constituencies. Each constituency communicates a vector of proportions to a central office, which is in charge of computing the overall proportions. However, in some constituencies (hopefully a small fraction only) the results are rigged. Hence, the set of observed vectors contains outliers.

We intend to provide non-asymptotic upper and lower bounds on the minimax risk that match up to numerical constants. In addition, we will provide confidence regions of the form $B_{\square}(\widehat{\boldsymbol{\theta}}_n, r_{n,\varepsilon,\delta}) = \{\boldsymbol{\theta} : d_{\square}(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}) \leq r_{n,\varepsilon,\delta}\}$ containing the true parameter with probability at least $1-\delta$ and such that the radius $r_{n,\varepsilon,\delta}$ goes to zero at the same rate as the corresponding minimax risk.

When there is no outlier, *i.e.*, $\varepsilon = 0$, it is well known that the sample mean

$$\bar{\boldsymbol{X}}_n := \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i$$

is minimax-rate-optimal and the rates corresponding to various distances are

$$\mathfrak{R}_{\mathbb{L}^2}(n, k, s, 0) \asymp (1/n)^{1/2} \quad \text{and} \quad \mathfrak{R}_{\square}(n, k, s, 0) \asymp (s/n)^{1/2} \quad \text{for } \square \in \{\mathsf{TV}, \mathsf{H}\}.$$

This raises several questions in the setting where data contains outliers. In particular, the following three questions will be answered in this work:

**Q1.** How do the risks $\mathfrak{R}_{\square}$ depend on $\varepsilon$? What is the largest proportion of outliers for which the minimax rate is the same as in the outlier-free case ?

**Q2.** Does the sample mean remain optimal in the contaminated setting?

**Q3.** What does happen if the unknown parameter $\boldsymbol{\theta}$ is $s$-sparse ?

The most important step for answering these questions is to show that

$$\begin{aligned}\mathfrak{R}_{\mathsf{TV}}(n,k,s,\varepsilon) &\asymp (s/n)^{1/2}+\varepsilon,\\ \mathfrak{R}_{\mathsf{H}}(n,k,s,\varepsilon) &\asymp (s/n)^{1/2}+\varepsilon^{1/2},\\ \mathfrak{R}_{\mathbb{L}^2}(n,k,s,\varepsilon) &\asymp (1/n)^{1/2}+\varepsilon.\end{aligned}$$

It is surprising to see that all the three rates are different leading to important discrepancies in the answers to the second part of question Q1 for different distances. Indeed, it turns out that the minimax rate is not deteriorated if the proportion of the outliers is smaller than $(s/n)^{1/2}$ for the TV-distance, $s/n$ for the Hellinger distance and $(1/n)^{1/2}$ for the $\mathbb{L}^2$ distance. Furthermore, we prove that the sample mean is minimax rate optimal. Thus, even when the proportion of outliers $\varepsilon$ and the sparsity $s$ are known, it is not possible to improve upon the sample mean. In addition, we show that all these claims hold true for the adversarial contamination and we provide corresponding confidence regions.

The rest of the paper is organized as follows. Section 2.2 introduces different possible ways of modeling data sets contaminated by outliers. Pointers to relevant prior work are given in Section 2.3. Main theoretical results and their numerical illustration are reported in Section 2.4 and Section 2.6, respectively. Section 2.7 contains a brief summary of the obtained results and their consequences, whereas the proofs are postponed to the appendix.

## 2.2 Various models of contamination

Different mathematical frameworks have been used in the literature to model the outliers. We present here five of them, from the most restrictive one to the most general, and describe their relationship. We present these frameworks in the general setting when the goal is to estimate the parameter $\boldsymbol{\theta}^*$ of a reference distribution $\boldsymbol{P}_{\boldsymbol{\theta}^*}$ when $\varepsilon$ proportion of the observations are outliers.

### 2.2.1 Huber's contamination

The most popular framework for studying robust estimation methods is perhaps the one of Huber's contamination. In this framework, there is a distribution $\boldsymbol{Q}$ defined on the same space as the reference distribution $\boldsymbol{P}_{\boldsymbol{\theta}^*}$ such that all the observations $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ are independent and drawn from the mixture distribution $\boldsymbol{P}_{\varepsilon,\boldsymbol{\theta}^*,\boldsymbol{Q}} := (1-\varepsilon)\boldsymbol{P}_{\boldsymbol{\theta}^*}+\varepsilon\boldsymbol{Q}$.

This corresponds to the following mechanism: one decides with probabilities $(1-\varepsilon,\varepsilon)$ whether a given observation is an inlier or an outlier. If the decision is made in favor of being inlier, the observation is drawn from $\boldsymbol{P}_{\boldsymbol{\theta}^*}$, otherwise it is drawn from $\boldsymbol{Q}$. More formally, if we denote by $\widehat{O}$ the random set of outliers, then conditionally to $\widehat{O}=O$,

$$\{\boldsymbol{X}_i : i\notin O\} \overset{\text{iid}}{\sim} \boldsymbol{P}_{\boldsymbol{\theta}^*},\ \{\boldsymbol{X}_i : i\in O\} \overset{\text{iid}}{\sim} \boldsymbol{Q},\ \{\boldsymbol{X}_i : i\in O\} \perp\!\!\!\perp \{\boldsymbol{X}_i : i\notin O\}, \tag{2.2}$$

for every $O \subset \{1,\dots,n\}$. Furthermore, for every subset $O$ of the observations, we have $\boldsymbol{P}(\widehat{O} = O) = (1-\varepsilon)^{n-|O|}\varepsilon^{|O|}$. We denote by[2] $\mathcal{M}_n^{\mathsf{HC}}(\varepsilon, \boldsymbol{\theta}^*)$ the set of joint probability distributions $\boldsymbol{P}_n$ of the random variables $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ satisfying the foregoing condition.

### 2.2.2 Huber's deterministic contamination

The set of outliers as well as the number of outliers in Huber's model of contamination are random. This makes it difficult to compare this model to the others that will be described later in this section. To cope with this, we define here another model, termed Huber's deterministic contamination. As its name indicates, this new model has the advantage of containing a deterministic number of outliers, in the same time being equivalent to Huber's contamination in a sense that will be made precise below.

We say that the distribution $\boldsymbol{P}_n$ of $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ belongs to the Huber's deterministic contamination model denoted by $\mathcal{M}_n^{\mathsf{HDC}}(\varepsilon, \boldsymbol{\theta}^*)$, if there are a set $O \subset \{1,\dots,n\}$ of cardinality at most $n\varepsilon$ and a distribution $\boldsymbol{Q}$ such that (2.2) is true. The apparent similarity of models $\mathcal{M}_n^{\mathsf{HC}}(\varepsilon, \boldsymbol{\theta}^*)$ and $\mathcal{M}_n^{\mathsf{HDC}}(\varepsilon, \boldsymbol{\theta}^*)$ can also be formalized mathematically in terms of the orders of magnitude of minimax risks. To ease notation, we let $R_d^{\square}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}})$ to be the worst-case risk of an estimator $\widehat{\boldsymbol{\theta}}$, where $\square$ is either HC or HDC. More precisely, for $\mathcal{M}_n^{\square}(\varepsilon, \Theta) := \cup_{\boldsymbol{\theta}\in\Theta}\mathcal{M}_n^{\square}(\varepsilon, \boldsymbol{\theta})$, we set[3]

$$R_d^{\square}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}) := \sup_{P_n \in \mathcal{M}_n^{\square}(\varepsilon,\Theta)} \mathbf{E}[d(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)].$$

This definition assumes that the parameter space $\Theta$ is endowed with a pseudo-metric $d : \Theta\times\Theta \to \mathbb{R}_+$. When $\Theta = \{\boldsymbol{\theta}^*\}$ is a singleton, we write $R_{d,n}^{\square}(\varepsilon, \boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}})$ instead of $R_d^{\square}(n, \varepsilon, \{\boldsymbol{\theta}^*\}, \widehat{\boldsymbol{\theta}})$.

**Proposition 3.** *Let $\widehat{\boldsymbol{\theta}}_n$ be an arbitrary estimator of $\boldsymbol{\theta}^*$. For any $\varepsilon \in (0, 1/2)$,*

$$R_d^{\mathit{HC}}(n, \varepsilon, \boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}_n) \le R_{d,n}^{\mathit{HDC}}(2\varepsilon, \boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}_n) + e^{-n\varepsilon/3} R_{d,n}^{\mathit{HDC}}(1, \boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}_n), \tag{2.3}$$

$$\sup_{\boldsymbol{P}_n \in \mathcal{M}_n^{\mathit{HC}}(\varepsilon,\boldsymbol{\theta}^*)} r\mathbf{P}\big(d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}^*) > r\big) \le R_{d,n}^{\mathit{HDC}}(2\varepsilon, \boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}_n) + re^{-n\varepsilon/3}. \tag{2.4}$$

*Proof in the appendix, page 82*

Denote by $\mathcal{D}_\Theta$ the diameter of $\Theta$, $\mathcal{D}_\Theta := \max_{\boldsymbol{\theta},\boldsymbol{\theta}'} d(\boldsymbol{\theta}, \boldsymbol{\theta}')$. Proposition 3 implies that

$$\inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathsf{HC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) \le \inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathsf{HDC}}(n, 2\varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) + e^{-n\varepsilon/3}\mathcal{D}_\Theta. \tag{2.5}$$

When $\Theta$ is bounded, the last term is typically of smaller order than the minimax risk over $\mathcal{M}_n^{\mathsf{HDC}}(2\varepsilon, \Theta)$. Therefore, the minimax rate of estimation in Huber's model is not slower than the minimax rate of estimation in Huber's deterministic contamination model. This entails that a lower bound on the minimax risk established in HC-model furnishes a lower bound in HDC-model.

[2]The superscript HC refers to the Huber's contamination

[3]The subscript $d$ refers to the distance $d$ used in the definition of the risk.

### 2.2.3 Oblivious contamination

A third model of contamination that can be of interest is the oblivious contamination. In this model, it is assumed that the set $O$ of cardinality $o$ and the joint distribution $\boldsymbol{Q}_O$ of outliers are determined in advance, possibly based on the knowledge of the reference distribution $\boldsymbol{P}_{\boldsymbol{\theta}^*}$. Then, the outliers $\{\boldsymbol{X}_i : i \in O\}$ are drawn randomly from $\boldsymbol{Q}_O$ independently of the inliers $\{\boldsymbol{X}_i : i \in O^c\}$. The set of all the joint distributions $\boldsymbol{P}_n$ of random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ generated by such a mechanism will be denoted by $\mathcal{M}_n^{\mathsf{OC}}(\varepsilon, \boldsymbol{\theta}^*)$. The model of oblivious contamination is strictly more general than that of Huber's deterministic contamination, since it does not assume that the outliers are iid. Therefore, the minimax risk over $\mathcal{M}_n^{\mathsf{OC}}(\varepsilon, \Theta)$ is larger than the minimax risk over $\mathcal{M}_n^{\mathsf{HDC}}(\varepsilon, \Theta)$:

$$\inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathsf{HDC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) \leq \inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathsf{OC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n).$$

The last inequality holds true for any set $\Theta$, any contamination level $\varepsilon \in (0, 1)$ and any sample size.

### 2.2.4 Parameter contamination

In the three models considered above, the contamination acts on the observations. One can also consider the case where the parameters of the distributions of some observations are contaminated. More precisely, for some set $O \subset \{1, \ldots, n\}$ selected in advance (but unobserved), the outliers $\{\boldsymbol{X}_i : i \in O\}$ are independent and independent of the inliers $\{\boldsymbol{X}_i : i \in O^c\}$. Furthermore, each outlier $\boldsymbol{X}_i$ is drawn from a distribution $\boldsymbol{Q}_i = \boldsymbol{P}_{\boldsymbol{\theta}_i}$ belonging to the same family as the reference distribution, but corresponding to a contaminated parameter $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}^*$. Thus, the joint distribution of the observations can be written as $(\bigotimes_{i \in O^c} \boldsymbol{P}_{\boldsymbol{\theta}^*}) \otimes (\bigotimes_{i \in O} \boldsymbol{P}_{\boldsymbol{\theta}_i})$. The set of all such distributions $\boldsymbol{P}_n$ will be denoted by $\mathcal{M}_n^{\mathsf{PC}}(\varepsilon, \boldsymbol{\theta}^*)$, where PC refers to "parameter contamination".

### 2.2.5 Adversarial contamination

The last model of contamination we describe in this work, the adversarial contamination, is the most general one. It corresponds to the following two-stage data generation mechanism. In a first stage, iid random variables $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are generated from a reference distribution $\boldsymbol{P}_{\boldsymbol{\theta}^*}$. In a second stage, an adversary having access to $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ chooses a (random) set $\widehat{O}$ of (deterministic) cardinality $s$ and arbitrarily modifies data points $\{\boldsymbol{Y}_i : i \in \widehat{O}\}$. The resulting sample, $\{\boldsymbol{X}_i : i = 1, \ldots, n\}$, is revealed to the Statistician. In this model, we have $\boldsymbol{X}_i = \boldsymbol{Y}_i$ for $i \notin \widehat{O}$. However, since $\widehat{O}$ is random and potentially dependent of $\boldsymbol{Y}_{1:n}$, it is not true that conditionally to $\widehat{O} = O$, $\{\boldsymbol{X}_i : i \in O^c\}$ are iid drawn from $\boldsymbol{P}_{\boldsymbol{\theta}^*}$ (for any deterministic set $O$ of cardinality $o$).

We denote by $\mathcal{M}_n^{\mathsf{AC}}(\varepsilon, \boldsymbol{\theta}^*)$ the set of all the joint distributions $\boldsymbol{P}_n$ of all the sequences

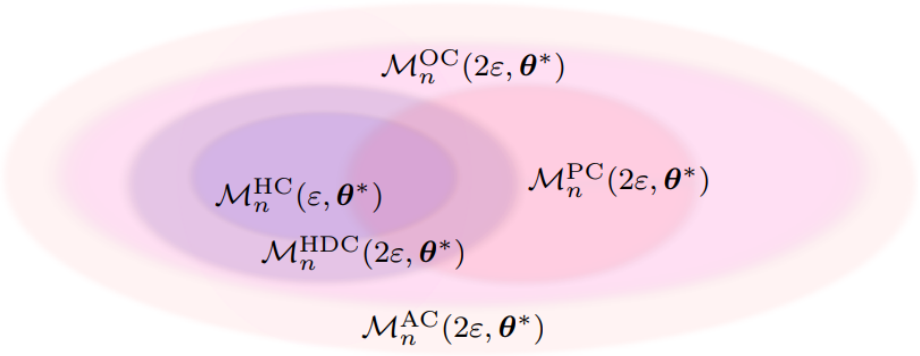


Figure 2.1: Visual representation of the hierarchy between various contamination models. Note that the inclusion of $\mathcal{M}_n^{\mathsf{HC}}(\varepsilon, \boldsymbol{\theta}^*)$ in $\mathcal{M}_n^{\mathsf{HDC}}(2\varepsilon, \boldsymbol{\theta}^*)$ is somewhat heuristic, based on the relation on the worst-case risks reported in Proposition 3.

$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ generated by the aforementioned two-stage mechanism. This set $\mathcal{M}_n^{\mathsf{AC}}(\varepsilon, \boldsymbol{\theta}^*)$ is larger than all the four sets of contamination introduced in this section. Therefore, the following inequalities hold:

$$\inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathsf{PC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) \leq \inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathrm{OC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) \leq \inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathsf{AC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n),$$

for any $n$, $\varepsilon$, $\Theta$ and any distance $d$.

### 2.2.6 Minimax risk "in expectation" versus "in deviation"

Most prior work on robust estimation focused on establishing upper bounds on the minimax risk in deviation[4], as opposed to the minimax risk in expectation defined by (2.1). One of the reasons for dealing with the deviation is that it makes the minimax risk meaningful for models[5] having random number of outliers and unbounded parameter space $\Theta$. The formal justification of this claim is provided by the following result.

**Proposition 4.** *Let $\Theta$ be a parameter space such that $\mathcal{D}_\Theta = \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta} d(\boldsymbol{\theta}, \boldsymbol{\theta}') = +\infty$. Then, for every estimator $\widehat{\theta}_n$, every $\varepsilon > 0$ and $n \in \mathbb{N}$, we have $R_d^{\mathsf{HC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) = +\infty$.*

*Proof in the appendix, page 83*

This result shows, in particular, that the last term in (2.5), involving the diameter of $\Theta$ is unavoidable. Such an explosion of the minimax risk occurs because Huber's model allows the number of outliers to be as large as $n/2$ with a strictly positive probability. One approach to overcome this shortcoming is to use the minimax risk in deviation. Another approach is to limit

[4] We call a risk bound in deviation any bound on the distance $d(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ that holds true with a probability close to one, for any parameter value $\boldsymbol{\theta}^* \in \Theta$.

[5] This is the case, for instance, of the Gaussian model with Huber's contamination.

theoretical developments to the models HDC, PC, OC or AC, in which the number of outliers is deterministic.

## 2.3 Prior work

Robust estimation is an area of active research in Statistics since at least five decades (Donoho and Gasko, 1992; Donoho and Huber, 1983; Huber, 1964; Rousseeuw and Hubert, 1999; Tukey, 1975). Until very recently, theoretical guarantees were almost exclusively formulated in terms of the notions of breakdown point, sensitivity curve, influence function, etc. These notions are well suited for accounting for gross outliers, observations that deviate significantly from the data points representative of an important fraction of data set.

More recently, various authors investigated (Chen et al., 2013; Dalalyan and Chen, 2012; Nguyen and Tran, 2013) the behavior of the risk of robust estimators as a function of the rate of contamination $\varepsilon$. A general methodology for parametric models subject to Huber's contamination was developed in Chen et al. (2016, 2018). This methodology allowed for determining the rate of convergence of the minimax risk as a function of the sample size $n$, dimension $k$ and the rate of contamination $\varepsilon$. An interesting phenomenon was discovered: in the problem of robust estimation of the Gaussian mean, classic robust estimators such as the coordinatewise median or the geometric median do not attain the optimal rate $(k/n)^{1/2} + \varepsilon$. This rate is attained by Tukey's median, the maximaizer of Tukey's halfspace depth, the computation of which is costly in a high dimensional setting. Detailed analysis of Tukey's halfspace depth was conducted in (Brunel, 2019).

In the model analyzed in this paper, we find the same minimax rate, $(k/n)^{1/2} + \varepsilon$, only when the total-variation distance is considered. A striking difference is that this rate is attained by the sample mean which is efficiently computable in any dimension. This property is to some extent similar to the problem of robust density estimation (Liu and Gao, 2019), in which the standard kernel estimators are minimax optimal in contaminated setting. Note that the fact that in the sparse setting the improvement from $(k/n)^{1/2}$ to $(s/n)^{1/2}$ can be achieved without any penalization, and that the constraint of belonging to the probability simplex acts as a sparsity favoring penalty, was already known in the literature, see (Dalalyan and Sebbar, 2018; Xia and Koltchinskii, 2016). Interestingly, similar phenomenon is observed in problems of estimation under shape constraints (Bellec, 2018; Guntuboyina and Sen, 2018). It is an interesting avenue of future research to analyze the robustness of the maximum likelihood estimator in this context.

Computational intractability of Tukey's median motivated a large number of studies that aimed at designing computationally tractable methods with nearly optimal statistical guarantees. Many of these works went beyond Huber's contamination by considering parameter contamination models (Bhatia et al., 2017; Carpentier et al., 2018; Collier and Dalalyan, 2019), oblivious contamination (Feng et al., 2014; Lai et al., 2016) or adversarial contami-

nation (Balakrishnan et al., 2017; Dalalyan and Minasyan, 2020; Diakonikolas et al., 2016a, 2017, 2018b). Interestingly, in the problem of estimating the Gaussian mean, it was proven that the minimax rates under adversarial contamination are within a factor at most logarithmic in $n$ and $k$ of the minimax rates under Huber's contamination[6]. While each of the aforementioned papers introduced clearly the conditions on the contamination, to our knowledge, none of them described different possible models and the relationship between them.

Another line of growing literature on robust estimation aims at robustifying estimators and prediction methods to heavy tailed distributions, see (Audibert and Catoni, 2011; Chinot et al., 2018; Devroye et al., 2016; Donoho and Montanari, 2016; Joly et al., 2017; Lecué and Lerasle, 2017; Lugosi and Mendelson, 2019b; Minsker, 2015, 2018a). The results of those papers are of a different nature, as compared to the present work, not only in terms of the goals, but also in terms of mathematical and algorithmic tools.

In the case of the discrete distributions, Braess and Sauer (2004) established the minimax rates under the Kullback-Leibler divergence. More recently, Kamath et al. (2015) determined the minimax rates under other distances such as $\mathbb{L}^2$, TV and the general family of $f$-divergence (including the $\chi^2$-divergence and the Hellinger distance). The estimator proposed in (Kamath et al., 2015), achieving the minimax rate for $\mathbb{L}^2$ and TV distances, is the sample mean while different estimators are proposed for the other distances. Concerning the robust estimation of discrete distributions, Chen et al. (2020); Jain and Orlitsky (2019); Qiao and Valiant (2018) studied the case of group-contamination. The distinctive feature of this setting is a better dependence of the minimax rate on the contamination rate $\varepsilon$. More precisely, if each group contains $m$ samples, and $\varepsilon$ fraction of groups are contaminated, the rates are obtained by replacing $\varepsilon$ by $\varepsilon/\sqrt{m}$. The estimators achieving these rates are much more sophisticated than the sample mean.

## 2.4 Minimax rates on the "sparse" simplex and confidence regions

We now specialize the general setting of Section 2.2 to a reference distribution $\boldsymbol{P}$, with expectation $\boldsymbol{\theta}^*$, defined on the simplex $\Delta^{k-1}$. Along with this reference model describing the distribution of inliers, we will use different models of contamination. More precisely, we will establish upper bounds on worst-case risks of the sample mean in the most general, adversarial, contamination setting. Then, matching lower bounds will be provided for minimax risks under Huber's contamination.

### 2.4.1 Upper bounds: worst-case risk of the sample mean

We denote by $\Delta_s^{k-1}$ the set of all $\boldsymbol{v} \in \Delta^{k-1}$ having at most $s$ non-zero entries.

[6] All these papers consider the risk in deviation, so that the minimax risk under Huber's contamination is finite.

**Theorem 7.** *For every triple of positive integers* $(k, s, n)$ *and for every* $\varepsilon \in [0, 1]$*, the sample mean* $\bar{\boldsymbol{X}}_n := \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$ *satisfies*

$$R^{\mathsf{AC}}_{\mathsf{TV}}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{X}}_n) \le (s/n)^{1/2} + 2\varepsilon,$$
$$R^{\mathsf{AC}}_{\mathsf{H}}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{X}}_n) \le (s/n)^{1/2} + \sqrt{2}\,\varepsilon^{1/2},$$
$$R^{\mathsf{AC}}_{\mathbb{L}^2}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{X}}_n) \le (1/n)^{1/2} + \sqrt{2}\,\varepsilon.$$

*Proof in the appendix, page 84*

An unexpected and curious phenomenon unveiled by this theorem is that all the three rates are different. As a consequence, the answer to the question "what is the largest possible number of outliers, $o_d^*(n, s)$, that does not impact the minimax rate of estimation of $\boldsymbol{\theta}^*$?" crucially depends on the considered distance $d$. Taking into account the relation $\varepsilon = o/n$, we get

$$o^*_{\mathsf{TV}}(n, s) \asymp (ns)^{1/2}, \quad o^*_{\mathsf{H}}(n, s) \asymp s, \quad o^*_{\mathbb{L}^2}(n, s) \asymp n^{1/2}.$$

Furthermore, all the claims concerning the total variation distance, in the considered model, yield corresponding claims for the Wasserstein distances $W_q$, for every $q \ge 1$. Indeed, one can see an element $\boldsymbol{\theta} \in \Delta^{k-1}$ as the probability distribution of a random vector $\boldsymbol{X}$ taking values in the finite set $\mathcal{A} = \{\boldsymbol{e}_1, \dots, \boldsymbol{e}_k\}$ of vectors of the canonical basis of $\mathbb{R}^k$. Since these vectors satisfy $\|\boldsymbol{e}_j - \boldsymbol{e}_{j'}\|_2^2 = 2\mathbb{1}(j \ne j')$, we have

$$\begin{aligned} W_q^q(\boldsymbol{\theta}, \boldsymbol{\theta}') &= \inf_{\Gamma} \mathbf{E}_{(\boldsymbol{X}, \boldsymbol{X}') \sim \Gamma}[\|\boldsymbol{X} - \boldsymbol{X}'\|_2^q] \\ &= \inf_{\Gamma} 2^{q/2} \boldsymbol{P}(\boldsymbol{X} \ne \boldsymbol{X}') = 2^{q/2} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_{\mathsf{TV}}, \end{aligned} \tag{2.6}$$

where the *inf* is over all joint distributions $\Gamma$ on $\mathcal{A} \times \mathcal{A}$ having marginal distributions $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. This implies that

$$R^{\mathsf{AC}}_{W_q}(n, \varepsilon, \Delta_s^{k-1}) \le \sqrt{2}\big\{(s/n)^{1/2} + 2\varepsilon\big\}^{1/q}, \qquad \forall q \ge 1. \tag{2.7}$$

In addition, since the $\mathbb{L}_2$ norm is an upper bound on the $\mathbb{L}_\infty$-norm, we have $R^{\mathsf{AC}}_{\mathbb{L}_\infty}(n, \varepsilon, \Delta^{k-1}) \le (1/n)^{1/2} + \sqrt{2}\,\varepsilon$. Thus, we have obtained upper bounds on the risk of the sample mean for all commonly used distances on the space of probability measures.

### 2.4.2 Lower bounds on the minimax risk

A natural question, answered in the next theorem, is how tight are the upper bounds obtained in the last theorem. More importantly, one can wonder whether there is an estimator that has a worst-case risk of smaller order than that of the sample mean.

**Theorem 8.** *There are universal constants* $c > 0$ *and* $n_0$*, such that for any integers* $k \ge 3$*,*

$s \leq k \wedge n$, $n \geq n_0$ *and for any* $\varepsilon \in [0,1]$, *we have*

$$\inf_{\bar{\boldsymbol{\theta}}_n} R^{HC}_{TV}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{\theta}}_n) \geq c\{(s/n)^{1/2} + \varepsilon\},$$
$$\inf_{\bar{\boldsymbol{\theta}}_n} R^{HC}_{H}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{\theta}}_n) \geq c\{(s/n)^{1/2} + \varepsilon^{1/2}\},$$
$$\inf_{\bar{\boldsymbol{\theta}}_n} R^{HC}_{\mathbb{L}^2}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{\theta}}_n) \geq c\{(1/n)^{1/2} + \varepsilon\},$$

*where* $\inf_{\bar{\boldsymbol{\theta}}_n}$ *stands for the infimum over all measurable functions* $\bar{\boldsymbol{\theta}}_n$ *from* $(\Delta^{k-1})^n$ *to* $\Delta^{k-1}$.

*Proof in the appendix, page 86*

The main consequence of this theorem is that whatever the contamination model is (among those described in Section 2.2), the rates obtained for the MLE in Theorem 7 are minimax optimal. Indeed, Theorem 8 yields this claim for Huber's contamination. For Huber's deterministic contamination and and the TV-distance, on the one hand, we have

$$\begin{aligned} R^{\mathsf{HDC}}_{\mathsf{TV}}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{\theta}}_n) &\geq R^{\mathsf{HDC}}_{\mathsf{TV}}(n, 0, \Delta_s^{k-1}, \bar{\boldsymbol{\theta}}_n) \\ &\overset{(1)}{=} R^{\mathsf{HC}}_{\mathsf{TV}}(n, 0, \Delta_s^{k-1}, \bar{\boldsymbol{\theta}}_n) \overset{(2)}{\geq} c(s/n)^{1/2}, \end{aligned}$$

where (1) uses the fact that for $\varepsilon = 0$ all the sets $\mathcal{M}_n^{\square}(\varepsilon, \boldsymbol{\theta}^*)$ are equal, while (2) follows from the last theorem. On the other hand, in view of Proposition 3, for $\varepsilon \geq (6/n)\log(8n/c)$ (implying that $2e^{-n\varepsilon/6} \leq (c/4)\varepsilon$),

$$\begin{aligned} R^{\mathsf{HDC}}_{\mathsf{TV}}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{\theta}}_n) &\geq R^{\mathsf{HC}}_{\mathsf{TV}}(n, \varepsilon/2, \Delta_s^{k-1}, \bar{\boldsymbol{\theta}}_n) - 2e^{-n\varepsilon/6} \\ &\geq (c/4)\big\{(s/n)^{1/2} + \varepsilon\big\}. \end{aligned}$$

Combining these two inequalities, for $n \geq (10 + 2\log(1/c))^2$, we get

$$R^{\mathsf{HDC}}_{\mathsf{TV}}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{\theta}}_n) \geq (c/4)\big\{(s/n)^{1/2} + \varepsilon\big\}$$

for every $k \geq 1$ and every $\varepsilon \in [0,1]$. The same argument can be used to show that all the inequalities in Theorem 8 are valid for Huber's deterministic contamination model as well. Since the inclusions $\mathcal{M}_n^{\mathsf{HDC}}(\varepsilon, \boldsymbol{\theta}^*) \subset \mathcal{M}_n^{\mathsf{OC}}(\varepsilon, \boldsymbol{\theta}^*) \cap \mathcal{M}_n^{\mathsf{PC}}(\varepsilon, \boldsymbol{\theta}^*) \subset \mathcal{M}_n^{\mathsf{AC}}(\varepsilon, \boldsymbol{\theta}^*)$ hold true, we conclude that the lower bounds obtained for HC remain valid for all the other contamination models and are minimax optimal.

The main tool in the proof of Theorem 8 is the following result (Chen et al., 2018, Theorem 5.1). There is a universal constant $c_1 > 0$ such that

$$\inf_{\bar{\boldsymbol{\theta}}_n} \sup_{\mathcal{M}_n^{\mathrm{HC}}(\varepsilon,\Delta)} \boldsymbol{P}\big(d(\bar{\boldsymbol{\theta}}_n, \boldsymbol{\theta}^*) \geq w_d(\varepsilon, \Delta)\big) \geq c_1, \quad \forall \varepsilon \in [0,1),$$

where $w_d(\varepsilon, \Delta)$ is the modulus of continuity defined by $w_d(\varepsilon, \Delta) = \sup\{d(\boldsymbol{\theta}, \boldsymbol{\theta}') : d_{\mathsf{TV}}(\boldsymbol{\theta}, \boldsymbol{\theta}') \leq$

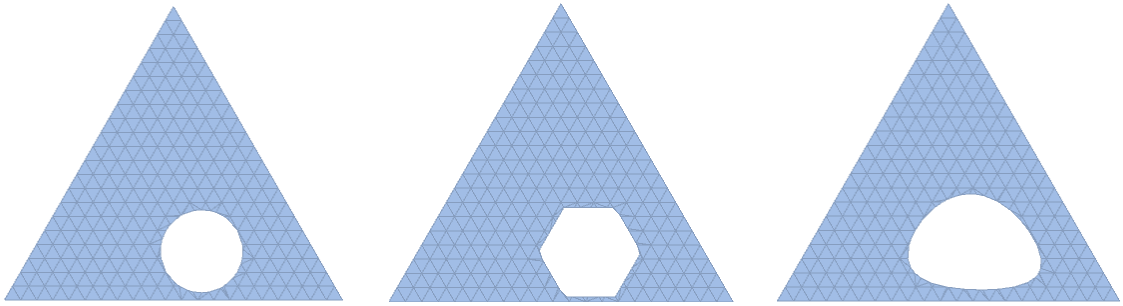

Figure 2.2: The shape of confidence sets (white regions) for the distances $\mathbb{L}^2$ (left), TV (center), and Hellinger (right) when the sample mean is $(1/3, 1/2, 1/6)$.

$\varepsilon/(1-\varepsilon)\}$. Choosing $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ to differ on only to coordinates, one can check that, for any $\varepsilon \leq 1/2$, $w_{\mathrm{TV}}(\varepsilon, \Delta_s^{k-1}) \geq \varepsilon$, $w_{\mathrm{H}}(\varepsilon, \Delta_s^{k-1}) \geq \varepsilon^{1/2}$ and $w_{\mathbb{L}^2}(\varepsilon, \Delta_s^{k-1}) \geq \sqrt{2}\varepsilon$. Combining with the lower bounds in the non-contaminated setting, this result yields the claims of Theorem 8. In addition, (2.6) combined with the results of this section implies that the rate in (2.7) is minimax optimal.

### 2.4.3 Confidence regions

In previous sections, we established bounds for the expected value of estimation error. The aim of this section is to present bounds on estimation error of the sample mean holding with high probability. This also leads to constructing confidence regions for the parameter vector $\boldsymbol{\theta}^*$. To this end, the contamination rate $\varepsilon$ and the sparsity $s$ are assumed to be known. It is an interesting open question whether one can construct optimally shrinking confidence regions for unknown $\varepsilon$ and $s$.

**Theorem 9.** *Let $\delta \in (0,1)$ be the tolerance level. If $\boldsymbol{\theta}^* \in \Delta_s^{k-1}$, then under any contamination model, the regions of $\Delta^{k-1}$ defined by each of the following inequalities*

$$
\begin{aligned}
d_{\mathbb{L}^2}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}) &\leq (1/n)^{1/2} + \sqrt{2}\,\varepsilon + \big(\log(1/\delta)/n\big)^{1/2}, \\
d_{\mathrm{TV}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}) &\leq (s/n)^{1/2} + 2\,\varepsilon + \big(2\log(1/\delta)/n\big)^{1/2}, \\
d_{\mathrm{H}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}) &\leq \sqrt{5}\big((s/n)\log(2s/\delta)\big)^{1/2} + \varepsilon^{1/2} + \big((1/2n)\log(2/\delta)\big)^{1/2},
\end{aligned}
$$

*contain $\boldsymbol{\theta}^*$ with probability at least $1-\delta$.*

*Proof in the appendix, page 88*

To illustrate the shapes of these confidence regions, we depicted them in Figure 2.2 for a three dimensional example, projected onto the plane containing the probability simplex. The sample mean in this example is equal to $(1/3, 1/2, 1/6)$.

While the estimator, $\bar{\boldsymbol{X}}_n$, used in the construction of confidence regions is adaptive both to the sparsity $s$ and to the contamination rate $\varepsilon$, the confidence region itself is not adaptive.

Indeed, the radius of the confidence region depends on $s$ and/or on $\varepsilon$. Constructing adaptive confidence regions and adaptive tests is an important question in statistics, we refer to (Cai and Low, 2006; Hoffmann and Nickl, 2011) for some precise results.

In the setting of shape constrained regression, Bellec (2016) proposed a method of constructing adaptive confidence regions that could be of interest for our setting as well. In particular, if there is no contamination, *i.e.*, $\varepsilon = 0$, we can define $\widehat{s} = \text{Card}(j : (\bar{\boldsymbol{X}}_n)_j \neq 0)$, the observed sparsity. It is clear that $\widehat{s} \leq s$. The analog of Bellec's method in our setting would consist in replacing $s$ by $\widehat{s}$ in the radius of the ball given by Theorem 9. Of course, this does not inflate the ball. However, we did not manage to adapt the argument in (Bellec, 2016, Theorem 4.1) for proving that the modified region (with $\widehat{s}$ instead of $s$) has still the required coverage property and, therefore, is adaptive to $s$. The main difficulty, at a very high-level, comes from the fact that the distance we consider, $d_{TV}$, is not induced by an inner product. Thus, constructing adaptive confidence regions even when there is no contamination is an open question.

## 2.5 Instance based bounds

When the dimension $k$ is not finite, we can provide bounds which depend on the reference distribution $\boldsymbol{\theta}^*$ or the sample mean $\bar{\boldsymbol{X}}_n$. We restrict $\boldsymbol{X}_i$'s to take value in $\{\boldsymbol{e}_1, \boldsymbol{e}_2 \dots\}$, and assume $\boldsymbol{e}_j$ occurs with probability $\boldsymbol{\theta}^*_j$. We define $\alpha_n(\boldsymbol{\theta}) := 2\sum_{\boldsymbol{\theta}_j < 1/n} \boldsymbol{\theta}_j$ and $\beta_n(\boldsymbol{\theta}) := \frac{1}{\sqrt{n}} \sum_{\boldsymbol{\theta} \geq 1/n} \sqrt{\boldsymbol{\theta}_j}$. Using the results of Berend and Kontorovich (2013) the following upper and lower bounds are obtained for the error of the sample mean under the TV distance and adversarial model with $\varepsilon$-contamination.

**Proposition 5.** *Suppose $\boldsymbol{X}_i$'s take value in $\{\boldsymbol{e}_1, \boldsymbol{e}_2 \dots\}$, and for $j \in \mathbb{N}$, $\boldsymbol{e}_j$ occurs with probability $\boldsymbol{\theta}^*_j$. For every $n$ and for every $\varepsilon \in [0,1]$, the sample mean $\bar{\boldsymbol{X}}_n$ satisfies*

$$\frac{\alpha_n(\boldsymbol{\theta}^*_j) + \beta_n(\boldsymbol{\theta}^*_j)}{4} - \frac{1}{4\sqrt{n}} - 2\varepsilon \leq \mathbf{E} d_{\mathsf{TV}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*) \leq \alpha_n(\boldsymbol{\theta}^*_j) + \beta_n(\boldsymbol{\theta}^*_j) + 2\varepsilon.$$

*Proof in the appendix, page 90*

These bounds need the knowledge of the reference distribution. The next theorem represent bounds based on the sample mean.

**Theorem 10.** *Suppose $\boldsymbol{X}_i$'s take value in $\{\boldsymbol{e}_1, \boldsymbol{e}_2 \dots\}$, and for $j \in \mathbb{N}$, $\boldsymbol{e}_j$ occurs with probability $\boldsymbol{\theta}^*_j$. For every $n$ and for every $\varepsilon \in [0,1]$, the sample mean $\bar{\boldsymbol{X}}_n$ satisfies*

$$d_{\mathsf{TV}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*) \leq \frac{1}{\sqrt{n}} \left\| \bar{\boldsymbol{X}}_n^{1/2} \right\|_1 + 2\varepsilon + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

*with probability at least* $1-\delta$*, where* $\delta \in (0,1)$*. We also have*

$$\mathbf{E} d_{\mathsf{TV}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*) \leq \frac{1}{\sqrt{n}} \mathbf{E} \left\| \bar{\boldsymbol{X}}_n^{1/2} \right\|_1 + 2\varepsilon.$$

*Proof in the appendix, page 91*

## 2.6 Illustration on a numerical example

We provide some numerical experiments which illustrate theoretical results of Section 2.4. The data set is the collection of 38 books written by Alexandre Dumas (1802-1870) and 38 books written by Emile Zola (1840-1902)[7]. To each author, we assign a parameter vector corresponding to the distribution of the number of words contained in the sentences used in the author's books. To be more clear, a sentence containing $l$ words is represented by vector $\boldsymbol{e}_l$, and if the parameter vector of an author is $(\theta_1, \ldots, \theta_k)$, it means that a sentence used by the author is of size $l \in \{1, \ldots, k\}$ with probability $\theta_l$. We carried out synthetic experiments in which the reference parameter to estimate is the probability vector of Dumas, while the distribution of outliers is determined by the probability vector of Zola. Ground truths for these parameters are computed from the aforementioned large corpus of their works. Only the dense case $s = k$ were considered. For various values of $\varepsilon$ and $n$, a contaminated sample was generated by randomly choosing $n$ sentences either from Dumas' works (with probability $1-\varepsilon$) or from Zola's works (with probability $\varepsilon$). The sample mean was computed for this corrupted sample, and the error with respect to Dumas' parameter vector was measured by the three distances TV, $\mathbb{L}^2$ and Hellinger. This experiment was repeated $10^4$ times for each special setting to obtain information on error's distribution. Furthermore, by grouping nearby outcomes we created samples of different dimensions for illustrating the behavior of the error as a function of $k$.

The error of $\bar{X}_n$ as a function of the sample size $n$, dimension $k$, and contamination rate $\varepsilon$ is plotted in Figures 2.3 and 2.4. These plots are conform to the theoretical results. Indeed, the first plot in Figure 2.3 shows that the errors for the three distances is decreasing w.r.t. $n$. Furthermore, we see that up to some level of $n$ this decay is of order $n^{-1/2}$. The second plot in Figure 2.3 confirms that the risk grows linearly in $k$ for the TV and Hellinger distances, while it is constant for the $\mathbb{L}^2$ error.

Left panel of Figure 2.4 suggests that the error grows linearly in terms of contamination rate. This is conform to our results for the TV and $\mathbb{L}^2$ errors. But it might seem that there is a disagreement with the result for the Hellinger distance, for which the risk is shown to increase at the rate $\varepsilon^{1/2}$ and not $\varepsilon$. This is explained by the fact that the rate $\varepsilon^{1/2}$ corresponds to the worst-case risk, whereas here, the setting under experiment does not necessarily represent the worst case. When the parameter vectors of the reference and contamination distributions,

[7] The works of both authors are available from https://www.gutenberg.org/

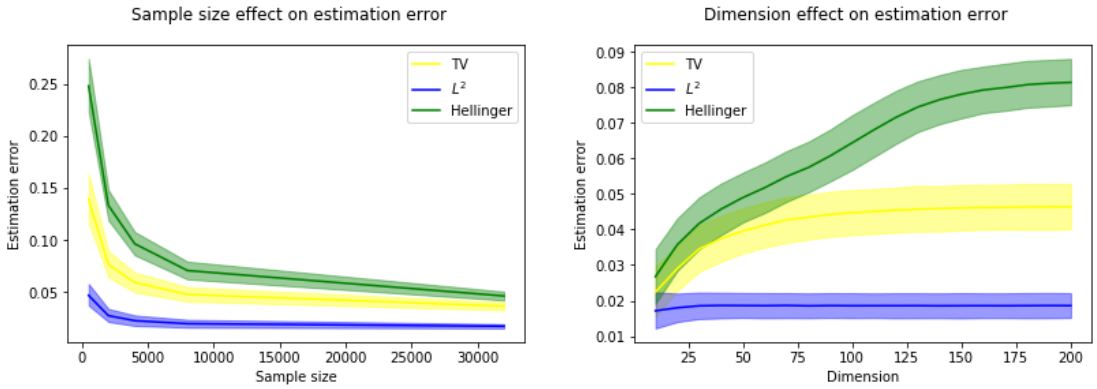


Figure 2.3: Estimation error of $\bar{\boldsymbol{X}}_n$ measured by total variation, Hellinger, and $\mathbb{L}^2$ distances as a function of (left panel) number of observations with contamination rate 0.2 and dimension $10^2$ and (right panel) dimension with contamination rate 0.2 and $10^4$ samples. The interval between 5th and 95th quantiles of the error, obtained from $10^4$ repetitions, is also depicted for every graph.

respectively, are $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ with $i \neq j$ (*i.e.*, when these two distributions are at the largest possible distance, which we call an extreme case), the graph of the error as a function of $\varepsilon$ (right panel of Figure 2.4) is similar to that of square-root function.

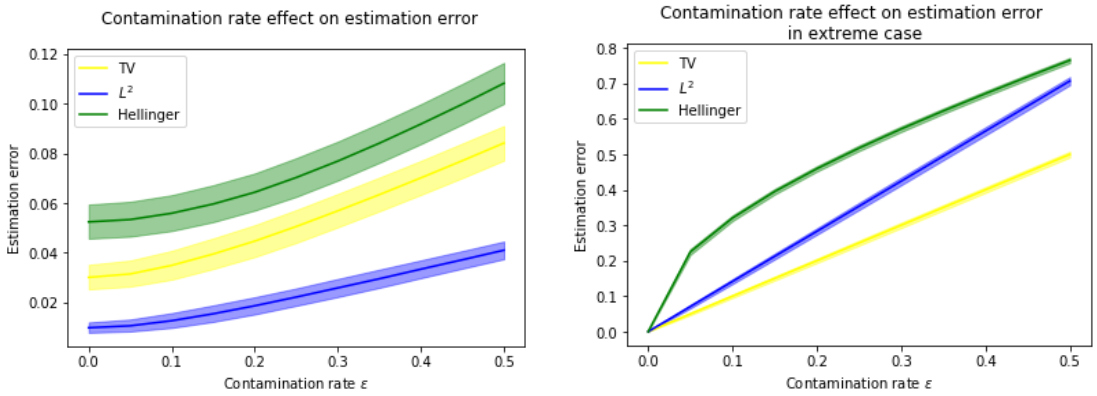


Figure 2.4: The estimation error of $\bar{\boldsymbol{X}}_n$ measured by total variation, Hellinger, and $\mathbb{L}^2$ distances in terms of the contamination rate (with dimension $10^2$ and $10^4$ samples). At right, we plotted the error with respect to the contamination rate for an extreme case, where the reference and contamination distributions have the largest distance. The interval between 5th and 95th quantiles of the error, obtained from $10^4$ trials, is also depicted.

## 2.7 Summary and conclusion

We have analyzed the problem of robust estimation of the mean of a random vector belonging to the probability simplex. Four measures of accuracy have been considered: total variation, Hellinger, Euclidean and Wasserstein distances. In each case, we have established the minimax rates of the expected error of estimation under the sparsity assumption. In addition,

confidence regions shrinking at the minimax rate have been proposed.

An intriguing observation is that the choice of the distance has much stronger impact on the rate than the nature of contamination. Indeed, while the rates for the aforementioned distances are all different, the rate corresponding to one particular distance is not sensitive to the nature of outliers (ranging from Huber's contamination to the adversarial one). While the rate obtained for the TV-distance coincides with the previously known rate of robustly estimating a Gaussian mean, the rates we have established for the Hellinger and for the Euclidean distances appear to be new. Interestingly, when the error is measured by the Euclidean distance, the quality of estimation does not get deteriorated with increasing dimension.

# Chapter 3

# Robust Estimation of Gaussian Mean

We study the problem of robust estimation of the mean vector of a sub-Gaussian distribution. We introduce an estimator based on spectral dimension reduction (SDR) and establish a finite sample upper bound on its error that is minimax-optimal up to a logarithmic factor. Furthermore, we prove that the breakdown point of the SDR estimator is equal to $1/2$, the highest possible value of the breakdown point. In addition, the SDR estimator is equivariant by similarity transforms and has low computational complexity. More precisely, in the case of $n$ vectors of dimension $p$—at most $\varepsilon n$ out of which are adversarially corrupted—the SDR estimator has a squared error of order $(\mathbf{r}_{\mathbf{\Sigma}}/n + \varepsilon^2 \log(1/\varepsilon))\log p$ and a running time of order $p^3 + np^2$. Here, $\mathbf{r}_{\mathbf{\Sigma}} \leq p$ is the effective rank of the covariance matrix of the reference distribution. Another advantage of the SDR estimator is that it does not require knowledge of the contamination rate and does not involve sample splitting. We also investigate extensions of the proposed algorithm and of the obtained results in the case of (partially) unknown covariance matrix.

---



---

## 3.1 Introduction

Robust estimation of a finite-dimensional parameter is a classical problem in statistics. The broad goal of robust estimation is to design statistical procedures that are not very sensitive to small changes in data or to small departures from the modeling assumptions. A typical example, extensively studied in the literature, and considered in the present work, is when the data set contains outliers.

The literature on robustness to outliers in parametric estimation is very rich; it would be impossible to review here all the important contributions. For an in-depth exposition of by now classical results and approaches, such as the influence function, the breakdown point and the efficiency, we refer to the books (Huber and Ronchetti, 2011; Maronna et al., 2006; Rousseeuw et al., 2011). In their vast majority, these well-established approaches treated the dimension of the parameter as a fixed and small constant. This simple setting was convenient for mathematical analysis and for computational purposes, but somewhat disconnected from many practical situations. Furthermore, it was hiding some fascinating phenomena that emerge only when the dimension is considered as a parameter that might be large, in the same way as the sample size.

More recently, (Chen et al., 2018) considered the problem of estimating the mean and the covariance matrix of a Gaussian distribution in the high-dimensional setting. The authors namely uncovered a new phenomenon: under the Huber contamination, the componentwise median is not minimax-rate optimal whereas the Tukey median is. More precisely, if a $p$-dimensional mean vector is to be estimated from $n$ independent vectors drawn from the mixture distribution $(1-\varepsilon)\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \varepsilon\mathbf{Q}$ (where $\varepsilon \in (0, 1/2)$ is the rate of contamination and $\mathbf{Q}$ is the unknown distribution of outliers), then the mean squared error of the componentwise median is of order $p/n + p\varepsilon^2$ while that of Tukey's median is of order $p/n + \varepsilon^2$. This extra factor $p$ in front of $\varepsilon^2$ has been proven in (Lai et al., 2016) to be present in the error of another widely used robust estimator of the mean, the geometric median (Minsker, 2015). Thus, as long as only statistical properties of the estimators are considered, Tukey's median is thus superior to its competitors, the componentwise and the geometric medians. However, the componentwise and the geometric medians are better than the Tukey's median in terms of the breakdown point: their breakdown point is equal to $1/2$ (Lopuhaa and Rousseeuw, 1991) whereas that of Tukey's median is $1/3$ (Donoho and Gasko, 1992). This is one of the appealing phenomena taking place in the high dimensional setting.

Another specificity of the high dimensional setting uncovered by (Chen et al., 2018) was the lack of computational tractability of the estimators that are statistically optimal. Indeed, Tukey's median is computationally intractable, but minimax-rate optimal, whereas the componentwise and the geometric medians are computationally tractable but statistically suboptimal. This observation led to the development of a number of computationally tractable estimators having an error with a better dependence on dimension than that of Tukey's median (Dalalyan and Minasyan, 2020; Diakonikolas et al., 2016a, 2017, 2018a; Dong et al., 2019; Lai et al., 2016).

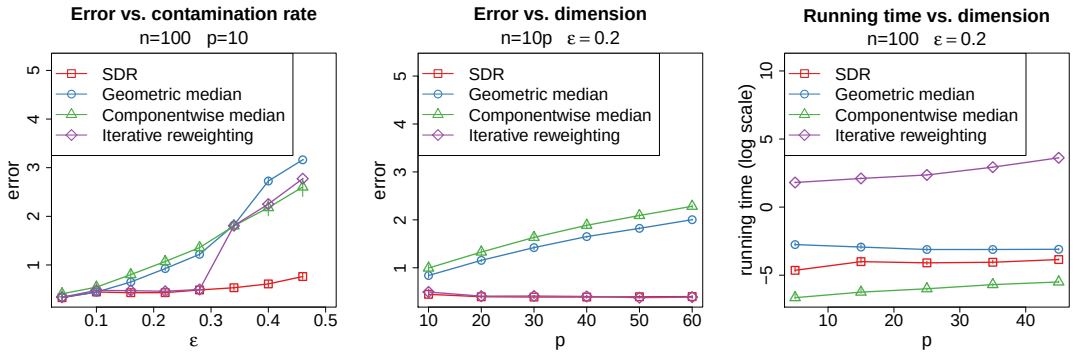


Figure 3.1: Plots that help to visually compare four robust estimators: SDR (our estimator), geometric median (GM) given by (3.1), componentwise median (CM), iteratively reweighted mean (IRM) of (Dalalyan and Minasyan, 2020). The first two plots show that SDR is as accurate as IRM for small $\varepsilon$, with SDR outperforming IRM for $\varepsilon$ close to $1/2$. IRM and SDR are naturally much more accurate than GM and CM. The last plot shows that the running time of SDR is comparable to that of GM and is much smaller than that of IRM. More details on these experiments are provided in Section 3.5.

In particular, most estimators introduced in these papers allow to conciliate computational tractability (*i.e.*, are computable in time polynomial in $n, p, 1/\varepsilon$) and statistical optimality up to logarithmic factors.

The goal of this paper is to make a step forward by designing an estimator which is not only nearly rate optimal and computationally tractable, but also has a breakdown point equal to $1/2$, which is the highest possible value of the breakdown point. To construct the estimator, termed iterative spectral dimension reduction or SDR, we combine and suitably adapt ideas from (Lai et al., 2016) and (Diakonikolas et al., 2017). The main underlying observation is that if we remove some clear outliers and restrict our attention to the subspace spanned by the eigenvectors of the sample covariance matrix corresponding to small eigenvalues, then the sample mean of the projected data points is a rate-optimal estimator. This allows us to iteratively reduce the dimension and eventually to estimate the remaining low-dimensional component of the mean by a standard robust estimator such as the componentwise median or the trimmed mean, see Algorithm 2.

The main contributions of this paper are methodological and theoretical. The SDR estimator, thoroughly defined in Section 3.2, is a fast and accurate method for robustly estimating the mean of a set of points. It depends on one tuning parameter, the threshold used for identifying and removing clear outliers, and on the dimension reduction regime. Our theoretical considerations provide some recommendations for their choices and our numerical experiments reported in Section 3.5 confirm the relevance of these choices. Importantly, the SDR estimator does not require as input the rate of contamination $\varepsilon$ but only an upper bound on $\varepsilon$. As for theoretical contributions of this paper, we state in Section 3.3 an upper bound on the error of the SDR estimator, showing that it is nearly minimax-rate optimal and has a breakdown point equal to $1/2$. This is done in the general case of a sub-Gaussian distribution with heterogeneous covariance matrix contaminated by adversarial noise. In Section 3.4, we further investigate the error of the SDR estimator in the case where only an approximation to the

| | Comput. tractable | Breakdown point | known $\varepsilon$ | Squared error rate | known $\mathbf{\Sigma}$ or $\mathbf{\Sigma} \propto \mathbf{I}$ |
|---|---|---|---|---|---|
| | Gaussian distribution | | | | |
| C./G. Median | yes | 0.5 | no | $\mathbf{r}_{\mathbf{\Sigma}}/n + \varepsilon^2 p$ | no |
| Tukey's Median | no | 0.33 | no | $\mathbf{r}_{\mathbf{\Sigma}}/n + \varepsilon^2$ | no |
| Agnostic Mean | yes | – | yes | $(p/n)\log^3 p + \varepsilon^2 \log p$ | yes |
| | Gaussian and sub-Gaussian distribution | | | | |
| Iter. Rew. Mean | yes | 0.28 | yes | $(\mathbf{r}_{\mathbf{\Sigma}}/n) + \varepsilon^2 \log(1/\varepsilon)$ | yes |
| Iterative Filtering | yes | – | yes | $(p/n)\log^a p + \varepsilon^2 \log(1/\varepsilon)$ | yes |
| SDR (this work) | yes | 0.5 | no | $(\mathbf{r}_{\mathbf{\Sigma}}/n + \varepsilon^2 \log(1/\varepsilon)) \log p$ | yes |

Table 3.1: Properties of various robust estimators. Agnostic mean, iteratively reweighted mean and iterative filtering are the estimators studied in (Lai et al., 2016), (Dalalyan and Minasyan, 2020) and (Diakonikolas et al., 2017), respectively. The error rates reported for Tukey's median, componentwise median, geometric median and the agnostic mean have been proved for non-adversarial contamination. The squared error rate is provided in the case of a covariance matrix satisfying $\|\mathbf{\Sigma}\|_{\mathrm{op}} = 1$.

covariance matrix is available.

The papers that are the closest to the present one are (Lai et al., 2016), (Diakonikolas et al., 2017) and (Dalalyan and Minasyan, 2020). The spectral dimension reduction scheme was proposed by (Lai et al., 2016) along with an initial sample splitting step ensuring the independence of the estimators over different subspaces. In the case of spherical Gaussian distribution contaminated by non-adversarial outliers, the paper states that the proposed estimator has a squared error at most of order $p \log^2 p \log(p/\varepsilon)/n + \varepsilon^2 \log p$. Compared to this, our results are valid in the more general setting of sub-Gaussian distribution, with arbitrary covariance matrix and adversarial contamination. In addition, our estimator does not rely on sample splitting and, therefore, has a risk with a better dependence on $p$. As compared to the filtering method of (Diakonikolas et al., 2017), our estimator has the advantage of being independent of $\varepsilon$ and our error bound is valid for every covariance matrix and every confidence level. On the down side, our error bound has an extra factor $\log p$ in front of $\varepsilon^2$. We believe that this factor is an artifact of the proof, but we were unable to remove it. Finally, compared to the iteratively reweighted mean (Dalalyan and Minasyan, 2020), the SDR estimator studied in the present paper has a higher breakdown point, does not require the knowledge of $\varepsilon$ and is much faster to compute. The advantages and shortcomings of these estimators are summarized in Table 3.1 and Figure 3.1.

*Notation.* For any pair of integers $k$ and $d$ such that $1 \le k \le d$, we denote by $\mathscr{V}_k^d$ the set of all $k$-dimensional linear subspaces $V$ of $\mathbb{R}^d$. For $V \in \mathscr{V}_k^d$, we write $k = \dim(V)$ and denote by $\mathrm{P}_V$ the orthogonal projection matrix onto $V$. $\mathbb{S}^{d-1}$ stands for the unit sphere in $\mathbb{R}^d$. For a $d \times d$ symmetric matrix $\mathbf{M}$, we denote by $\lambda_1(\mathbf{M}), \ldots, \lambda_d(\mathbf{M})$ its eigenvalues sorted in increasing order, and use the notation $\lambda_{\min}(\mathbf{M}) = \lambda_1(\mathbf{M})$, $\lambda_{\max}(\mathbf{M}) = \lambda_d(\mathbf{M})$, $\|\mathbf{M}\|_{\mathrm{op}} =$

$\max(|\lambda_{\min}(\mathbf{M})|, |\lambda_{\max}(\mathbf{M})|)$, $\mathsf{Tr}(\mathbf{M}) = (\lambda_1 + \ldots + \lambda_d)(\mathbf{M})$ and $\mathbf{r}_{\mathbf{M}} = \mathsf{Tr}(\mathbf{M})/\|\mathbf{M}\|_{\mathsf{op}}$. For any integer $n > 0$, we set $[n] = \{1, \ldots, n\}$. We will denote by $\mathcal{O} \subset [n]$ the subscripts of the outliers and by $\mathcal{I} = [n] \setminus \mathcal{O}$ the subscripts of inliers. We also use notation $\log_+(x) = \max\{0, \log(x)\}$. For a matrix $\mathbf{M}$, we denote by $\sigma_{\min}(\mathbf{M})$ and $\sigma_{\max}(\mathbf{M})$ respectively its smallest and largest singular values.

**Algorithm 2** SDR($\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; \boldsymbol{\Sigma}, t$)

**let** $p$ be the dimension of $\boldsymbol{X}_1$
**let** $\widehat{\boldsymbol{\mu}}^{\mathsf{GM}}$ be the geometric median of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$
**let** $\mathcal{S} \leftarrow \{i : \|\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}^{\mathsf{GM}}\| \le t\sqrt{p}\}$
**let** $\overline{\boldsymbol{X}}_{\mathcal{S}}$ be the sample mean of the filtered sample $\{\boldsymbol{X}_i : i \in \mathcal{S}\}$
**let** $\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}}$ be the covariance matrix of the filtered sample $\{\boldsymbol{X}_i : i \in \mathcal{S}\}$
**if** $p > 1$ **then**
  **let** $V$ be the span of the top $\lceil p/e \rceil$ principal components of $\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}} - \boldsymbol{\Sigma}$
  **let** $\mathtt{P}_V$ be the orth. projection onto $V$
  **let** $\mathtt{P}_{V^\perp}$ be the orth. projection onto the orth. complement of $V$
  **let** $\widehat{\boldsymbol{\mu}} \leftarrow \mathtt{P}_{V^\perp}\overline{\boldsymbol{X}}_{\mathcal{S}} + \mathsf{SDR}(\mathtt{P}_V\boldsymbol{X}_1, \ldots, \mathtt{P}_V\boldsymbol{X}_n; \mathtt{P}_V\boldsymbol{\Sigma}\,\mathtt{P}_V, t)$
**else**
  **let** $\widehat{\boldsymbol{\mu}} \leftarrow \widehat{\boldsymbol{\mu}}^{\mathsf{GM}}$
**end if**
**return** $\widehat{\boldsymbol{\mu}}$

## 3.2 Adversarially corrupted sub-Gaussian model and spectral dimension reduction

We assume that a set $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of $n$ data points drawn from a distribution $\mathsf{P}_n$ is given. This set is assumed to contain at least $n - [n\varepsilon]$ inliers, the remaining points being outliers. All the points lie in the $p$-dimensional Euclidean space and the inliers are independently drawn from a reference distribution, assumed to be sub-Gaussian with mean $\boldsymbol{\mu}^* \in \mathbb{R}^p$ and covariance matrix $\boldsymbol{\Sigma}$. To state the assumptions imposed on the observations in a more precise way, let us recall that the random vector $\boldsymbol{\zeta}$ is said to be sub-Gaussian with zero mean and identity covariance matrix, if $\mathbb{E}[\boldsymbol{\zeta}] = 0$, $\mathbb{E}[\boldsymbol{\zeta}\boldsymbol{\zeta}^\top] = \mathbf{I}_p$ and for some $\mathfrak{s} > 0$, we have

$$\mathbb{E}\big[e^{\boldsymbol{v}^\top\boldsymbol{\zeta}}\big] \le \exp\big\{\mathfrak{s}\|\boldsymbol{v}\|^2/2\big\}, \quad \forall \boldsymbol{v} \in \mathbb{R}^p.$$

The parameter $\mathfrak{s}$ is commonly called the variance proxy and the writing $\boldsymbol{\zeta} \sim \mathsf{SG}_p(\mathfrak{s})$ is used.

**Definition 2.** *We say that the data generating distribution $\mathsf{P}_n$ is an adversarially corrupted sub-Gaussian distribution with mean $\boldsymbol{\mu}^*$, covariance matrix $\boldsymbol{\Sigma}$, variance proxy $\mathfrak{s}$ and contamination rate $\varepsilon$, if there is a probability space on which we can define a sequence of random vectors $(\boldsymbol{X}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{Y}_n)$ such that*

1. *$\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are independent and $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}^*) \sim \mathsf{SG}_p(\mathfrak{s})$ for every $i \in [n]$.*

2. *the cardinality of* $\mathcal{O} = \{i \in [n] : \boldsymbol{Y}_i \neq \boldsymbol{X}_i\}$ *is at most equal to* $n\varepsilon$.

3. *the distribution of* $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ *is* $\mathsf{P}_n$.

*We write then*[1] $\mathsf{P}_n \in \mathsf{SGAC}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \mathfrak{s}, \varepsilon)$. *In the particular case where all* $\boldsymbol{Y}_i$ *are Gaussian, we will write* $\mathsf{P}_n \in \mathsf{GAC}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon)$.

For an overview of various kind of contamination models we refer the interested reader to (Bateni and Dalalyan, 2020). The adversarial contamination considered throughout this work is perhaps the most general one considered in the literature as the elements of the set $\mathcal{O}$—called outliers—may be chosen using $\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon$ but also $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ by an omniscient adversary. Note that in this setting, even the set $\mathcal{O}$ is random and depends on $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$. Therefore, the inliers $\{\boldsymbol{X}_i : i \in \mathcal{I}\}$ cannot be considered as independent random variables. The problem studied in this work consists in estimating the mean $\boldsymbol{\mu}^*$ of the reference distribution from the adversarially corrupted observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

The estimator we analyze in this work is termed iterative spectral dimension reduction and denoted by $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}$. It is closely related to the agnostic mean (Lai et al., 2016) and to iterative filtering (Diakonikolas et al., 2017) estimators. We will prove that SDR enjoys most of desired properties in the setting of robust estimation of the sub-Gaussian mean.

The parameters given as input to the iterative spectral dimension reduction algorithm are a strictly decreasing sequence of positive integers $p_0, \ldots, p_L$ such that $p_0 = p$ and a positive threshold $t > 0$. We recall that the geometric median is defined by

$$\widehat{\boldsymbol{\mu}}^{\mathsf{GM}} \in \arg\min_{\boldsymbol{\mu} \in \mathbb{R}^p} \sum_{i=1}^{n} \|\boldsymbol{X}_i - \boldsymbol{\mu}\|_2. \tag{3.1}$$

The algorithm for computing the SDR estimator reads as follows.

1. Start by setting $\mathbf{V}_0 = \mathbf{I}_p$.
2. For $\ell = 0, \ldots L-1$ do
   (a) Define $\bar{\boldsymbol{\mu}}^{(\ell)} \in \mathbb{R}^{p_\ell}$ as the geometric median of $\{\mathbf{V}_\ell^\top \boldsymbol{X}_i : i \in [n]\}$.
   (b) Define the set $\mathcal{S}^{(\ell)} = \big\{i \in [n] : \|\mathbf{V}_\ell^\top \boldsymbol{X}_i - \bar{\boldsymbol{\mu}}^{(\ell)}\|_2 \leq t\sqrt{p_\ell}\big\}$ of filtered data points.
   (c) Let $\overline{\boldsymbol{X}}^{(\ell)}$ and $\widehat{\boldsymbol{\Sigma}}^{(\ell)}$ be the mean vector and the covariance matrix of the filtered sample $\{\boldsymbol{X}_i : i \in \mathcal{S}^{(\ell)}\}$, that is
   $$\overline{\boldsymbol{X}}^{(\ell)} = \frac{1}{|\mathcal{S}^{(\ell)}|} \sum_{i \in \mathcal{S}^{(\ell)}} \boldsymbol{X}_i, \qquad \widehat{\boldsymbol{\Sigma}}^{(\ell)} = \frac{1}{|\mathcal{S}^{(\ell)}|} \sum_{i \in \mathcal{S}^{(\ell)}} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})^{\otimes 2}.$$
   (d) Set $\widehat{\boldsymbol{\mu}}^{(\ell)} = \mathbf{V}_\ell \mathbf{U}_\ell^\top \mathbf{U}_\ell \mathbf{V}_\ell^\top \overline{\boldsymbol{X}}^{(\ell)}$, where $\mathbf{U}_\ell$ is a $(p_\ell - p_{\ell+1}) \times p_\ell$ orthogonal matrix the rows of which are the eigenvectors of $\mathbf{V}_\ell^\top (\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma}) \mathbf{V}_\ell$ corresponding to its $(p_\ell - p_{\ell+1})$ smallest eigenvalues.

[1] SGAC stands for sub-Gaussian with adversarial contamination.

(e) Set $\mathbf{V}_{\ell+1} = \mathbf{V}_\ell(\mathbf{U}_\ell^\perp)^\top \in \mathbb{R}^{p\times p_{\ell+1}}$, where $\mathbf{U}_\ell^\perp$ is a $p_{\ell+1}\times p_\ell$ orthogonal matrix orthogonal to $\mathbf{U}_\ell$, that is $\mathbf{U}_\ell^\perp\mathbf{U}_\ell^\top = \mathbf{0}$.

3. Define $\bar{\boldsymbol{\mu}}^{(L)}$ as the geometric median of $\mathbf{V}_L^\top \boldsymbol{X}_i$ for $i = 1,\dots,n$ and set $\mathcal{S}^{(L)} = \big\{i \in [n] : \|\mathbf{V}_L^\top \boldsymbol{X}_i - \bar{\boldsymbol{\mu}}^{(L)}\|_2 \le t\sqrt{p_L}\big\}$.

4. Define $\widehat{\boldsymbol{\mu}}^{(L)} = \mathbf{V}_L\mathbf{V}_L^\top\, \overline{\boldsymbol{X}}^{(L)}$, the average of filtered and projected vectors.

5. Return $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}} = \widehat{\boldsymbol{\mu}}^{(0)} + \widehat{\boldsymbol{\mu}}^{(1)} + \ldots + \widehat{\boldsymbol{\mu}}^{(L)}$.

The steps described above can be summarised as follows. At each iteration $\ell < L$, we start by determining a filtered subsample $\mathcal{S}^{(\ell)}$ and a "nearly-outlier-orthogonal" subspace $\mathscr{U}_\ell = \mathrm{Im}(\mathbf{V}_\ell\mathbf{U}_\ell^\top)$ of $\mathbb{R}^p$ of dimension $p_\ell - p_{\ell+1}$. We define the projection of $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}$ onto $\mathscr{U}_\ell$ as the sample mean of the filtered and projected subsample, and we move to the next step for determining the projection of $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}$ onto the remaining part of the space. At the last iteration $L$, when the dimension is well reduced, the projection of $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}$ onto the subspace $\mathscr{U}_L$ is defined as the average of the filtered subsample projected onto $\mathscr{U}_L$. The subspaces $\mathscr{U}_\ell$ are two-by-two orthogonal and span the whole space $\mathbb{R}^p$. Each subspace is determined from the spectral decomposition of the covariance matrix of the data points projected onto $(\mathscr{U}_0 \oplus \ldots \oplus \mathscr{U}_{\ell-1})^\perp$, after removing the points lying at an abnormally large distance from the geometric median.

### 3.2.1 Choice of the dimension reduction regime

The analysis of the error of the SDR estimator conducted in this work leads to an upper bound in which the sequence $(p_0,\dots,p_L)$ is involved only through the expression

$$F(p_0,\dots,p_L) = \sum_{\ell=1}^{L} \frac{p_{\ell-1}}{p_\ell}.$$

Therefore, an appealing way of choosing this sequence is to minimize the function $F$ under the constraint that the sequence is decreasing and $p_0 = p$ and $p_L = 1$. It follows from the inequality between the arithmetic and geometric means that $F(p_0,\dots,p_L) \ge Lp^{1/L}$. Furthermore, the equality is achieved[2] in the case when all the terms in the definition of $F$ are equal, *i.e.*, when for some $c > 0$ we have $p_{\ell-1} = cp_\ell$ for every $\ell \in [L]$. Since $p_0 = p$ and $p_L = 1$, this yields $c = p^{1/L}$ or, equivalently, $L = \log p/\log c$. Using these relations, we find that the function $F$ is lower bounded by $Lc = (c/\log c)\log p$. The last step is to find the minimum of the function $c \mapsto c/\log c$ over the interval $(1,\infty)$. One easily checks that this function has a unique minimum at $c = e$. All these considerations advocate for using the dimension reduction regime defined by

$$p_0 = p, \quad p_\ell = \lfloor p_{\ell-1}/e \rfloor + 1,\ \ell \in [L], \quad p_L = 1, \tag{3.2}$$

[2]We relax here the assumption that all the entries $p_\ell$ are integers.

where $\lfloor x \rfloor$ is the largest integer strictly smaller than $x$. Such a definition of $(p_\ell)$ ensures that $p_{\ell-1}/p_\ell \le e$ and that[3] $L \le 2\log p$. In the rest of the paper, we assume that the sequence $(p_\ell)$ is chosen as in (3.2).

### 3.2.2 Choice of the threshold

The SDR procedure has one important tuning parameter: the threshold $t$ used to discard clearly outlying data points. Let us introduce the auxiliary notation

$$\bar{\mathsf{r}}_n = \frac{\sqrt{\mathbf{r}_{\mathbf{\Sigma}}} + \sqrt{2\log(2/\delta)}}{\sqrt{n}}, \qquad \text{and} \qquad \tau = \frac{1}{4} \bigwedge \frac{\bar{\mathsf{r}}_n}{\sqrt{\log_+(2/\bar{\mathsf{r}}_n)}}. \tag{3.3}$$

Note that $\bar{\mathsf{r}}_n$ is essentially the quantile, up to a universal constant factor, of order $1-\delta$ of the distribution of $\|\bar{\boldsymbol{Y}}_n - \boldsymbol{\mu}^*\|_2$ where $\boldsymbol{Y}_i$'s are independently drawn from $\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{\Sigma})$ with $\|\mathbf{\Sigma}\|_{\mathrm{op}} = 1$. Our theoretical results advocate for using the value $t = t_1 + t_2$, where

$$t_1 = \frac{2(1+\bar{\mathsf{r}}_n)}{1-2\varepsilon^*}, \qquad t_2 = 1 + \frac{\bar{\mathsf{r}}_n}{\sqrt{\tau}} + \sqrt{2 + \log(2/\tau)},$$

where $\varepsilon^* < 1/2$ is the largest value of the contamination rate that the algorithm may handle.

Let $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n$ be independent Gaussian with zero mean and covariance $\mathbf{\Sigma}$. The expression of $t_1$ is obtained as an upper bound on the quantile of order $1-\delta/2$ of the distribution of the random variable

$$T_1 = \sup_V \frac{2}{n(1-2\varepsilon)\dim(V)} \sum_{i=1}^n \|\mathrm{P}_V \boldsymbol{\xi}_i\|_2,$$

see Lemma 2 and its proof for further details. Similarly, $t_2$ is defined so that the event

$$\sup_V \sum_{i=1}^n \mathbb{1}\big(\|\mathrm{P}_V \boldsymbol{\xi}_i\|_2^2 > t_2^2 \dim(V)\big) \le n\tau$$

has a probability at lest $1-\delta/2$. The related computations are deferred to Appendix B.1.3. Although we tried to get sharp values for these thresholds $t_1$ and $t_2$, it is certainly possible to improve these values either by better mathematical arguments or by empirical considerations. Of course, smaller values of the thresholds $t_1$ and $t_2$ satisfying aforementioned conditions lead to an SDR estimator having smaller error.

## 3.3 Assessing the error of the SDR estimator

The iterative spectral dimension reduction estimator defined in previous sections has some desirable properties of a robust estimator that are easy to check. In particular, it is clearly

[3] To check this inequality, one can use the fact that $3 \le p_{L-2} \le pe^{2-L} + e/(e-1)$. This implies $L \le 2\log p$ for $p \ge 6$. For smaller values of $p$, the inequality can be checked by direct computations.

equivariant by translation, orthogonal linear transform and global scaling. Furthermore, the breakdown point of the estimator is equal to that of the geometric median, that is to $1/2$. This means that even if almost the half of data points are chosen to be infinitely large, the estimator will not "break down" in the sense of becoming infinitely large. However, the fact that the estimated value does not become infinitely large, it might be not very close to the true mean. The next theorem shows that this is not the case and that the error of the SDR estimator has a nearly rate-optimal behavior even when the contamination rate is close to $1/2$. The adverb "nearly" is used here to reflect the presence of the $\sqrt{\log p}$ factor in the error bound, which is not present in the minimax rate.

**Theorem 11.** *Let $\varepsilon^* \in (0, 1/2)$, and $\delta \in (0, 1/2)$. Define $\bar{\mathsf{r}}_n$ and $\tau$ as in (3.3). For every $\varepsilon \leq \varepsilon^*$, let $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}$ be the estimator returned by Algorithm 3.2 with*

$$t = \frac{3 - 2\varepsilon^*}{1 - 2\varepsilon^*}\left(1 + \frac{\bar{\mathsf{r}}_n}{\sqrt{\tau}}\right) + \sqrt{2 + 2\log\left(1/\tau\right)}.$$

*There exists a universal constant $\mathsf{C}$ such that for every $\mathsf{P}_n \in \mathsf{GAC}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon)$ with $\varepsilon \leq \varepsilon^*$ and[4] $\|\boldsymbol{\Sigma}\|_{\mathrm{op}} = 1$, the probability of the event*

$$\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}} - \boldsymbol{\mu}^*\right\|_2 \leq \frac{\mathsf{C}\sqrt{\log p}}{1 - 2\varepsilon^*}\left(\sqrt{\frac{\mathbf{r}_{\boldsymbol{\Sigma}}}{n}} + \varepsilon\sqrt{\log(2/\varepsilon)} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$$

*is at least $1 - \delta$. Moreover, the constant $\mathsf{C}$ from the last display can be made explicit by replacing the effective rank $\mathbf{r}_{\boldsymbol{\Sigma}}$ by the dimension $p$ in the definition of $\bar{\mathsf{r}}_n$: That is, for every $\delta \in (0, 1/5)$ the inequality*

$$\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}} - \boldsymbol{\mu}^*\right\|_2 \leq \frac{156\sqrt{2\log p}}{1 - 2\varepsilon^*}\left(\sqrt{\frac{2p}{n}} + \varepsilon\sqrt{\log(2/\varepsilon)} + \sqrt{\frac{3\log(2/\delta)}{n}}\right)$$

*holds with probability at least $1 - 5\delta$.*

*Proof in the appendix, page 92*

If we compare this result with its counterpart established in (Dalalyan and Minasyan, 2020) for the iteratively reweighted mean, besides the extra $\log p$ factor, we see that the above error bound does not reduce to the error of the empirical mean when the contamination rate goes to zero. We do not know whether this is just a drawback of our proof, or it is an intrinsic property of the estimator. Our numerical experiments reported later on suggest that is might be a property of the estimator.

There is another logarithmic factor, $\sqrt{\log(2/\varepsilon)}$, present in the second term of the error bounds provided by the last theorem, which does not appear in the minimax rate. There are computationally intractable robust estimators of the Gaussian mean, such as the Tukey median, that have an error bound free of this factor. However, all the known error bounds

[4] Since in this theorem $\boldsymbol{\Sigma}$ is assumed to be known, we can always divide all the data points $\boldsymbol{X}_i$ by $\|\boldsymbol{\Sigma}\|_{\mathrm{op}}^{1/2}$ to get a data set with a covariance matrix satisfying $\|\boldsymbol{\Sigma}\|_{\mathrm{op}} = 1$.

provably valid for polynomial time algorithms has this extra $\sqrt{\log(2/\varepsilon)}$ factor. Furthermore, this factor is known to be unavoidable in the case of sub-Gaussian model with adversarial contamination[5], see (Lugosi and Mendelson, 2021, Section 2).

As shows the next theorem, the claims of Theorem 11 carry over the sub-Gaussian reference distributions with some slight modifications. These modifications mainly stem from the following lemma assessing the tail behavior of the singular values of a matrix having independent and sub-Gaussian columns.

**Lemma 1** (Vershynin (2012), Theorem 5.39)**.** *Let $\boldsymbol{\xi}_{1:n}$ be a matrix consisting of sub-Gaussian vectors with variance proxy $\mathfrak{s}$. There is a universal constant $\mathsf{C}_0$ such that for every $t > 0$ and for every pair of positive integers $n$ and $p$, we have*

$$\mathbf{P}\big(\sigma_{\min}(\boldsymbol{\xi}_{1:n}) \le \sqrt{n} - \mathsf{C}_0\mathfrak{s}(\sqrt{p} + t)\big) \le e^{-t^2},$$
$$\mathbf{P}\big(\sigma_{\max}(\boldsymbol{\xi}_{1:n}) \ge \sqrt{n} + \mathsf{C}_0\mathfrak{s}(\sqrt{p} + t)\big) \le e^{-t^2}.$$

Note that in the Gaussian case $\mathfrak{s} = 1$ and the constant $\mathsf{C}_0$ can be chosen equal to $\sqrt{2}$. The last lemma leads to the following adaptations in the values of the thresholds used in the SDR estimator. First, we introduce auxiliary definitions

$$\tau = \frac{1}{4} \bigwedge \frac{\bar{\mathsf{r}}_{n,\mathfrak{s}}}{\sqrt{\log_+(2/\bar{\mathsf{r}}_{n,\mathfrak{s}})}}, \qquad \text{with} \qquad \bar{\mathsf{r}}_{n,\mathfrak{s}} = \frac{3\sqrt{\mathfrak{s}}\big(\sqrt{p} + 2\sqrt{\log(2/\delta)}\big)}{\sqrt{n}}. \tag{3.4}$$

Then, we set $t = t_1 + t_2$ with

$$t_1 = \frac{2(1 + \mathsf{C}_0\,\bar{\mathsf{r}}_{n,\mathfrak{s}}\sqrt{\mathfrak{s}})}{1 - 2\varepsilon^*}, \qquad t_2 = 1 + \mathsf{C}_0\sqrt{\mathfrak{s}}\Big(\frac{\bar{\mathsf{r}}_{n,\mathfrak{s}}}{\sqrt{\tau}} + \sqrt{2 + \log(1/\tau)}\Big),$$

where $\mathsf{C}_0$ is the same as in Lemma 1.

Now we are ready to state the theorem for the sub-Gaussian distributions showing that the SDR estimator with the threshold depending on the variance proxy $\mathfrak{s}$ yields the same upper bound on $\ell_2$ distance between our estimator $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}$ and the true value $\boldsymbol{\mu}^*$ replacing the effective rank $\mathbf{r}_{\boldsymbol{\Sigma}}$ with the dimension $p$.

**Theorem 12** (Sub-Gaussian version)**.** *Let $\varepsilon^* \in (0, 1/2)$, and $\delta \in (0, 1/2)$. Define $\bar{\mathsf{r}}_{n,\mathfrak{s}}$ and $\tau$ as in (3.4). For every $\varepsilon \le \varepsilon^*$, let $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}$ be the estimator returned by Algorithm 3.2 with*

$$t = \frac{3 - 2\varepsilon^*}{1 - 2\varepsilon^*}\left(1 + \mathsf{C}_0\,\bar{\mathsf{r}}_{n,\mathfrak{s}}\sqrt{\frac{\mathfrak{s}}{\tau}}\right) + \mathsf{C}_0\,\mathfrak{s}\sqrt{2 + 2\log\big(1/\tau\big)},$$

*where $\mathsf{C}$ is a universal constant. Then, there exists a constant $\mathsf{C}_{\mathfrak{s}}$ depending only on the variance proxy $\mathfrak{s}$ such that for every $\mathsf{P}_n \in \mathsf{SGAC}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \mathfrak{s}, \varepsilon)$ with $\varepsilon \le \varepsilon^*$ and $\|\boldsymbol{\Sigma}\|_{\mathrm{op}} = 1$, the*

[5]Both sub-Gaussianity of the reference distribution and the adversarial nature of the contamination are important for getting the extra $\sqrt{\log(2/\varepsilon)}$ factor in the minimax rate.

*probability of the event*

$$\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}} - \boldsymbol{\mu}^*\right\|_2 \leq \frac{\mathsf{C}_{\mathfrak{s}}\sqrt{\log p}}{1-2\varepsilon^*}\left(\sqrt{\frac{p}{n}} + \varepsilon\sqrt{\log(2/\varepsilon)} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$$

*is at least* $1-\delta$.

*Proof in the appendix, page 111*

## 3.4 The case of unknown covariance matrix

The SDR estimator, as defined in Algorithm 2, requires the knowledge of covariance matrix $\boldsymbol{\Sigma}$. In this section we consider the case where the matrix $\boldsymbol{\Sigma}$ is unknown, but an approximation of the latter is available. Namely, we assume that we have access to a matrix $\widetilde{\boldsymbol{\Sigma}}$ and to a real number $\gamma > 0$ such that $\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathsf{op}} \leq \gamma\|\boldsymbol{\Sigma}\|_{\mathsf{op}}$. In such a situation, we can replace in the SDR estimator the true covariance matrix by its approximation $\widetilde{\boldsymbol{\Sigma}}$. This will necessarily require to adjust the threshold $t$ accordingly. The goal of the present section is to propose a suitable choice of $t$ and to show the impact of the approximation error $\gamma$ on the estimation accuracy.

As mentioned, the parameter $t$ used in Algorithm 2 needs to be properly tuned in order to account for the approximation error in the covariance matrix. To this end, we introduce the following auxiliary notation similar to those presented in (3.3):

$$\widetilde{\mathsf{r}}_n = \frac{\sqrt{\mathsf{C}_\gamma \mathbf{r}_{\widetilde{\boldsymbol{\Sigma}}}} + \sqrt{2\log(2/\delta)}}{\sqrt{n}} \qquad \text{and} \qquad \widetilde{\tau} = \frac{1}{4} \bigwedge \frac{\widetilde{\mathsf{r}}_n}{\sqrt{\log_+(2/\widetilde{\mathsf{r}}_n)}}, \tag{3.5}$$

where $\mathsf{C}_\gamma = (1+\gamma)/(1-\gamma)$. Compared to (3.3), the main difference here is the presence of the factor $\mathsf{C}_\gamma$ (which is equal to one if $\gamma = 0$) and the substitution of the effective rank of $\boldsymbol{\Sigma}$ by that of its approximation $\widetilde{\boldsymbol{\Sigma}}$. In the rest of this section, we assume that $\boldsymbol{\Sigma}$ is invertible.

**Theorem 13.** *Let* $\varepsilon^* \in (0, 1/2)$, $\delta \in (0, 1/2)$ *and define* $\widetilde{\mathsf{r}}_n$ *and* $\tau$ *as in* (3.5). *Assume that* $\widetilde{\boldsymbol{\Sigma}}$ *satisfies* $\|\boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2} - \mathbf{I}_p\|_{\mathsf{op}} \leq \gamma$ *for some* $\gamma \in (0, 1/2]$. *Let* $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}$ *be the output of* $\mathsf{SDR}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; \widetilde{\boldsymbol{\Sigma}}, \widetilde{t}_\gamma)$, *see Algorithm 2, with*

$$\widetilde{t}_\gamma = \frac{\|\widetilde{\boldsymbol{\Sigma}}\|_{\mathrm{op}}}{1-\gamma}\left\{\frac{3-2\varepsilon^*}{1-2\varepsilon^*}\left(1 + \frac{\widetilde{\mathsf{r}}_n}{\sqrt{\widetilde{\tau}}}\right) + \sqrt{2 + \log\left(2/\tau\right)}\right\}.$$

*Then, there exists a universal constant* $\mathsf{C}$ *such that for every data generating distribution* $\mathsf{P}_n \in \mathsf{GAC}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon)$ *with* $\varepsilon \leq \varepsilon^*$, *the probability of the event*

$$\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}} - \boldsymbol{\mu}^*\right\|_2 \leq \frac{\mathsf{C}\,\|\boldsymbol{\Sigma}\|_{\mathsf{op}}^{1/2}\sqrt{\log p}}{1-2\varepsilon^*}\left(\sqrt{\frac{\mathbf{r}_{\boldsymbol{\Sigma}}}{n}} + \varepsilon\sqrt{\log(2/\varepsilon)} + \sqrt{\varepsilon\gamma} + \sqrt{\frac{\log(1/\delta)}{n}}\right) \tag{3.6}$$

*is at least* $1-\delta$.

*Proof in the appendix, page 106*

On the one hand, if the value of $\gamma$ is at most of order $\sqrt{(\mathbf{r}_{\mathbf{\Sigma}}/n)\log(1/\varepsilon)} + \varepsilon\log(1/\varepsilon)$ then Theorem 13 implies that the estimation error is of the same order as in the case of known covariance matrix $\mathbf{\Sigma}$ (Theorem 11). For instance, if the matrix $\mathbf{\Sigma}$ is assumed to be diagonal, one can defined $\widetilde{\mathbf{\Sigma}}$ as the diagonal matrix composed of robust estimators of the variances of univariate contaminated Gaussian samples; see, for instance, Section 2 in (Comminges et al., 2021). For recent advances on robust estimation of (non-diagonal) covariance matrices by computationally tractable algorithms we refer the reader to (Cheng et al., 2019b).

On the other hand, if the value of $\gamma$ for which the condition $\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\mathrm{op}} \leq \gamma\|\mathbf{\Sigma}\|_{\mathrm{op}}$ is known to be true is of larger order than $\sqrt{(\mathbf{r}_{\mathbf{\Sigma}}/n)\log(1/\varepsilon)} + \varepsilon\log(1/\varepsilon)$, then $\sqrt{\varepsilon\gamma}$ dominates the other terms appearing in the error bound (3.6). Moreover, if $\gamma$ is of constant order, then we get the error rate $\sqrt{\frac{\mathbf{r}_{\mathbf{\Sigma}}}{n}} + \sqrt{\varepsilon}$, which is in line with previously known bounds for computationally tractable estimators; see for example (Lai et al., 2016, Theorem 1.1), (Diakonikolas et al., 2017, Theorem 3.2), (Dalalyan and Minasyan, 2020, Theorem 4).

## 3.5 Numerical experiments

We conducted numerical experiments on synthetic contaminated data to corroborate our theoretical results. The main goal of these experiments is to display statistical and computational features of the SDR and their dependence on various parameters. Moreover, we compared SDR to some other estimators proposed in the literature as well as to the oracle (empirical mean of the inliers). To do so, we selected componentwise median (CM), geometric median (GM) and Tukey's median (TM) as the three classic estimators of the context, and the iteratively reweighted mean (IRM), introduced in (Dalalyan and Minasyan, 2020), as an example of optimization based method.

### 3.5.1 Implementation details

The experiments were run on a laptop with a 1.8 GHz Intel Core i7 and 8 GB of RAM. R codes of the experiments are freely available on the last author's website. For GM and TM the R packages Gmedian[6] (Cardot et al. (2013)) and TukeyRegion[7] (Liu et al. (2019)) were used. IRM had been already implemented in R using Mosek[8].

To optimize SDR, several choices were made. First, since geometric median is used in SDR as a rough estimator of the location, we limited it to at most 15 iterations and to stop at an accuracy of order 1. See the reference manual of Gmedian to have more details on these parameters. Second, since at the last step of the SDR one can use any estimator which is robust in low-dimensional setting, we chose to use the median of the projected data points.

[6] https://cran.r-project.org/package=Gmedian
[7] https://cran.r-project.org/web/packages/TukeyRegion/index.html
[8] www.mosek.com

Finally, we adjusted the numerical constant in the threshold $t$. In all the experiments, we assumed that the true value of $\varepsilon$ is known and used $\varepsilon^* = \varepsilon$.

### 3.5.2 Experimental setup

Experiments were conducted on synthetic data sets obtained by applying a contamination scheme to $n$ i.i.d. samples drawn from $\mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$. The following contamination schemes were considered.

- *Contamination by uniform outliers* (CUO): the locations of $n\varepsilon$ outliers are chosen at random independently of the inliers. The outliers are independent Gaussian with identity covariance matrix and with means having coordinates independently drawn from the uniform in $[0, 3]$ distribution.
- *Gaussian mixture contamination* (GMC): the locations of $n\varepsilon$ outliers are chosen at random independently of the inliers. The outliers are independent $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{I}_p)$. In our experiments, we chose $\boldsymbol{\mu}$ such that $\|\boldsymbol{\mu}\| = 15$.
- *Contamination by "smallest" eigenvector* (CSE): We replace the $n\varepsilon$ samples most correlated with the smallest principal eigenvector $\boldsymbol{v}_p$ of the sample covariance matrix, by $n\varepsilon$ vectors all equal to $\sqrt{p}\,\boldsymbol{v}_p$ ($\boldsymbol{v}_p$ is assumed to be a unit vector). In contrast with the two previous schemes, this one is adversarial.

Each experiment was repeated 50 times for SDR, CM, GM, the oracle and 10 times for IRM and TM. The tolerance probability $\delta$ was set to 0.1 in all the experiments. In the figures, points on the curves are median values of the error or of the running time for these trials whereas vertical bars overlaid on the points show the spread between the first and third quartiles. Since the computation of TM is prohibitively costly and is possible only for small sample sizes and dimensions, it is excluded from most of the experiments.

### 3.5.3 Statistical accuracy

At the upper left panel of Figure 3.2, we illustrate the behavior of the risk when the sample size increases for four different contamination levels: $\varepsilon \in \{0.1, 0.2, 0.3, 0.4\}$. The data are of dimension 60 and generated by the GMC scheme. The median estimation error converges respectively to the values 0.18, 0.36, 0.62 and 1.06. According to our theoretical result, the limit of the error should be proportional to $\frac{\varepsilon \log(1/\varepsilon)}{1-2\varepsilon}$. This is confirmed by the experimental results, since the ratio between the empirical limit of the median error and $\frac{\varepsilon \log(1/\varepsilon)}{1-2\varepsilon}$ for each $\varepsilon$ is between 0.58 and 0.69.

At the upper right panel of Figure 3.2, the dependence of the error on the dimension is displayed. To better illustrate the effect of the dimension on the estimation error, we carried out our experiment on data sets of small sample size $n = 100$ with CUO contamination.

We compared the error of GM, CM, IRM and the oracle. In this plot, we clearly observe the supremacy of SDR and IRM as compared to GM and CM, which is in line with theoretical results. An important observation is that the error of the SDR estimator is very close to those of the IRM estimator and the oracle. This suggests that the factor $\sqrt{\log p}$ present in our theoretical results might be an artifact of the proof rather than an intrinsic property of the estimator, at least for nonadversarial contamination.

The last experiment aiming to display the behavior of the estimation error is depicted in the lower left panel of Figure 3.2. The examined synthetic datasets were generated by the CSE scheme with $\varepsilon = 0.2$. We measured the error for different values of the dimension and for sample size $n = 10p$ proportional to the dimension. In this case, the term $\sqrt{p/n}$ in the risk bound remains unchanged and we may perceive if the dimension virtually effects the term dependent on $\varepsilon$ in the bound. The plot clearly confirms that the error is stable for SDR as it is for IRM and the oracle, in sharp contrast with GM and CM. The last point, of course, is not surprising since the risks of GM and CM scale as $\varepsilon\sqrt{p}$. Once again, this plot suggests that the factor $\sqrt{\log p}$ present in the SDR's risk bound might be unnecessary.

### 3.5.4 Computational efficiency

We conducted another experiment in order to better understand the computational complexity of SDR. Note that the computational cost of SDR comes from two operations done at each iterations: SVD of sample covariance matrix and computation of geometric median. We see that SDR can be computed in a reasonable time even in high dimensions. For instance, for $n = 10000$ and $p = 1000$ it takes nearly 26 seconds (tested over 20 trials).

At the lower right panel of Figure 3.2, we plotted the running times (in seconds) of GM, CM, IRM and SDR for various dimensions. Sample size in this experiment was set to 100, contamination rate was $\varepsilon = 0.2$ and CSE contamination scheme was used. As expected, IRM has substantially larger running time compared to SDR, GM and CM; this is due to the semidefinite programming solver running at each iteration of IRM. The fact that SDR is faster than GM (even though GM is deployed at each iteration of SDR) is explained by our choice of computing only a rough approximation of GM within SDR (limiting to 15 iterations and a tolerance parameter set to 1).

### 3.5.5 Breakdown point

A natural measure of robustness of an estimator is its resistance to a large fraction of outliers. The goal here is to demonstrate empirically our theoretical result showing that the breakdown point of the SDR estimator is 1/2.

In Figure 3.3, at the left panel, we evaluated the error of the estimators on samples of size 100 and dimension 10 generated by the CSE scheme, for various values of $\varepsilon$. We can observe that SDR preserves its robustness with large contamination rates and outperforms

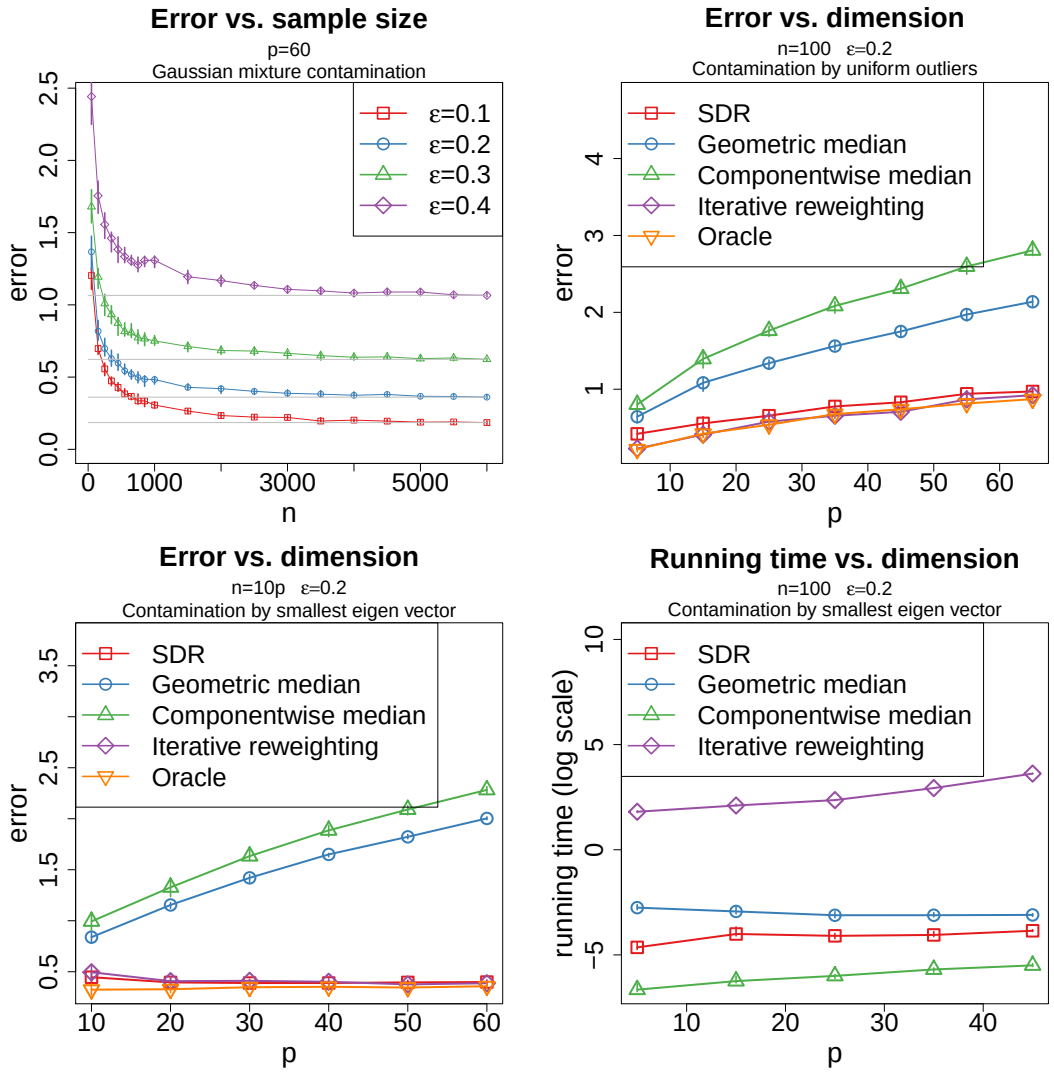


Figure 3.2: The upper left panel illustrates the convergence of SDR's median error when the sample size tends to infinity, for various contamination rates. The limiting values are shown by gray lines. The upper right panel shows the effect of dimension on the error. We see that in the case of SDR this effect is almost the same as for IRM and the oracle. The lower left panel plots the quantities when the sample-size increases proportionally to the dimension. Once again, we see that SDR is almost as accurate as IRM and the oracle. The lower right panel plots the running times of different estimators for various dimensions. It shows the huge computational gain of the SDR estimator as compared to IRM.

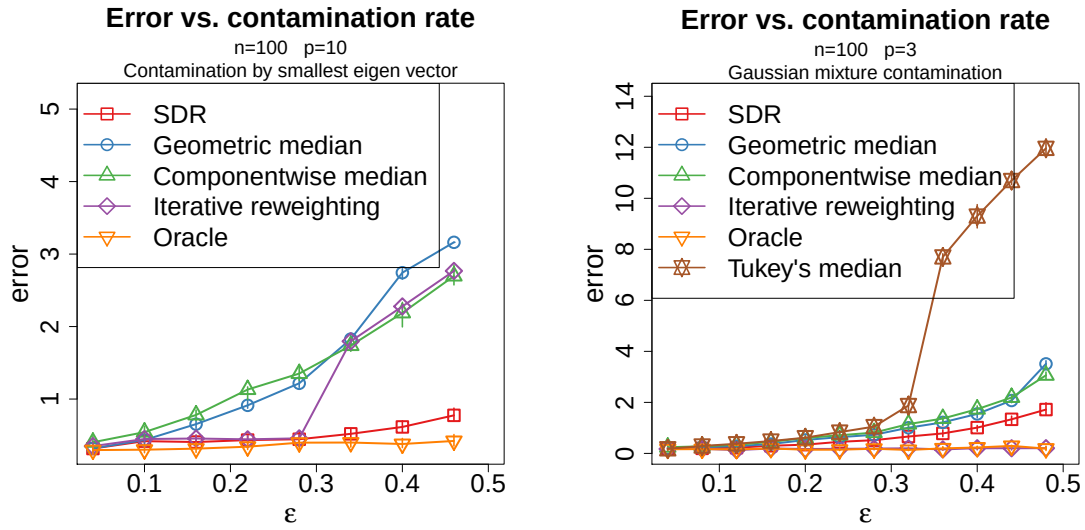


Figure 3.3: The left panel compares the robustness of various estimators by displaying the estimation error for different contamination rates under SCE scheme. SDR outperforms other estimators (even the oracle). Results displayed in the right panel are obtained by a similar experiment conducted for the GMC scheme. SDR is remarkably stable for different contamination schemes, while we see that SDR and Tukey's median may behave poorly for $\varepsilon > 1/3$.

other estimators, excepted the oracle. More precisely, SDR and IRM have roughly the same error up to $\varepsilon = 0.28$. Starting from this value, the error of IRM starts a steep deterioration joining CM and GM.

At the right panel of Figure 3.3, we plotted the error as a function of the contamination rate for TM, CM, GM, IRM and SDR. Data used in this experiment were of size 100 and dimension 3, corrupted by GMC scheme. For this type of contamination, we observe that the IRM estimator remains robust even for $\varepsilon$ close to $1/2$, whereas the error of TM deteriorates significantly for $\varepsilon > 1/3$. As a conclusion, for two contamination schemes which are challenging for iteratively reweighted mean and Tukey's median, SDR shows very stable behavior.

## 3.6 Summary, related work and conclusion

We have proved that the multivariate mean estimator obtained by the iterative spectral dimension reduction method enjoys several appealing properties in the setting of sub-Gaussian observations subject to adversarial contamination. More precisely, in addition to being rigid transform equivariant and having breakdown point equal to $1/2$, the estimator has been shown to achieve the nearly minimax rate. Furthermore, the SDR estimator has low computational complexity, confirmed by reported numerical experiments. Indeed, its computational complexity is of the same order as that of computing the sample covariance matrix and performing a SVD on it. Presumably, at the cost of a moderate drop in accuracy, further speed-ups can be obtained by randomization (Halko et al., 2011) in the spirit of the prior work (Cheng et al., 2019a; Depersin and Lecué, 2022).

Notably, we have proved that the SDR estimator achieves the nearly optimal error rate without requiring the precise knowledge of the contamination rate. It however requires the knowledge of the covariance matrix. To alleviate this constraint, we have also established estimation guarantees in the case where an approximation of the covariance matrix is used instead of the true one. We have conducted numerical tests that show that the SDR is both fast and accurate.

Many recent works studied the problem of robust estimation in more complex high dimensional settings such linear regression or sparse mean and covariance estimation; see (Balakrishnan et al., 2017; Cheng et al., 2021; Chinot, 2020; Chinot et al., 2020; Collier and Dalalyan, 2019; Dalalyan and Thompson, 2019; Goes et al., 2020; Liu et al., 2020b; Pensia et al., 2020) and the references therein. It is under current investigation whether the results of the present paper can be extended to these settings.

Another interesting avenue for future research is to find an estimator that is rate-optimal, computationally tractable, with breakdown point equal to $1/2$ and, in the same time, asymptotically optimal in the sense that its risk is of order $\sqrt{p/n}$ when $\varepsilon$ tends to zero. On a related note, it would be interesting to push further the exploration of second-order properties of the risk started in (Minasyan, 2020). Finally, an open question is how the minimax risk blows-up when the contamination rate tends to 1/2. For the SDR estimator studied in this work, we established an upper bound of order $1/(1-2\varepsilon)$. However, we have no clue whether this is optimal. Our intuition is that it is not as the lower bound provided in Chapter 1 is of order $\log^{1/2}\big(1/(1-2\varepsilon)\big)$ when $\varepsilon$ is close to 1/2.

# Chapter 4

# Discussion

In this manuscript, we investigated the problem of the robust estimation of the mean against outliers for two types of distributions: distributions supported by probability simplex and multivariate sub-Gaussian distributions. At the beginning, we presented existing approaches for dealing with outliers. We reviewed some folklore results and in particular we outlined the main challenge of the problem in high-dimensional settings which is the reconciliation of the statistical accuracy and computational efficiency.

We started our study by the simpler case of the distributions supported by the probability simplex, a special case of which is the the category of the discrete distributions. We established the minimax rates under three different distances: total variation, Hellinger and Euclidean distances. We discovered that the convergence rate is different under each distance. Then, we proposed confidence regions shrinking at the minimax rate and instance based bounds. Finally, we provided some experimental results corroborating the established theoretical rates. In addition, throughout this study, we presented various contamination models and relations among them.

In the next stage, we considered the problem of the robust estimation of the multivariate sub-Gaussian mean under the adversarial contamination model. Unlike to the case of the distributions supported by probability simplex, the sample mean is not a robust estimator for the Gaussian mean. We designed a new robust estimator (called SDR), based on the idea of spectral dimension reduction proposed in (Lai et al., 2016). SDR enjoys a near optimal risk rate, a high breakdown point (1/2), a low computational complexity and is equivariant by similarity transforms. While SDR has the advantage of not requiring the knowledge of the contamination rate, it requires the knowledge of the covariance matrix, as it is the case for all the proposed computational tractable estimators attaining a near optimal error rate (we outlined this fact in Section 1.3.4). However, to alleviate this constraint, we established estimation guarantees in the case where an approximation of the covariance matrix is used instead of the true one. At the end, we presented some practical evidences for the theoretical properties of SDR by carrying out different experiments on synthetic data and comparing the performance of SDR to that of some other existing estimators in the literature.

An avenue for future research is to extend the SDR for robust realization of other statistical tasks such as linear regression, covariance matrix estimation, etc. It is also interesting to investigate the properties of SDR with respect to the pseudo-norms of the form $\boldsymbol{x} \in \mathbb{R}^p \mapsto \|\boldsymbol{x}\|_S = \sup_{v \in S} \langle \boldsymbol{v}, \boldsymbol{x} \rangle$ where $S$ is a symmetric subset of $\mathbb{R}^p$.

Despite the existence several tractable robust estimators achieving a near optimal risk rate (with dependency $\varepsilon\sqrt{\log(1/\varepsilon)}$), and the evidence provided by Diakonikolas et al. (2016b) on the necessity of the factor $\varepsilon\sqrt{\log(1/\varepsilon)}$ for tractable robust estimators under the adversarial contamination, there are still some open questions. Is there an estimator for the adversarial contamination model that is rate-optimal, computationally tractable, with breakdown point equal to $1/2$, not requiring the knowledge of the contamination rate, and in the same time, asymptotically optimal in the sense that its risk is equivalent to $\sqrt{p/n}$ when $\varepsilon$ tends to zero? Is there a tractable estimator for Huber's contamination which attains the optimal error rate $\sqrt{p/n} + \varepsilon$?

On another note, Tukey's median which is considered as an optimal estimator in the robust estimation context has still unknown properties to be explored. For example, in the case of the non-spherical Gaussian distributions, where the covariance matrix of the reference distribution is the matrix $\boldsymbol{\Sigma}$, does Tukey's median enjoy the optimal risk rate $\sqrt{\mathsf{Tr}(\boldsymbol{\Sigma})/n} + \|\boldsymbol{\Sigma}\|_{\mathrm{op}}^{1/2}\varepsilon$? How does it perform in robust estimation of the mean for other distributions such as distributions with bounded moments? Let $\varepsilon^*$ denote the finite-sample breakdown point of Tukey's median. In Chapter 1, we showed that for multivariate spherical Gaussian data, $\varepsilon^*$ satisfies

$$C\frac{p}{n} \leq \frac{1}{3} - \varepsilon^* \leq C'\sqrt{\frac{p}{n}},$$

where $C$ and $C'$ are positive constants. Can we improve the lower bound of $\frac{1}{3} - \varepsilon^*$ to $\sqrt{\frac{p}{n}}$? And there are questions on the computational aspects of Tukey's median. Tukey's median is a NP-hard problem, but, is there an approximation algorithm returning a point with an acceptable high Tukey's halfspace depth? We hope that these questions yield new directions and creative approaches in statistics and computer science.

# Chapter 5

# Introduction en français



## 5.1 Définition du problème

On observe des données indépendamment identiquement distribuées (iid) selon une loi de référence. On suppose que cette loi admet une moyenne et on souhaite estimer cette moyenne à partir des observations. Dans le cas des distributions sous-gaussiennes, on sait que la moyenne empirique est un estimateur optimal dans le sens minimax, c'est-à-dire que l'erreur d'estimation de la moyenne empirique dans son pire cas est au moins aussi bonne que celle de tout autre estimateur dans son pire cas respectif. Mais que se passe-t-il si parmi nos données, il y a quelques-unes qui ne suivent pas la loi de référence ? Ces données désobéissantes que l'on appellera *données aberrantes* ou *outlier*, peuvent avoir des effets significatifs et indésirables sur notre estimations. Par exemple, la présence d'une donnée aberrante extrêmement éloignée de la moyenne peut considérablement dévier la moyenne empirique et entraîner une grande erreur d'estimation. Ainsi, la moyenne empirique n'est pas robuste contre les données aberrantes.

Dans ce travail, on étudie le problème de l'estimation robuste de la moyenne. On suppose que $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ sont $n$ échantillons iid selon une loi $\boldsymbol{P}_{\boldsymbol{\theta}}$ où $\boldsymbol{\theta}$ est la moyenne de cette loi. Étant donné $\varepsilon \in [0, 1/2)$, on observe des échantillons $\varepsilon$-contaminés $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ ce qu'il veut dire

que $\boldsymbol{X}_i = \boldsymbol{Y}_i$ si $i \in \mathcal{I}$ et que $\boldsymbol{X}_i$ prend une valeur arbitraire si $i \in \mathcal{O}$ où $|\mathcal{I}| = n(1-\varepsilon)$ et $|\mathcal{O}| = n\varepsilon$[1]. On pose $\mathcal{S} := \mathcal{I} \cup \mathcal{O} = \{1, \dots, n\}$. Le but est d'estimer $\boldsymbol{\theta}$ à partir de $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$.

## 5.2 Variables discrètes

Pour commencer, on considère le cas où la loi de référence $\boldsymbol{P}_{\boldsymbol{\theta}}$ est une distribution discrète à $k$ valeurs distinctes. On représente les variables aléatoires issue de cette loi par des vecteurs de dimension $k$ en utilisant l'encodage one-hot. Ainsi, $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ et $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ appartiennent à l'ensemble $E_k := \{\boldsymbol{e}_1, \dots, \boldsymbol{e}_k\}$, l'ensemble des vecteurs de la base canonique. Pour mesurer l'erreur de notre estimation $\widehat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$, on utilise trois différentes distances : celle[2] de variation totale, de Hellinger, et de $\mathbb{L}^2$

$$
\begin{aligned}
d_{\mathsf{TV}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &:= {}^1\!/_2 \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1, \\
d_{\mathsf{H}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &:= {}^1\!/_{\sqrt{2}} \|\widehat{\boldsymbol{\theta}}^{1/2} - \boldsymbol{\theta}^{1/2}\|_2, \\
d_{\mathbb{L}^2}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &:= \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2.
\end{aligned}
$$

On s'intéresse à l'erreur minimax définie par

$$
\mathfrak{R}_{\square}(n, k, \varepsilon) := \inf_{\bar{\boldsymbol{\theta}}_n} \sup_{P_n \in \mathcal{M}_n(\varepsilon, \theta)} \mathbf{E}[d_{\square}(\bar{\boldsymbol{\theta}}_n, \boldsymbol{\theta})],
$$

où *inf* porte sur tous les estimateurs $\bar{\boldsymbol{\theta}}_n$ construits à partir des observations $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ et *sup* porte sur toutes les distributions jointes $\boldsymbol{P}_n$ de $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ déterminées par notre modèle de contamination $\mathcal{M}_n(\varepsilon, \theta)$. L'indice $\square$ de $\mathfrak{R}$ ci-dessus fait référence à la distance employée dans le risque, et donc $\square$ est TV, H, ou $\mathbb{L}^2$.

$\mathcal{M}_n(\varepsilon, \theta)$ détermine la nature des outliers dans un modèle où la loi de référence est $\boldsymbol{P}_{\boldsymbol{\theta}}$ et les outliers constituent une $\varepsilon$ fraction des données. Le modèle de contamination le plus général est le *modèle adversarial* dans lequel un adversaire omniscient observe les variables initiales $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ et remplace $n\varepsilon$ d'entre eux par des valeurs arbitraires. Le modèle de contamination le plus restrictif dans la littérature est celui de *Huber* dans lequel les observations sont des variables iid issues de la loi de mélange suivante

$$
(1-\varepsilon)\boldsymbol{P}_{\theta} + \varepsilon \boldsymbol{Q},
$$

où $\boldsymbol{Q}$ désigne la distribution des outliers. On détaille les différents modèles de contamination dans la section 2.2.

[1] Sans perte de généralité, on suppose que $n\varepsilon$ est un entier.

[2] On écrit $\|\boldsymbol{u}\|_q = (\sum_{j=1}^k |u_j|^q)^{1/q}$ et $\boldsymbol{u}^q = (u_1^q, \dots, u_k^q)$ pour tout $\boldsymbol{u} \in \mathbb{R}_+^k$ et $q > 0$.

Concentrons maintenant sur la distance euclidienne. On sait que la moyenne empirique

$$\bar{\boldsymbol{X}}_n := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$$

est l'estimateur minimax quand il n'y a pas de donnée aberrante. On va vérifier si cet estimateur reste robuste lors que des outliers contaminent les données. Nous avons

$$\begin{aligned} d_{\mathbb{L}^2}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*) &= \left\|\bar{\boldsymbol{X}}_n - \boldsymbol{\theta}^*\right\|_2 = \left\|\bar{\boldsymbol{Y}}_n - \boldsymbol{\theta}^* + \frac{1}{n}\sum_{i\in O}(\boldsymbol{X}_i - \boldsymbol{Y}_i)\right\|_2 \\ &\leq \left\|\bar{\boldsymbol{Y}}_n - \boldsymbol{\theta}^*\right\|_2 + \frac{|O|}{n}\sup_{\boldsymbol{x},\boldsymbol{y}\in E_k}\|\boldsymbol{x}-\boldsymbol{y}\|_2 \\ &= \left\|\bar{\boldsymbol{Y}}_n - \boldsymbol{\theta}^*\right\|_2 + \sqrt{2}\varepsilon. \end{aligned}$$

On voit que l'erreur dans le cas contaminé est majorée par l'erreur dans le cas non-contaminé plus un terme de l'ordre de $\varepsilon$. Ceci montre que la contamination a au plus un impact de l'ordre de $\varepsilon$ sur notre erreur d'estimation. Remarquons que cette relation est valable sous n'importe quel modèle de contamination. Par ailleurs, on peut montrer qu'il y a une borne inférieure de même ordre de grandeur pour l'erreur en question dans le modèle de Huber, *i.e.* dans le modèle le plus restrictif de la littérature. Cela nous aide à établir les vitesses de convergence minimax pour les différents modèles de contaminations. En effet, dans le chapitre 2, en généralisant cet argument, on prouve que

$$\begin{aligned} \mathfrak{R}_{\mathsf{TV}}(n,k,\varepsilon) &\asymp (k/n)^{1/2} + \varepsilon, \\ \mathfrak{R}_{\mathsf{H}}(n,k,\varepsilon) &\asymp (k/n)^{1/2} + \varepsilon^{1/2}, \\ \mathfrak{R}_{\mathbb{L}^2}(n,k,\varepsilon) &\asymp (1/n)^{1/2} + \varepsilon. \end{aligned}$$

C'est curieux de voir que toutes ces vitesses de convergence sont distinctes. En fait, il s'avère que l'erreur minimax n'est pas détériorée si la proportion des outliers est plus petite que $(k/n)^{1/2}$ pour la distance de variation totale, $k/n$ pour la distance de Hellinger, et $(1/n)^{1/2}$ pour la distance euclidienne. De plus, on en déduit que la moyenne empirique est l'estimateur minimax pour les trois distances.

Dans le même chapitre, on étudie la relation entre les différents modèles de contamination, et on généralise ces résultats pour toutes les distributions à support dans le simplexe de dimension $k$, soit $\Delta^{k-1} = \{\boldsymbol{v} \in \mathbb{R}_+^k : v_1 + \ldots + v_k = 1\}$. En outre, on y établit des régions de confiance qui rétrécissent à vitesses minimax.

## 5.3 Variables gaussiennes

Le cas des variables à support dans un simplexe est simple tandis que si la loi de référence est à support infini, il est un peu délicat de trouver un estimateur robuste contre les outliers.

La raison est que la présence d'un seul outlier avec une valeur arbitrairement grande peut dévier largement un estimateur tel que la moyenne empirique. Dans cette partie, on étudie le problème de l'estimation de la moyenne d'une loi normale à partir des observations contaminées par des outliers sous le modèle adversarial. On commence par le cas uni-dimensionnel et puis on aborde la version multidimensionnelle du problème.

### 5.3.1 Problème en dimension 1

On garde le cadre introduit dans la section 5.1, et on suppose désormais que la loi de référence est $\mathcal{N}(\boldsymbol{\mu}, \sigma^2)$. On regarde brièvement deux approches possibles pour estimer $\boldsymbol{\mu}$ de manière robuste. L'erreur d'estimation est mesurée par la valeur absolue.

#### Filtrage

Comme nous avons évoqué le problème avec la moyenne empirique est sa sensibilité aux outliers extrêmement éloignés de $\boldsymbol{\mu}$. Pour régler ce problème, on peut calculer la moyenne empirique sans considérer les plus petits et les plus grands points. En fait, on trie les observations $\boldsymbol{X}_{(1)} \leq \cdots \leq \boldsymbol{X}_{(n)}$ et on retourne $\widehat{\boldsymbol{\mu}}_F := \sum_{i \in F} \boldsymbol{X}_i / |F|$ où $F = \{i \in \mathcal{I} \cup \mathcal{O} | \boldsymbol{X}_{(2n\varepsilon)} < \boldsymbol{X}_i < \boldsymbol{X}_{(n-2n\varepsilon)}\}$. Même s'il y a des outliers parmi les échantillons après ce filtrage, on peut être sûr qu'ils sont au plus aussi éloignés que $\boldsymbol{Y}_{(n\varepsilon)}$ et $\boldsymbol{Y}_{(n-n\varepsilon)}$ (où les échantillons initiaux sont triés : $\boldsymbol{Y}_{(1)} \leq \cdots \leq \boldsymbol{Y}_{(n)}$). Ainsi, cet estimateur, appelé la *moyenne tronquée*, n'est pas tant influencé par les outliers avec des valeurs extrêmes. Regarder la figure 5.1.

Dans la section 1.2.1, on donne des clés de preuve pour montrer que $\widehat{\boldsymbol{\mu}}_F$ estime $\boldsymbol{\mu}$ avec une erreur de l'ordre de

$$\frac{\sigma}{\sqrt{n}} + \sigma\varepsilon\sqrt{\log(1/\varepsilon)}$$

avec grande probabilité, si $\varepsilon < 1/4$.

#### Moyenne en tant que le centre de symétrie

La loi normale est une loi symétrique, et la moyenne n'est pas seulement le barycentre de la distribution, mais elle est également le centre de symétrie. Ainsi, une autre idée pour estimer la moyenne serait d'estimer le centre de symétrie. Il y a de différentes notions de symétrie. La notion que l'on adopte ici c'est celle induite par la masse : il existe la même quantité de masse dans les deux côtés du centre. Plus précisément, dans cette symétrie, le centre est la médiane. Donc, on peut estimer le centre, *i.e.* la moyenne, par la médiane de l'échantillon, qui en fait le centre de la symétrie induite par la masse empirique.

Dans la section 1.2.2, on prouve que la médiane de l'échantillon, sous une condition non

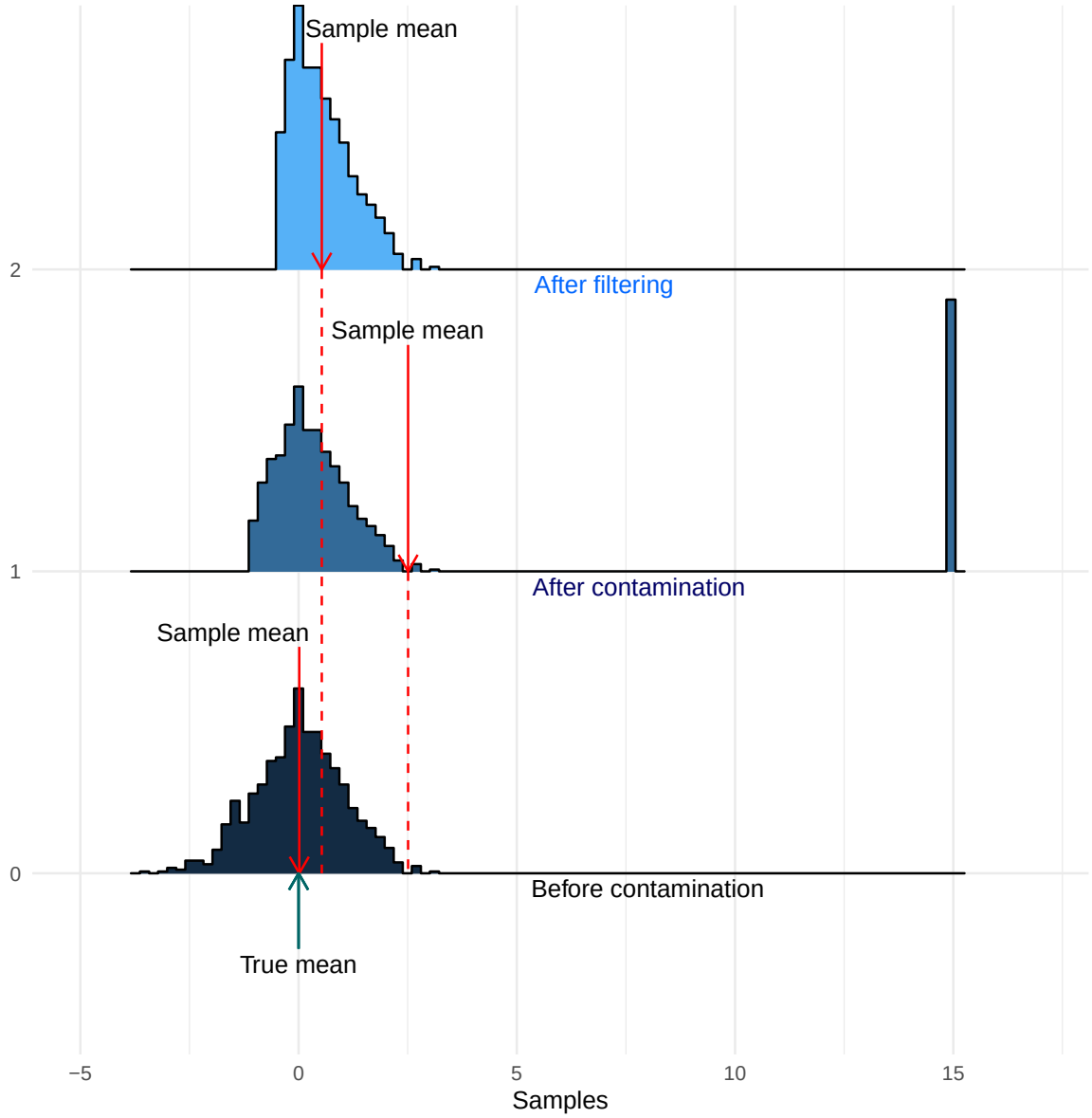


Figure 5.1: Au premier étage, l'histogramme des échantillons initiaux $(\boldsymbol{Y}_i)_{i\in\mathcal{S}}$ est affiché. Au deuxième étage, nous avons l'histogramme des échantillons contaminés $(\boldsymbol{X}_i)_{i\in\mathcal{S}}$ où les $n\varepsilon$ plus petits échantillons sont remplacées par $n\varepsilon$ échantillons de valeur 15. Au troisième étage, on voit l'histogramme des échantillons après le filtrage où les $2n\varepsilon$ plus grands et les $2n\varepsilon$ plus petits échantillons parmi $(\boldsymbol{X}_i)_{i\in\mathcal{S}}$ sont mis de côté. La moyenne empirique pour chaque jeu de données est marquée par des lignes rouges. L'impact du filtrage sur la performance de la moyenne empirique est visible.

très restrictive, estime $\boldsymbol{\mu}$ avec une erreur de l'ordre de

$$\frac{\sigma}{\sqrt{n}} + \sigma\varepsilon$$

avec grande probabilité. Par ailleurs, Chen et al. (2018) montre une borne inférieure du même ordre pour l'erreur minimax, ce qui implique que la médiane de l'échantillon est un estimateur minimax pour ce problème. Pour plus de détails sur cette borne inférieure regarder la section 1.2.4.

À part les deux méthodes mentionnées ci-dessus, il y a aussi la méthode de tournoi qui consiste à sélectionner parmi un ensemble fini de fonctions de densité, celle qui s'adapte mieux aux observations. On aborde cette méthode dans la section 1.2.3.

### 5.3.2 Problème en grande dimension

On suppose que nos données vivent dans $\mathbb{R}^p$ et que les échantillons initiaux $(\boldsymbol{Y}_i)_{i\in\{1,\dots,n\}}$ suivent la loi normale multidimensionnelle $\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_p)$. Tout comme le cas précédant, nous observons $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$, une version $\varepsilon$-contaminée des données sous le modèle adversarial. On quantifie l'erreur d'estimation par la distance euclidienne $\|.\|_2$.

Une solution naïve pour le cas multidimensionnel serait d'estimer séparément chaque coordonnée de $\boldsymbol{\mu}$ par une méthode robuste unidimensionnelle. Par la borne d'union, on pourrait prouver qu'en appliquant par exemple la médiane de l'échantillon à chaque coordonnée, on obtient un estimateur avec une erreur de l'ordre de

$$\sigma\sqrt{\frac{p\log(p)}{n}} + \sigma\varepsilon\sqrt{p}$$

avec grande probabilité. Le problème lié à ce genre d'estimateurs est que le terme $\varepsilon\sqrt{p}$ pourrait être problématique lorsque la valeur de $p$ est grande. Chen et al. (2018) montre que sous le modèle de contamination de Huber, l'erreur de la médiane coordonnée par coordonnée ne peut pas être meilleure que $\sigma(\sqrt{p/n} + \varepsilon\sqrt{p})$ quand la distribution de contamination $Q$ est définie comme un Dirac sur le point $\boldsymbol{\mu} + \sigma(1, \dots, 1)^\top$. La même contamination peut être utilisée pour montrer que les autres méthodes robustes pour dimension 1 n'auront pas de meilleur comportement en grande dimension si on les applique coordonnée par coordonnée. Maintenant, la question est si on peut avoir une meilleure dépendance en $p$ dans l'expression de l'erreur de notre estimation.

#### Filtrage (cas multidimensionnel)

Dans le cas multidimensionnel, il n'y a pas de notion d'ordre. Par conséquent, ce n'est pas possible d'employer la méthode de filtrage uni-dimensionnelle proposée ci-dessus. Pourtant, on sait que les échantillons gaussiens multidimensionnels se concentrent sur une sphère autour de la moyenne de rayon $\sigma\sqrt{p}$ (remarquer le contraste avec le cas unidimensionnel où

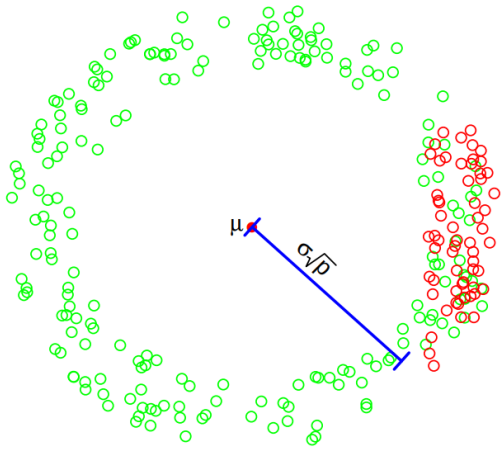


Figure 5.2: En grande dimension, les échantillons se concentrent sur une sphère de rayon $\sigma\sqrt{p}$ autour de la moyenne. Ici, les points rouges sont les outliers et les points verts sont les bons échantillons. Les simples outils de filtrage ne peuvent pas détecter ces outliers, et cette contamination dévie la moyenne empirique par un terme de l'ordre de $\varepsilon\sqrt{p}$.

les échantillons gaussiens se concentrent près de la moyenne). Plus précisément, il existe $c > 0$ tel que pour tout $t > 0$

$$\mathbf{P}\big(\big|\|\boldsymbol{Y}_i - \boldsymbol{\mu}\|_2 - \sigma\sqrt{p}\big| > t\big) \leq 2\exp(-ct^2/\sigma^2),$$

(regarder *e.g.*, (Vershynin, 2018, Theorem 3.1.1)).

Cette propriété nous aide à construire des procédures de filtrage capables de détecter les échantillons extrêmement éloignés de la moyenne. En voir un exemple dans la section 1.3.1. Le problème est que ce genre de filtrages ne peuvent enlever que les outliers extrêmes. Dans ce cas, les outliers peuvent se concentrer sur une même direction dans la sphère indiquée plus haut sans être filtrés, et après dévier la moyenne empirique par un terme de l'ordre de $\varepsilon\sqrt{p}$. Regarder la figure 5.2. En conséquence, on voit qu'un filtrage simple ne nous permet pas non plus de nous débarrasser du terme $\varepsilon\sqrt{p}$ dans l'expression de l'erreur de notre estimation. Pourtant, comme on verra c'est un outil utile.

#### La moyenne en tant que le centre de symétrie (cas multidimensionnel)

Pour généraliser la notion de la médiane de l'échantillon au cas multivarié, il faut adopter une notion de symétrie. Les différentes notions de symétrie donnent lieu à de différentes définitions de la médiane. Comme on analyse dans la section 1.3.2, la notion de la *symétrie centrale* définit la moyenne empirique comme la médiane de l'échantillon, ce qui n'est pas un estimateur robuste de la moyenne ; la notion de la *symétrie angulaire* définit la médiane géométrique, et ce dernier a le même défaut que les estimateurs précédents, à savoir le terme

$\varepsilon\sqrt{p}$ dans l'expression de son erreur dans le pire scénario.

En dimension 1, on a défini la médiane en tant que le centre de la symétrie induite par la masse. Cette notion de symétrie se généralise au cas multidimensionnel par la notion de la *symétrie de demi-espace*. Un vecteur aléatoire $\boldsymbol{Y}$ respecte la symétrie de demi-espace autour le centre $\boldsymbol{\theta}$ si l'inégalité $\mathbf{P}(\boldsymbol{Y} \in H) \geq 1/2$ est valide pour tout demi-espace fermé $H$ contenant $\boldsymbol{\theta}$. Autrement dit, $\boldsymbol{\theta}$ est le centre de la symétrie induite par la masse dans toute direction. Cette notion de symétrie définit la *médiane de Tukey* qui est un point appartenant à l'ensemble

$$\arg\max_{\boldsymbol{x}\in\mathbb{R}^p} \min_{\boldsymbol{v}\in\mathbb{R}^{p-1}} \sum_{i=1}^{n} \mathbb{1}(\boldsymbol{v}^\top \boldsymbol{X}_i \leq \boldsymbol{v}^\top \boldsymbol{x}).$$

Pour plus de précisions sur cette définition, regarder la section 1.3.2.

Chen et al. (2018) prouvent que sous certaines conditions, le risque de la médiane de Tukey est de l'ordre de

$$\sigma\sqrt{p/n} + \sigma\varepsilon$$

avec grande probabilité, et qu'en plus il s'agit de l'erreur optimale. C'est à dire que la médiane de Tukey est un estimateur optimal dans le sens minimax. Cependant, la médiane de Tukey n'est pas calculable en temps polynomial car son calcul nécessite le traitement de toutes les directions, soit un nombre d'opérations qui est exponentiel en $p$.

#### Utiliser la matrice de covariance

Comparant $\varepsilon\sqrt{p}$ et $\varepsilon$, on remarque qu'en grande dimension, il y a un écart significatif entre l'erreur des méthodes calculables déjà citées (telles que la médiane coordonnée par coordonnée, la médiane géométrique, le filtrage, etc.) et celle de la médiane de Tukey ou de la méthode de tournoi expliquée dans la section 1.3.3, qui ne sont pas calculables efficacement. Ce phénomène a stimulé beaucoup de recherches afin de trouver un compromis entre la précision statistique et l'efficacité de calcul. Le point commun entre presque toutes les méthodes proposant un tel compromis, c'est qu'elles se servent de la connaissance de la matrice de covariance de la loi de référence.

On suppose désormais que la loi de référence est $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Soit $\widehat{\boldsymbol{\Sigma}}$ la matrice de covariance empirique. En nous appuyant sur notre connaissance de $\boldsymbol{\Sigma}$, on essaie de filtrer ou pondérer les observations de sorte que $\widehat{\boldsymbol{\Sigma}}$ s'approche de $\boldsymbol{\Sigma}$. En effet, on se sert ici d'une propriété essentielle : $\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|_2$ est contrôlé par $\varepsilon\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\|_{\mathrm{op}}$ où $\widehat{\boldsymbol{\mu}}$ et $\widehat{\boldsymbol{\Sigma}}$ sont respectivement la moyenne empirique et la matrice de covariance empirique des observations filtrées ou pondérées. Ainsi, notre but consiste à minimiser $\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\|_{\mathrm{op}}$ en filtrant ou pondérant les observations.

La méthode que nous construisons dans le chapitre 3, appelée *Spectral Dimension Reduction* (SDR), appartient à ce paradigme des méthodes employant la matrice de covariance

de la loi de référence. Il s'agit d'une méthode qui filtre les observations de manière itérative où dans chaque itération elle réduit la dimension en regardant le spectre de la matrice $\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$. Elle est basée sur l'observation qu'il y a des directions dans lesquelles l'impact de la contamination n'est pas significatif. Autrement dit, au lieu de regarder toutes les directions comme c'est nécessaire pour le calcul de la médiane de Tukey, on se focalise sur les directions les plus impactées par la contamination.

En fait, nous pouvons partitionner les vecteurs propres de la matrice $\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$ en deux groupes : les vecteurs propres supérieurs (correspondant aux grandes valeurs propres) et les vecteurs propres inférieurs (correspondant aux petites valeurs propres). Puis, nous projetons orthogonalement les échantillons sur les sous-espaces engendrés par chaque groupe. Il s'avère que la moyenne empirique est un estimateur robuste dans le sous-espace engendré par les vecteurs propres inférieurs. Donc, nous pouvons nous concentrer sur le sous-espace engendré par les vecteurs propres supérieurs. Sur ce sous-espace, nous répétons récursivement la même démarche. L'idée de la réduction itérative de dimension pour l'estimation robuste est originellement introduite dans (Lai et al., 2016). Cette procédure peut se schématiser dans l'algorithme "SDR" suivant :

- Si $p = 1$ on retourne la médiane de l'échantillon
- Sinon :
    1. Étape de filtrage : on supprime des échantillons afin d'avoir tous les échantillons dans une distance de l'ordre de $\sqrt{p}$ de $\boldsymbol{\mu}$
    2. Soit $V$ le sous-espace engendré par les $p/2$ vecteurs propres supérieurs de $\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$
    3. On projette orthogonalement les échantillons sur $V$ et $V^\top$
    4. Sur $V^\top$ on applique la moyenne empirique
    5. Sur $V$ on applique récursivement SDR

La version détaillée de cet algorithme se trouve dans le chapitre 3. On y prouve que le risque de SDR est de l'ordre de

$$\sqrt{\log p}\Big(\sqrt{\mathsf{Tr}(\mathbf{\Sigma})/n} + \|\mathbf{\Sigma}\|_{\mathrm{op}}^{1/2}\varepsilon\sqrt{\log(1/\varepsilon)}\Big)$$

avec grande probabilité. Donc, SDR est une méthode calculable en temps polynomial avec une erreur presque optimale (optimale modulo des facteurs logarithmiques). Cette méthode peut être également généralisée pour toutes les distributions sous-gaussiennes.

L'avantage de la méthode SDR, ce qu'elle est rapide, qu'elle ne requière pas la connaissance de la proportion de contamination ($\varepsilon$), et que son breakdown point vaut 1/2 (breakdown point est la proportion maximale des outliers pour laquelle le risque de notre estimateur reste borné). De plus, on montre à la fin que SDR est adaptable au cas où notre connaissance de $\mathbf{\Sigma}$ est partielle.

## 5.4 Organisation du manuscrit

Ce manuscrit est constitué de quatre chapitres et deux appendices. Le chapitre 1 est une introduction sur l'estimation robuste de la moyenne gaussienne. On y présent les approches générales et on prouve informellement quelques résultats folklores du domaine. À la fin du chapitre, on expose les contributions principales de ce travail. Le chapitre 2 aborde le problème de l'estimation robuste des distributions à support dans le simplexe de probabilité. Dans ce chapitre, on étudie aussi les différents modèles de contamination. Cette partie du manuscrit est publiée dans (Bateni and Dalalyan, 2020). Dans le chapitre 3, on propose une méthode calculatoirement efficace avec une erreur presque optimale pour estimer la moyenne d'une loi gaussienne sous le modèle adversarial. Cette partie du travail se trouve dans (Bateni et al., 2022). On finit avec le chapitre 4 où on présente une conclusion et une perspective sur les directions possibles pour les prochains travaux. Les deux appendices contiennent les preuves formelles des résultats exposés dans les chapitres 2 et 3.

# Appendix A

# Proofs for Chapter 2



## A.1 Proofs of propositions

*Proof of Proposition 3 on page 41.* Recall that $\widehat{O}$ is the set of outliers in the Huber model. Let $O$ be any subset of $\{1, \dots, n\}$. It follows from the definition of Huber's model that if $\boldsymbol{P}_n^O$ stands for the conditional distribution of $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ given $\widehat{O} = O$, when $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ is drawn from $\boldsymbol{P}^n \in \mathcal{M}_n^{\mathsf{HC}}(\varepsilon, \boldsymbol{\theta}^*)$, then $\boldsymbol{P}_n^O \in \mathcal{M}_n^{\mathsf{HDC}}(|O|/n, \boldsymbol{\theta}^*)$. Therefore, for every $O$ of cardinality $o \geq 2\varepsilon n$, we have

$$\begin{aligned}\sup_{\mathcal{M}_n^{\mathsf{HC}}(\varepsilon,\boldsymbol{\theta}^*)} \mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}^*)\mathbb{1}(\widehat{O} = O)] &= \sup_{\mathcal{M}_n^{\mathsf{HC}}(\varepsilon,\boldsymbol{\theta}^*)} \mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}^*)|\widehat{O} = O]\mathbf{P}(\widehat{O} = O) \\ &\leq \sup_{\mathcal{M}_n^{\mathsf{HDC}}(o/n,\boldsymbol{\theta}^*)} \mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}^*)]\mathbf{P}(\widehat{O} = O) \qquad \text{(A.1)} \\ &\overset{(1)}{\leq} \sup_{\mathcal{M}_n^{\mathsf{HDC}}(1,\boldsymbol{\theta}^*)} \mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}^*)]\mathbf{P}(\widehat{O} = O).\end{aligned}$$

Inequality (1) above is a direct consequence of the inclusion $\mathcal{M}_n^{\mathsf{HDC}}(o/n, \boldsymbol{\theta}^*) \subset \mathcal{M}_n^{\mathsf{HDC}}(1, \boldsymbol{\theta}^*)$. Summing the obtained inequality over all sets $O$ of cardinality $\geq 2\varepsilon n$, we get

$$\sup_{\mathcal{M}_n^{\mathsf{HC}}(\varepsilon,\boldsymbol{\theta}^*)} \mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}^*)\mathbb{1}(|\widehat{O}| \geq 2\varepsilon n)] \leq \sup_{\mathcal{M}_n^{\mathsf{HDC}}(1,\boldsymbol{\theta}^*)} \mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}^*)]\mathbf{P}(|\widehat{O}| \geq 2\varepsilon n).$$

It follows from the multiplicative form of Chernoff's inequality that $\mathbf{P}(|\widehat{O}| \geq 2\varepsilon n) \leq e^{-n\varepsilon/3}$. This leads to the last term in inequality (2.3).

Using the same argument as for (A.1), for any $O$ of cardinality $o < 2n\varepsilon$, we get

$$\begin{aligned}\sup_{\mathcal{M}_n^{\mathsf{HC}}(\varepsilon,\boldsymbol{\theta}^*)} \mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*)\mathbb{1}(|\widehat{O}| < 2n\varepsilon)] &\leq \sup_{\mathcal{M}_n^{\mathsf{HDC}}(2\varepsilon,\boldsymbol{\theta}^*)} \mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*)] \sum_{|O|\leq 2n\varepsilon} \mathbf{P}(\widehat{O} = O) \\ &= \sup_{\mathcal{M}_n^{\mathsf{HDC}}(2\varepsilon,\boldsymbol{\theta}^*)} \mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*)].\end{aligned}$$

This completes the proof of (2.3).

One can use the same arguments along with the Tchebychev inequality to establish (2.4). Indeed, for every $S$ of cardinality $o \leq 2\varepsilon n$, we have

$$\begin{aligned}&\sup_{\mathcal{M}_n^{\mathsf{HC}}(\varepsilon,\boldsymbol{\theta}^*)} r\mathbf{P}\big(d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*) > r \text{ and } \widehat{O} = O\big) \\ &\qquad = \sup_{\mathcal{M}_n^{\mathsf{HC}}(\varepsilon,\boldsymbol{\theta}^*)} r\mathbf{P}\big(d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*) > r \,|\, \widehat{O} = O\big)\mathbf{P}(\widehat{O} = O) \\ &\qquad \leq \sup_{\mathcal{M}_n^{\mathsf{HDC}}(o/n,\boldsymbol{\theta}^*)} r\mathbf{P}\big(d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*) > r\big)\mathbf{P}(\widehat{O} = O) \\ &\qquad \leq \sup_{\mathcal{M}_n^{\mathsf{HDC}}(2\varepsilon,\boldsymbol{\theta}^*)} \mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*)]\mathbf{P}(\widehat{O} = O).\end{aligned}$$

Summing the obtained inequality over all sets $O$ of cardinality $o \leq 2\varepsilon n$, we get

$$\begin{aligned}\sup_{\mathcal{M}_n^{\mathsf{HC}}(\varepsilon,\boldsymbol{\theta}^*)} r\mathbf{P}\big(d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*) > r \text{ and } |\widehat{O}| \leq 2\varepsilon n\big) &\leq \sup_{\mathcal{M}_n^{\mathsf{HDC}}(2\varepsilon,\boldsymbol{\theta}^*)} \mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*)] \\ &= R_d^{\mathsf{HDC}}(n, 2\varepsilon, \boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}).\end{aligned}$$

On the other hand, it holds that

$$\sup_{\mathcal{M}_n^{\mathsf{HC}}(\varepsilon,\boldsymbol{\theta}^*)} r\mathbf{P}\big(d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*) > r \text{ and } |\widehat{O}| > 2\varepsilon n\big) \leq r\mathbf{P}\big(|\widehat{O}| > 2\varepsilon n\big) \leq re^{-n\varepsilon/3},$$

and the claim of the proposition follows. □

*Proof of Proposition 4 on page 43.* Let $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ be two points in $\Theta$. We have

$$\begin{aligned}2R_d^{\mathsf{HC}}(n,\varepsilon,\Theta,\widehat{\boldsymbol{\theta}}_n) &\geq R_d^{\mathsf{HC}}(n,\varepsilon,\boldsymbol{\theta}_1,\widehat{\boldsymbol{\theta}}_n) + R_d^{\mathsf{HC}}(n,\varepsilon,\boldsymbol{\theta}_2,\widehat{\boldsymbol{\theta}}_n) \\ &\geq \mathbf{E}_{(1-\varepsilon)\boldsymbol{P}_{\boldsymbol{\theta}_1}+\varepsilon\boldsymbol{P}_{\boldsymbol{\theta}_2}}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}_1)] + \mathbf{E}_{(1-\varepsilon)\boldsymbol{P}_{\boldsymbol{\theta}_1}+\varepsilon\boldsymbol{P}_{\boldsymbol{\theta}_2}}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}_2)].\end{aligned}$$

To ease writing, assume that $n$ is an even number. Let $O$ be any fixed set of cardinality $n/2$. It is clear that the set of outliers $\widehat{O}$ satisfies

$$p_O = \boldsymbol{P}(\widehat{O} = O) = \boldsymbol{P}(\widehat{O} = O^c) > 0.$$

Furthermore, if $\boldsymbol{X}_{1:n}$ is drawn from $((1-\varepsilon)\boldsymbol{P}_{\boldsymbol{\theta}_1} + \varepsilon \boldsymbol{P}_{\boldsymbol{\theta}_2})^{\otimes n} := \boldsymbol{P}_{\varepsilon}^{\otimes n}$, then its conditional distribution given $\widehat{O} = O$ is exactly the same as the conditional distribution of $\boldsymbol{X}_{1:n} \sim \boldsymbol{P}_{\varepsilon}^{\otimes n}$ given $\widehat{O} = O^c$. This implies that

$$\begin{aligned}
&2R_d^{\mathsf{HC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) \\
&\quad \geq p_O\big(\mathbf{E}_{\boldsymbol{P}_\varepsilon}[d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_1)|\widehat{O} = O] + \mathbf{E}_{\boldsymbol{P}_\varepsilon}[d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_2)|\widehat{O} = O^c]\big) \\
&\quad = p_O \mathbf{E}_{\boldsymbol{P}_\varepsilon}[d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_1) + d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_2)|\widehat{O} = S] \geq p_O d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2),
\end{aligned}$$

where in the last step we have used the triangle inequality. The obtained inequality being true for every $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, we can take the supremum to get

$$R_d^{\mathsf{HC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) \geq (p_O/2) \sup_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta} d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = +\infty.$$

This completes the proof. □

## A.2 Minimax upper bounds over the sparse simplex

This section is devoted to the proof of the upper bounds on minimax risks in the discrete model with respect to various distances.

*Proof of Theorem 7 on page 46.* To ease notation, we set

$$\bar{\boldsymbol{X}}_n = \frac{1}{n}\sum_i \boldsymbol{X}_i, \quad \bar{\boldsymbol{Y}}_n = \frac{1}{n}\sum_i \boldsymbol{Y}_i, \quad \bar{\boldsymbol{X}}_O = \frac{1}{o}\sum_{i \in O} \boldsymbol{X}_i, \quad \bar{\boldsymbol{Y}}_O = \frac{1}{o}\sum_{i \in O} \boldsymbol{Y}_i$$

In the adversarial model, we have $\boldsymbol{X}_i = \boldsymbol{Y}_i$ if $i \notin O$ where $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ are generated from the reference distribution $\boldsymbol{\theta}^*$.

$$\begin{aligned}
d_{\mathbb{L}^2}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*) &= \big\|\bar{\boldsymbol{X}}_n - \boldsymbol{\theta}^*\big\|_2 = \big\|\bar{\boldsymbol{Y}}_n - \boldsymbol{\theta}^* + \frac{1}{n}\sum_{i \in O}(\boldsymbol{X}_i - \boldsymbol{Y}_i)\big\|_2 \\
&\leq \big\|\bar{\boldsymbol{Y}}_n - \boldsymbol{\theta}^*\big\|_2 + \frac{|O|}{n} \sup_{\boldsymbol{x}, \boldsymbol{y} \in \Delta^{k-1}} \|\boldsymbol{x} - \boldsymbol{y}\|_2 \\
&= \big\|\bar{\boldsymbol{Y}}_n - \boldsymbol{\theta}^*\big\|_2 + \sqrt{2}\varepsilon,
\end{aligned}$$

which gives us

$$\sup_{\mathcal{M}_n^{\mathrm{AC}}(\varepsilon, \boldsymbol{\theta}^*)} \mathbf{E}[d_{\mathbb{L}^2}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*)] \leq \sup_{\boldsymbol{\theta}^*} \mathbf{E}[d_{\mathbb{L}^2}(\bar{\boldsymbol{Y}}_n, \boldsymbol{\theta}^*)] + \sqrt{2}\varepsilon.$$

And for a fixed $\boldsymbol{\theta}^*$ it is well known that

$$\begin{aligned}\mathbf{E}[d^2_{\mathbb{L}^2}(\bar{\boldsymbol{Y}}_n, \boldsymbol{\theta}^*)] &= \sum_{j=1}^k \mathbf{Var}[\bar{\boldsymbol{Y}}_{n,j}] = \sum_{j=1}^k \mathbf{Var}\left[\frac{1}{n}\sum_{i=1}^n \mathbb{1}(Y_i = \boldsymbol{e}_j)\right] \\ &= \frac{1}{n}\sum_{j=1}^k \mathbf{Var}[\mathbb{1}(Y_1 = \boldsymbol{e}_j)] \le \frac{1}{n}\sum_{j=1}^k \mathbf{E}[\mathbb{1}(Y_1 = \boldsymbol{e}_j)] = \frac{1}{n}.\end{aligned}$$

Hence, we obtain $R^{\mathrm{AC}}_{\mathbb{L}^2}(n, \varepsilon, \Delta^{k-1}) \le (1/n)^{1/2} + \varepsilon$. Similarly,

$$\begin{aligned}d_{\mathrm{TV}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*) &\le \left\|\bar{\boldsymbol{Y}}_n - \boldsymbol{\theta}^*\right\|_1 + \frac{|O|}{n}\sup_{\boldsymbol{x},\boldsymbol{y}\in\Delta^{k-1}} \|\boldsymbol{x} - \boldsymbol{y}\|_1 \\ &= \left\|\bar{\boldsymbol{Y}}_n - \boldsymbol{\theta}^*\right\|_1 + 2\varepsilon.\end{aligned}$$

This gives

$$\sup_{\mathcal{M}^{\mathrm{AC}}_n(\varepsilon,\boldsymbol{\theta}^*)} \mathbf{E}[d_{\mathrm{TV}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*)] \le \sup_{\boldsymbol{\theta}^*} \mathbf{E}[d_{\mathrm{TV}}(\bar{\boldsymbol{Y}}_n, \boldsymbol{\theta}^*)] + 2\varepsilon.$$

In addition, for every $\boldsymbol{\theta}^*$,

$$\begin{aligned}\mathbf{E}[d_{\mathrm{TV}}(\bar{\boldsymbol{Y}}_n, \boldsymbol{\theta}^*)] &= \frac{1}{2}\sum_{j=1}^k \mathbf{E}\big[\big|\bar{\boldsymbol{Y}}_{n,j} - \boldsymbol{\theta}^*_j\big|\big] \\ &\le \frac{1}{2}\sum_{j=1}^k \left(\mathbf{E}\big[\big|\bar{\boldsymbol{Y}}_{n,j} - \boldsymbol{\theta}^*_j\big|^2\big]\right)^{1/2} \\ &= \frac{1}{2}\sum_{j=1}^k \left(\mathbf{Var}\big[\bar{\boldsymbol{Y}}_{n,j}\big]\right)^{1/2} \\ &\overset{(1)}{=} \frac{1}{2}\sum_{j\in J} \big(\frac{1}{n}\boldsymbol{\theta}^*_j(1 - \boldsymbol{\theta}^*_j)\big)^{1/2} \\ &\overset{(2)}{\le} \frac{1}{2}\, s^{1/2}\left(\sum_{j=1}^k \frac{1}{n}\boldsymbol{\theta}^*_j(1 - \boldsymbol{\theta}^*_j)\right)^{1/2} \le \frac{1}{2}\big(s/n\big)^{1/2},\end{aligned}$$

where in (1) we have used the notation $J = \{j : \theta^*_j \neq 0\}$ and in (2) we have used the Cauchy-Schwarz inequality. This leads to

$$R^{\mathrm{AC}}_{\mathrm{TV}}(n, \varepsilon, \Delta^{k-1}) \le (k/n)^{1/2} + \varepsilon.$$

Finally, for the Hellinger distance

$$d_{\mathrm{H}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*) \le d_{\mathrm{H}}(\bar{\boldsymbol{X}}_n, \bar{\boldsymbol{Y}}_n) + d_{\mathrm{H}}(\bar{\boldsymbol{Y}}_n, \boldsymbol{\theta}^*) \le d_{\mathrm{TV}}(\bar{\boldsymbol{X}}_n, \bar{\boldsymbol{Y}}_n)^{1/2} + d_{\mathrm{H}}(\bar{\boldsymbol{Y}}_n, \boldsymbol{\theta}^*),$$

where we have already seen that

$$d_{\mathrm{TV}}(\bar{\boldsymbol{X}}_n, \bar{\boldsymbol{Y}}_n) = \frac{o}{n}\big\|\bar{\boldsymbol{X}}_O - \bar{\boldsymbol{Y}}_O\big\|_1 \le 2\varepsilon.$$

This yields

$$\sup_{\mathcal{M}_n^{\mathrm{AC}}(\varepsilon,\boldsymbol{\theta}^*)} \mathbf{E}[d_{\mathrm{H}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*)] \le \sup_{\boldsymbol{\theta}^*} \mathbf{E}[d_{\mathrm{H}}(\bar{\boldsymbol{Y}}_n, \boldsymbol{\theta}^*)] + \sqrt{2\varepsilon}.$$

Note that $\mathbf{E}[\bar{\boldsymbol{Y}}_{n,j}] = \theta_j^*$ implies that $\bar{\boldsymbol{Y}}_{n,j} = \theta_j^* = 0$ for every $j$ that does not belong to the sparsity pattern $J$. Furthermore, for every $j \in J$, we have

$$\big|\sqrt{\bar{\boldsymbol{Y}}_{n,j}} - \sqrt{\boldsymbol{\theta}_j^*}\big| = \frac{|\bar{\boldsymbol{Y}}_{n,j} - \boldsymbol{\theta}_j^*|}{\sqrt{\bar{\boldsymbol{Y}}_{n,j}} + \sqrt{\boldsymbol{\theta}_j^*}} \le \frac{|\bar{\boldsymbol{Y}}_{n,j} - \boldsymbol{\theta}_j^*|}{\sqrt{\boldsymbol{\theta}_j^*}}.$$

This implies that for every $\boldsymbol{\theta}^* \in \Delta_s^{k-1}$,

$$\begin{aligned}\mathbf{E}[d_{\mathrm{H}}^2(\bar{\boldsymbol{Y}}_n, \boldsymbol{\theta}^*)] &= \mathbf{E}\bigg[\frac{1}{2}\sum_{j=1}^k \Big(\sqrt{\bar{\boldsymbol{Y}}_{n,j}} - \sqrt{\boldsymbol{\theta}_j^*}\Big)^2\bigg] \le \frac{1}{2}\mathbf{E}\bigg[\sum_{j\in J}\frac{(\bar{\boldsymbol{Y}}_{n,j} - \boldsymbol{\theta}_j^*)^2}{\boldsymbol{\theta}_j^*}\bigg] \\ &= \frac{1}{2}\sum_{j\in J}\frac{1}{\boldsymbol{\theta}_j^*}\mathbf{Var}\big[\bar{\boldsymbol{Y}}_{n,j}\big] = \frac{1}{2}\sum_{j\in J}\frac{1}{\boldsymbol{\theta}_j^*} \times \frac{\boldsymbol{\theta}_j^*(1-\boldsymbol{\theta}_j^*)}{n} = \frac{s-1}{2n}.\end{aligned}$$

Hence, by Jensen's inequality $\mathbf{E}[d_{\mathrm{H}}(\bar{\boldsymbol{Y}}_n, \boldsymbol{\theta}^*)] < \sqrt{s/n}$. Therefore, we infer that

$$R_{\mathrm{H}}^{\mathrm{AC}}(n, \varepsilon, \Delta_s^{k-1}) \le (s/n)^{1/2} + \sqrt{2}\varepsilon^{1/2}$$

and the last claim of the theorem follows. □

## A.3 Minimax lower bounds over the sparse simplex

This section is devoted to the proof of the lower bounds on minimax risks in the discrete model with respect to various distances. Note that the rates over the high-dimensional "sparse" simplex $\Delta_s^{k-1}$ coincide with those for the dense simplex $\Delta^{s-1}$. For this reason, all the lower bounds will be proved for $\Delta^{s-1}$ only (for $s \ge 2$ an even integer). In addition, we will restrict our attention to the distributions $\boldsymbol{P}$, $\boldsymbol{Q}$ over $\Delta^{s-1}$ that are supported by the set $\mathcal{A}$ of the elements of the canonical basis that is $\boldsymbol{P}(\mathcal{A}) = \boldsymbol{Q}(\mathcal{A}) = 1$.

*Proof of Theorem 8 on page 46.* We denote by $\boldsymbol{e}_j$ the vector in $\mathbb{R}^s$ having all the coordinates equal to zero except the $j$th coordinate which is equal to one. Setting

$$\boldsymbol{\theta} = \boldsymbol{e}_1, \quad \text{and} \quad \boldsymbol{\theta}' = \Big(1 - \frac{\varepsilon}{1-\varepsilon}\Big)\boldsymbol{e}_1 + \frac{\varepsilon}{1-\varepsilon}\boldsymbol{e}_2$$

we have

$$d_{\mathrm{TV}}(\boldsymbol{\theta},\boldsymbol{\theta}') = \frac{\varepsilon}{1-\varepsilon}, \quad d_{\mathbb{L}^2}(\boldsymbol{\theta},\boldsymbol{\theta}') \geq \sqrt{2}\varepsilon, \quad \text{and} \quad d_{\mathrm{H}}(\boldsymbol{\theta},\boldsymbol{\theta}') \geq \varepsilon^{1/2}.$$

Therefore, modulus of continuity defined by

$$w_d(\varepsilon,\Delta) = \sup\left\{d(\boldsymbol{\theta},\boldsymbol{\theta}') : \boldsymbol{\theta},\boldsymbol{\theta}' \in \Delta,\ d_{\mathrm{TV}}(\boldsymbol{\theta},\boldsymbol{\theta}') \leq \varepsilon/(1-\varepsilon)\right\}$$

for a distance $d$ and a set $\Delta$, satisfies for any $\varepsilon \leq 1/2$,

$$w_{\mathrm{TV}}(\varepsilon,\Delta^{k-1}) \geq \varepsilon, \quad w_{\mathbb{L}^2}(\varepsilon,\Delta^{k-1}) \geq \sqrt{2}\varepsilon, \quad \text{and} \quad w_{\mathrm{H}}(\varepsilon,\Delta^{k-1}) \geq \varepsilon^{1/2}.$$

These bounds on the modulus of continuity are the first ingredient we need for lower bounding the minimax risk using (Chen et al., 2018, Theorem 5.1).

The second ingredient is the minimax rate in the non-contaminated case. are well known. For each of the considered distances, this rate is well-known. However, for the sake of completeness, we provide below a proof those lower bounds using Fano's method. For this, we use the Varshamov-Gilbert lemma (see e.g. Lemma 2.9 in (Tsybakov, 2009)) and Theorem 2.5 in (Tsybakov, 2009). The Varshamov-Gilbert lemma guarantees the existence of a set $\boldsymbol{\omega}^{(1)},\ldots,\boldsymbol{\omega}^{(M)} \in \{0,1\}^{\lfloor s/2 \rfloor}$ of cardinality $M \geq 2^{s/16}$ such that

$$\rho(\boldsymbol{\omega}^{(i)},\boldsymbol{\omega}^{(j)}) \geq \frac{s}{16}, \quad \text{for all} \quad i \neq j,$$

where $\rho(.,.)$ stands for the Hamming distance. Using these binary vectors $\{\boldsymbol{\omega}_j\}$, a parameter $\beta \in [0,1/s]$ to be specified later and the "baseline" vector $\boldsymbol{\theta}^{(0)} = (1/s,\ldots,1/s)$, we define hypotheses $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}$ by the relations

$$\boldsymbol{\theta}^{(i)}_{2j-1} = \boldsymbol{\theta}^{(0)}_{2j-1} + \boldsymbol{\omega}^{(i)}_j\beta \quad \text{and} \quad \boldsymbol{\theta}^{(i)}_{2j} = \boldsymbol{\theta}^{(0)}_{2j} - \boldsymbol{\omega}^{(i)}_j\beta \quad \forall j \in \{0,\ldots,\lfloor s/2 \rfloor\}.$$

Remark that $\boldsymbol{\theta}^{(0)},\ldots,\boldsymbol{\theta}^{(M)}$ are all probability vectors of dimension $s$. Denoting the Kullback-Leibler divergence by $d_{\mathrm{KL}}(.,.)$, one can check the conditions of Theorem 2.5 in Tsybakov (2009):

$$\begin{aligned} d_{\mathbb{L}^2}(\boldsymbol{\theta}^{(i)},\boldsymbol{\theta}^{(j)}) &\geq \beta\frac{\sqrt{2s}}{4} \qquad \forall j \neq i, \\ d_{\mathrm{TV}}(\boldsymbol{\theta}^{(i)},\boldsymbol{\theta}^{(j)}) &\geq \frac{\beta s}{16} \qquad \forall j \neq i, \end{aligned}$$

as well as

$$
\begin{aligned}
d_{\mathrm{KL}}(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(0)}) &\leq \sum_{j=1}^{\lfloor s/2 \rfloor} (\boldsymbol{\theta}^{(0)}_{2j-1} + \beta) \log \frac{\boldsymbol{\theta}^{(0)}_{2j-1} + \beta}{\boldsymbol{\theta}^{(0)}_{2j-1}} + (\boldsymbol{\theta}^{(0)}_{2j} - \beta) \log \frac{\boldsymbol{\theta}^{(0)}_{2j} - \beta}{\boldsymbol{\theta}^{(0)}_{2j}} \\
&\leq \sum_{j=1}^{\lfloor k/2 \rfloor} \frac{\beta}{\boldsymbol{\theta}^{(0)}_{2j-1}} (\boldsymbol{\theta}^{(0)}_{2j-1} + \beta) - \frac{\beta}{\boldsymbol{\theta}^{(0)}_{2j}} (\boldsymbol{\theta}^{(0)}_{2j} - \beta) \\
&= \beta^2 \sum_{j=1}^{k} \frac{1}{\boldsymbol{\theta}^{(0)}_{j}} = \beta^2 \sum_{j=1}^{s} s = \beta^2 s^2 \leq \frac{\alpha \log M}{n} \qquad \forall i \in \{1, \dots, M\},
\end{aligned}
$$

for $\beta = \sqrt{\alpha/(ns)}/4$, taking into account the fact that $2^{s/16} \leq M$. Now by applying the aforementioned theorem, we obtain for the non-contaminated setting ($\varepsilon = 0$)

$$
\begin{aligned}
\inf_{\bar{\boldsymbol{\theta}}_n} \sup_{\mathcal{M}_n^{\mathrm{HC}}(0, \Delta^{s-1})} \boldsymbol{P}\left( d_{\mathbb{L}^2}(\bar{\boldsymbol{\theta}}_n, \boldsymbol{\theta}^*) \geq \frac{\beta\sqrt{2s}}{2} \right) &\geq \frac{\sqrt{M}}{1+\sqrt{M}} \left( 1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}} \right), \\
\inf_{\bar{\boldsymbol{\theta}}_n} \sup_{\mathcal{M}_n^{\mathrm{HC}}(0, \Delta^{s-1})} \boldsymbol{P}\left( d_{\mathrm{TV}}(\bar{\boldsymbol{\theta}}_n, \boldsymbol{\theta}^*) \geq \frac{\beta s}{8} \right) &\geq \frac{\sqrt{M}}{1+\sqrt{M}} \left( 1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}} \right).
\end{aligned}
$$

Setting $M = 2^{s/16}$ and $\alpha = 1/32$, by Markov's inequality, one concludes

$$
\begin{aligned}
\inf_{\bar{\boldsymbol{\theta}}_n} R^{\mathsf{HC}}_{\mathbb{L}^2}(n, 0, \Delta^{s-1}, \bar{\boldsymbol{\theta}}_n) &\geq c(1/n)^{1/2}, \\
\inf_{\bar{\boldsymbol{\theta}}_n} R^{\mathsf{HC}}_{\mathrm{TV}}(n, 0, \Delta^{s-1}, \bar{\boldsymbol{\theta}}_n) &\geq c(s/n)^{1/2},
\end{aligned}
$$

where $c = 1/25600$. Since, $\sqrt{2}\, d_{\mathrm{H}}(\boldsymbol{\theta}, \boldsymbol{\theta}') \geq d_{\mathrm{TV}}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ for any pair of points $(\boldsymbol{\theta}, \boldsymbol{\theta}')$ on the simplex, we have

$$
\inf_{\bar{\boldsymbol{\theta}}_n} R^{\mathsf{HC}}_{\mathrm{H}}(n, 0, \Delta^{s-1}, \bar{\boldsymbol{\theta}}_n) \geq c(s/n)^{1/2}.
$$

Finally, we apply (Chen et al., 2018, Theorem 5.1) stating in our case for any distance $d$

$$
\inf_{\bar{\boldsymbol{\theta}}_n} R^{\mathsf{HC}}_{d}(n, \varepsilon, \Delta^{s-1}, \bar{\boldsymbol{\theta}}_n) \geq c\big\{ \inf_{\bar{\boldsymbol{\theta}}_n} R^{\mathsf{HC}}_{d}(n, 0, \Delta^{s-1}, \bar{\boldsymbol{\theta}}_n) + w_d(\varepsilon, \Delta^{s-1}) \big\},
$$

for an universal constant $c$, which completes the proof of Theorem 8. □

## A.4 Proofs of bounds with high probability

*Proof of Theorem 9 on page 48.* Suppose $\boldsymbol{X}_i = \boldsymbol{Y}_i$ if $i \notin O$ where $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ are independently generated from the reference distribution $\boldsymbol{P}$ so that $\mathbf{E}[\boldsymbol{Y}_i] = \boldsymbol{\theta}^*$. For any $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n$, let $\boldsymbol{\Phi}_{\square}(\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n) := d_{\square}(\sum_{i=1}^n \boldsymbol{Z}_i/n, \boldsymbol{\theta}^*)$, where $\square$ refers here to the distances $\mathbb{L}^2$ or TV. Given

$\boldsymbol{Y}'_1, \dots, \boldsymbol{Y}'_n \in \Delta^{k-1}$ we have for every $i$

$$\boldsymbol{\Phi}_\square(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_i, \dots, \boldsymbol{Y}_n) - \boldsymbol{\Phi}_\square(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_{i-1}, \boldsymbol{Y}'_i, \boldsymbol{Y}_{i+1}, \dots, \boldsymbol{Y}_n) \le \frac{1}{n} d_\square(\boldsymbol{Y}_i, \boldsymbol{Y}'_i).$$

Furthermore, it can easily be shown that the last term is bounded by $\sqrt{2}/n$ and $2/n$ for the distances $\mathbb{L}^2$ and TV, respectively. By bounded difference inequality (see for example Theorem 6.2 of Boucheron et al. (2013)) with probability at least $1 - \delta$

$$\begin{aligned}
\boldsymbol{\Phi}_{\mathbb{L}^2}(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n) &\le \mathbf{E}\boldsymbol{\Phi}_{\mathbb{L}^2}(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n) + \big(\log(1/\delta)/n\big)^{1/2} \\
&\le (1/n)^{1/2} + \big(\log(1/\delta)/n\big)^{1/2}, \\
\boldsymbol{\Phi}_{TV}(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n) &\le \mathbf{E}\boldsymbol{\Phi}_{TV}(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n) + \big(\log(2/\delta)/n\big)^{1/2} \\
&\le (s/n)^{1/2} + \big(\log(2/\delta)/n\big)^{1/2}.
\end{aligned}$$

Using $\boldsymbol{\Phi}_\square(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) \le \boldsymbol{\Phi}_\square(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n) + d_\square(\sum_{i\in O} \boldsymbol{X}_i/n, \sum_{i\in O} \boldsymbol{Y}_i/n)$, one can conclude the proof of the first two claims of the theorem.

For the Hellinger distance, the computations are more tedious. We have to separate the case of small $\theta^*_j$. To this end, let $J = \{j : 0 < \theta^*_j < (1/n)\log(2s/\delta)\}$ and $J' = \{j : \theta^*_j \ge (1/n)\log(2s/\delta)\}$. We have

$$\begin{aligned}
\sum_{j\in J} \Big(\sqrt{\bar{\boldsymbol{Y}}_{n,j}} - \sqrt{\theta^*_j}\Big)^2 &\le \sum_{j\in J} (\bar{\boldsymbol{Y}}_{n,j} + \theta^*_j) \\
&\le \frac{1}{n}\sum_{i=1}^n \Big(\sum_{j\in J} \boldsymbol{Y}_{i,j} - \theta^*_j\Big) + \frac{2s\log(2s/\delta)}{n}.
\end{aligned}$$

Since the random variables $U_i := \Big(\sum_{j\in J} \boldsymbol{Y}_{i,j}\Big)$ are iid, positive, bounded by $1$, the Bernstein inequality implies that

$$\frac{1}{n}\sum_{i=1}^n \big(U_i - \mathbf{E}[U_1]\big) \le \sqrt{\frac{2\mathsf{Var}(U_1)\log(2/\delta)}{n}} + \frac{\log(2/\delta)}{3n},$$

holds with probability at least $1 - \delta/2$ for $0 < \delta < 1$. One easily checks that $\sqrt{\mathsf{Var}(U_1)} \le \sum_{j\in J}\sqrt{\mathsf{Var}(\boldsymbol{Y}_{1,j})} \le s\sqrt{\log(2s/\delta)/n}$. Therefore, with probability at least $1 - \delta/2$, we have

$$\frac{1}{n}\sum_{i=1}^n \Big(\sum_{j\in J} \boldsymbol{Y}_{i,j} - \theta^*_j\Big) \le \frac{\sqrt{2}\, s\log(2s/\delta)}{n} + \frac{\log(2/\delta)}{3n}.$$

This yields

$$\sum_{j\in J} \Big(\sqrt{\bar{\boldsymbol{Y}}_{n,j}} - \sqrt{\theta^*_j}\Big)^2 \le \frac{3.5s\log(2s/\delta) + \log(2/\delta)}{n}, \tag{A.2}$$

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

On the other hand, we have

$$\sum_{j\in J'}\left(\sqrt{\bar{\boldsymbol{Y}}_{n,j}}-\sqrt{\theta^*_j}\right)^2 \le \sum_{j\in J'}\frac{(\bar{\boldsymbol{Y}}_{n,j}-\theta^*_j)^2}{\theta^*_j} \le s\max_{j\in J'}\frac{(\bar{\boldsymbol{Y}}_{n,j}-\theta^*_j)^2}{\theta^*_j}. \tag{A.3}$$

The Bernstein inequality and the union bound imply that, with probability at least $1-\delta/2$, for all $j\in J'$,

$$\begin{aligned}|\bar{\boldsymbol{Y}}_{n,j}-\theta^*_j| &\le \sqrt{\frac{2\mathrm{Var}(\boldsymbol{Y}_{1,j})\log(2s/\delta)}{n}}+\frac{\log(2s/\delta)}{n}\\ &\le \sqrt{\frac{2\theta^*_j\log(2s/\delta)}{n}}+\frac{\log(2s/\delta)}{n} \le 2.5\sqrt{\frac{\theta^*_j\log(2s/\delta)}{n}}.\end{aligned} \tag{A.4}$$

Combining (A.3) and (A.4), we obtain

$$\sum_{j\in J'}\left(\sqrt{\bar{\boldsymbol{Y}}_{n,j}}-\sqrt{\theta^*_j}\right)^2 \le 2.5^2\frac{s\log(2s/\delta)}{n}, \tag{A.5}$$

with probability at least $1-\delta/2$. Finally, inequalities (A.2) and (A.5) together lead to

$$d^2_{\mathsf{H}}(\bar{\boldsymbol{Y}}_n,\boldsymbol{\theta}^*)=\frac{1}{2}\sum_{j=1}^{n}\left(\sqrt{\bar{\boldsymbol{Y}}_{n,j}}-\sqrt{\theta^*_j}\right)^2 \le \frac{5s\,\log(2s/\delta)}{n}+\frac{\log(2/\delta)}{2n},$$

which is true with probability at least $1-\delta$. Using the triangle inequality and the fact that the Hellinger distance is smaller than the square root of the TV-distance, we get

$$\begin{aligned}d_{\mathsf{H}}(\bar{\boldsymbol{X}}_n,\boldsymbol{\theta}^*) &\le d_{\mathsf{H}}(\bar{\boldsymbol{X}}_n,\bar{\boldsymbol{Y}}_n)+d_{\mathsf{H}}(\bar{\boldsymbol{Y}}_n,\boldsymbol{\theta}^*)\\ &\le \sqrt{d_{\mathsf{TV}}(\bar{\boldsymbol{X}}_n,\bar{\boldsymbol{Y}}_n)}+d_{\mathsf{H}}(\bar{\boldsymbol{Y}}_n,\boldsymbol{\theta}^*)\\ &\le \varepsilon^{1/2}+\sqrt{\frac{5s\,\log(2s/\delta)}{n}}+\sqrt{\frac{\log(2/\delta)}{2n}},\end{aligned}$$

with probability at least $1-\delta$. This completes the proof of the theorem. □

## A.5 Proofs of instance based bounds

*Proof of Proposition 5 on page 49.* In the adversarial model, we have $\boldsymbol{X}_i=\boldsymbol{Y}_i$ if $i\notin O$ where $\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_n$ are generated from the reference distribution $\boldsymbol{\theta}^*$. By Proposition 3 and Lemma 6 in Berend and Kontorovich (2013) and using the triangle inequality

$$\begin{aligned}\left\|\bar{\boldsymbol{Y}}_n-\boldsymbol{\theta}^*\right\|_1-\frac{|O|}{n}\sup_{\boldsymbol{x},\boldsymbol{y}\in\Delta^\infty}\|\boldsymbol{x}-\boldsymbol{y}\|_1 &\le d_{\mathrm{TV}}(\bar{\boldsymbol{X}}_n,\boldsymbol{\theta}^*)\\ &\le \left\|\bar{\boldsymbol{Y}}_n-\boldsymbol{\theta}^*\right\|_1+\frac{|O|}{n}\sup_{\boldsymbol{x},\boldsymbol{y}\in\Delta^\infty}\|\boldsymbol{x}-\boldsymbol{y}\|_1,\end{aligned}$$

the proposition is proved. □

*Proof of Theorem 10 on page 49.* Using Theorem 2.1 in Cohen et al. (2020) and applying the same triangle inequality as in the proof of Proposition 5, we conclude the theorem. □

# Appendix B

# Proofs for Chapter 3



## B.1 Proof of Theorem 11

Before proving Theorem 11, we provide some auxiliary lemmas and propositions.

### B.1.1 Bounding the projected error of the average of filtered observations

For any $J \subset [n]$, we define $\overline{\boldsymbol{Z}}_J$ and $\widehat{\boldsymbol{\Sigma}}_J^Z$ as the sample average and sample covariance matrix of the subsample $\{\boldsymbol{Z}_i : i \in J\}$, that is

$$\overline{\boldsymbol{Z}}_J = \frac{1}{|J|}\sum_{i\in J} \boldsymbol{Z}_i, \qquad \widehat{\boldsymbol{\Sigma}}_J^Z = \frac{1}{|J|}\sum_{i\in J} \boldsymbol{Z}_i\boldsymbol{Z}_i^\top - \overline{\boldsymbol{Z}}_J\overline{\boldsymbol{Z}}_J^\top.$$

The main building block of the proof is the following result.

**Proposition 6.** *Let $\mathcal{S} \subset [n]$ be an arbitrary set. We define its subsets $\mathcal{S}_\mathcal{I} = \mathcal{S} \cap \mathcal{I}$ and $\mathcal{S}_\mathcal{O} = \mathcal{S} \cap \mathcal{O}$. Let $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n$ and $\boldsymbol{\mu}^Z$ be arbitrary points in $\mathbb{R}^q$ with $q \geq 2$. Let $\boldsymbol{\Sigma}^Z$ be an arbitrary $q \times q$*

*covariance matrix and let* $\mathtt{P}_k$ *be the projection matrix projecting onto the subspace spanned by the bottom* $k$ *eigenvectors of* $\widehat{\mathbf{\Sigma}}^Z_{\mathcal{S}} - \mathbf{\Sigma}^Z$*, for* $k = 1, \ldots, q-1$*. We have*

$$\big\|\mathtt{P}_k(\overline{\boldsymbol{Z}}_{\mathcal{S}} - \boldsymbol{\mu}^Z)\big\|_2 \le \left\{2\omega_{\mathcal{O}}\|\widehat{\mathbf{\Sigma}}^Z_{\mathcal{S}_{\mathcal{I}}} - \mathbf{\Sigma}^Z\|_{\mathrm{op}} + \frac{\omega_{\mathcal{O}}^2}{\omega_{\mathcal{I}}}\bigg((\lambda_q - \lambda_1)(\mathbf{\Sigma}^Z) + \frac{\delta_Z^2}{q-k}\bigg)\right\}^{1/2} + \big\|\mathtt{P}_k\overline{\boldsymbol{\xi}}^Z_{\mathcal{S}_{\mathcal{I}}}\big\|_2,$$

*where* $\omega_{\mathcal{O}} = |\mathcal{S}_{\mathcal{O}}|/|\mathcal{S}|$*,* $\omega_{\mathcal{I}} = 1 - \omega_{\mathcal{O}}$*,* $\boldsymbol{\xi}_i^Z = \boldsymbol{Z}_i - \boldsymbol{\mu}^Z$ *and* $\delta_Z = \inf_{\boldsymbol{\mu}} \max_{i\in\mathcal{S}} \|\boldsymbol{Z}_i - \boldsymbol{\mu}^Z\|_2$*. Furthermore, if* $|\mathcal{S}| \le q - k$*, then*

$$\big\|\mathtt{P}_k(\overline{\boldsymbol{Z}}_{\mathcal{S}} - \boldsymbol{\mu}^Z)\big\|_2 \le \left\{2\omega_{\mathcal{O}}\|\widehat{\mathbf{\Sigma}}^Z_{\mathcal{S}_{\mathcal{I}}} - \mathbf{\Sigma}^Z\|_{\mathrm{op}} + \frac{\omega_{\mathcal{O}}^2}{\omega_{\mathcal{I}}}(\lambda_q - \lambda_1)(\mathbf{\Sigma}^Z)\right\}^{1/2} + \big\|\mathtt{P}_k\overline{\boldsymbol{\xi}}^Z_{\mathcal{S}_{\mathcal{I}}}\big\|_2.$$

*Proof.* Since there is no risk of confusion, we remove the superscript $Z$ from $\mathbf{\Sigma}^Z$, $\widehat{\mathbf{\Sigma}}^Z_{\mathcal{S}}$ and so on. Since $\overline{\boldsymbol{Z}}_{\mathcal{S}} = \omega_{\mathcal{I}}\overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{I}}} + w_{\mathcal{O}}\overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{O}}}$ yields $\overline{\boldsymbol{Z}}_{\mathcal{S}} - \overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{I}}} = w_{\mathcal{O}}\big(\overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{O}}} - \overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{I}}}\big)$, the triangle inequality implies that

$$\begin{aligned}\big\|\mathtt{P}_k(\overline{\boldsymbol{Z}}_{\mathcal{S}} - \boldsymbol{\mu})\big\|_2 &\le \big\|\mathtt{P}_k(\overline{\boldsymbol{Z}}_{\mathcal{S}} - \overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{I}}})\big\|_2 + \big\|\mathtt{P}_k(\overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{I}}} - \boldsymbol{\mu})\big\|_2 \\ &\le w_{\mathcal{O}}\big\|\mathtt{P}_k(\overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{I}}} - \overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{O}}})\big\|_2 + \big\|\mathtt{P}_k\overline{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}}\big\|_2. \end{aligned} \tag{B.1}$$

Moreover, one can check that

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_{\mathcal{S}} &= \omega_{\mathcal{I}}\widehat{\mathbf{\Sigma}}_{\mathcal{S}_{\mathcal{I}}} + \omega_{\mathcal{O}}\widehat{\mathbf{\Sigma}}_{\mathcal{S}_{\mathcal{O}}} + \omega_{\mathcal{I}}\omega_{\mathcal{O}}(\overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{I}}} - \overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{O}}})^{\otimes 2} \\ &\succeq \omega_{\mathcal{I}}\widehat{\mathbf{\Sigma}}_{\mathcal{S}_{\mathcal{I}}} + \omega_{\mathcal{I}}\omega_{\mathcal{O}}(\overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{I}}} - \overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{O}}})^{\otimes 2}. \end{aligned} \tag{B.2}$$

Hence, multiplying from left and right by $\mathtt{P}_k$ and computing the largest eigenvalue of both sides, we get

$$\begin{aligned}\omega_{\mathcal{I}}\omega_{\mathcal{O}}\|\mathtt{P}_k(\overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{I}}} - \overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{O}}})\|_2^2 &\le \lambda_q\big(\mathtt{P}_k(\widehat{\mathbf{\Sigma}}_{\mathcal{S}} - \omega_{\mathcal{I}}\widehat{\mathbf{\Sigma}}_{\mathcal{S}_{\mathcal{I}}})\mathtt{P}_k^\top\big) \\ &\le \lambda_q\big(\mathtt{P}_k(\widehat{\mathbf{\Sigma}}_{\mathcal{S}} - \mathbf{\Sigma})\mathtt{P}_k^\top\big) + \lambda_q\big(\mathbf{\Sigma} - \omega_{\mathcal{I}}\widehat{\mathbf{\Sigma}}_{\mathcal{S}_{\mathcal{I}}}\big) \\ &\le \lambda_k(\widehat{\mathbf{\Sigma}}_{\mathcal{S}} - \mathbf{\Sigma}) + \omega_{\mathcal{I}}\,\lambda_q\big(\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_{\mathcal{S}_{\mathcal{I}}}\big) + \omega_{\mathcal{O}}\lambda_q(\mathbf{\Sigma}). \end{aligned} \tag{B.3}$$

On the other hand, using the Weyl inequality (several times) and the identity (B.2), we get

$$\begin{aligned}\lambda_k(\widehat{\mathbf{\Sigma}}_{\mathcal{S}} - \mathbf{\Sigma}) &\le \omega_{\mathcal{I}}\lambda_q\big(\widehat{\mathbf{\Sigma}}_{\mathcal{S}_{\mathcal{I}}} - \mathbf{\Sigma}\big) + \omega_{\mathcal{O}}\lambda_k\big(\widehat{\mathbf{\Sigma}}_{\mathcal{S}_{\mathcal{O}}} - \mathbf{\Sigma} + \omega_{\mathcal{I}}(\overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{I}}} - \overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{O}}})^{\otimes 2}\big) \\ &\le \omega_{\mathcal{I}}\lambda_q\big(\widehat{\mathbf{\Sigma}}_{\mathcal{S}_{\mathcal{I}}} - \mathbf{\Sigma}\big) + \omega_{\mathcal{O}}\lambda_{k+1}\big(\widehat{\mathbf{\Sigma}}_{\mathcal{S}_{\mathcal{O}}} - \mathbf{\Sigma}\big) + \omega_{\mathcal{I}}\lambda_{q-1}\big((\overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{I}}} - \overline{\boldsymbol{Z}}_{\mathcal{S}_{\mathcal{O}}})^{\otimes 2}\big) \\ &\le \omega_{\mathcal{I}}\lambda_q\big(\widehat{\mathbf{\Sigma}}_{\mathcal{S}_{\mathcal{I}}} - \mathbf{\Sigma}\big) + \omega_{\mathcal{O}}\lambda_{k+1}(\widehat{\mathbf{\Sigma}}_{\mathcal{S}_{\mathcal{O}}}) - \omega_{\mathcal{O}}\lambda_1(\mathbf{\Sigma}). \end{aligned} \tag{B.4}$$

For the middle term of the right hand side, we can use the following upper bound

$$\begin{aligned}\lambda_{k+1}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_\mathcal{O}}) &\leq \frac{\lambda_{k+1}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_\mathcal{O}}) + \ldots + \lambda_q(\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_\mathcal{O}})}{q-k} \\ &\leq \frac{\mathsf{Tr}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_\mathcal{O}})}{q-k} = \frac{1}{q-k}\,\mathsf{Tr}\left(\frac{1}{|\mathcal{S}_\mathcal{O}|}\sum_{i\in\mathcal{S}_\mathcal{O}}(\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_{\mathcal{S}_\mathcal{O}})^{\otimes 2}\right) \\ &= \frac{1}{q-k}\inf_{\boldsymbol{\mu}}\mathsf{Tr}\left(\frac{1}{|\mathcal{S}_\mathcal{O}|}\sum_{i\in\mathcal{S}_\mathcal{O}}(\boldsymbol{Z}_i - \boldsymbol{\mu})^{\otimes 2}\right) \\ &= \inf_{\boldsymbol{\mu}}\frac{1}{(q-k)|\mathcal{S}_\mathcal{O}|}\sum_{i\in\mathcal{S}_\mathcal{O}}\|\boldsymbol{Z}_i - \boldsymbol{\mu}\|_2^2 \leq \frac{\delta_Z^2}{q-k}.\end{aligned}$$

Combining (B.3), (B.4) and the last display, we get

$$\omega_\mathcal{I}\omega_\mathcal{O}\|\mathsf{P}_k(\overline{\boldsymbol{Z}}_{\mathcal{S}_\mathcal{I}} - \overline{\boldsymbol{Z}}_{\mathcal{S}_\mathcal{O}})\|_2^2 \leq 2\omega_\mathcal{I}\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_\mathcal{I}} - \boldsymbol{\Sigma}\|_{\mathrm{op}} + \omega_\mathcal{O}\big(\lambda_q(\boldsymbol{\Sigma}) - \lambda_1(\boldsymbol{\Sigma})\big) + \frac{\omega_\mathcal{O}\delta_Z^2}{q-k}.$$

Dividing both sides by $\omega_\mathcal{I}\omega_\mathcal{O}$, we arrive at

$$\|\mathsf{P}_k(\overline{\boldsymbol{Z}}_{\mathcal{S}_\mathcal{I}} - \overline{\boldsymbol{Z}}_{\mathcal{S}_\mathcal{O}})\|_2^2 \leq \frac{2}{\omega_\mathcal{O}}\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_\mathcal{I}} - \boldsymbol{\Sigma}\|_{\mathrm{op}} + \frac{1}{\omega_\mathcal{I}}\left(\lambda_q(\boldsymbol{\Sigma}) - \lambda_1(\boldsymbol{\Sigma}) + \frac{\delta_Z^2}{q-k}\right).$$

Combining this inequality with (B.1), we obtain the first claim of the proposition. To get the second claim, we simply remark that $\lambda_{k+1}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_\mathcal{O}}) = 0$ since the rank of the matrix $\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_\mathcal{O}}$ is less than $|\mathcal{S}_\mathcal{O}| \leq |\mathcal{S}| \leq q-k$. □

### B.1.2 Bounding the error of the geometric median of projected observations

We assume in this section that $V$ is a linear subspace of $\mathbb{R}^p$ of dimension $k$ and consider the geometric median $\widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}}$ of the projected vectors $\mathsf{P}_V\boldsymbol{X}_1, \ldots, \mathsf{P}_V\boldsymbol{X}_n$ that is

$$\widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}} \in \mathsf{arg}\min_{\boldsymbol{\mu}\in\mathbb{R}^p}\sum_{i=1}^n\|\mathsf{P}_V\boldsymbol{X}_i - \boldsymbol{\mu}\|_2.$$

**Lemma 2.** *With probability at least $1-\delta$, for all linear subspaces $V \subset \mathbb{R}^p$, we have*

$$\frac{\|\widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}} - \mathsf{P}_V\boldsymbol{\mu}^*\|_2}{\sqrt{\dim(V)}} \leq \frac{2\sqrt{\|\boldsymbol{\Sigma}\|_{\mathrm{op}}}}{1-2\varepsilon}\left(1 + \frac{\sqrt{\mathsf{r}_\boldsymbol{\Sigma}} + \sqrt{2\log(1/\delta)}}{\sqrt{n}}\right).$$

*Proof.* It follows from (Dalalyan and Minasyan, 2020, Lemma 2) that

$$\|\widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}} - \mathsf{P}_V\boldsymbol{\mu}^*\|_2 \leq \frac{2}{n(1-2\varepsilon)}\sum_{i=1}^n\|\mathsf{P}_V\boldsymbol{\xi}_i\|_2.$$

Upper bounding the last term using Cauchy-Schwartz's inequality, one obtains

$$\|\widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}} - \mathtt{P}_V \boldsymbol{\mu}^*\|_2 \leq \frac{2}{\sqrt{n}(1-2\varepsilon)} \Big( \sum_{i=1}^n \|\mathtt{P}_V \boldsymbol{\xi}_i\|_2^2 \Big)^{1/2}.$$

Let $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k$ be any orthonormal basis of $V$, with $k = \dim(V)$. We have

$$\begin{aligned}
\|\widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}} - \mathtt{P}_V \boldsymbol{\mu}^*\|_2 &\leq \frac{2}{\sqrt{n}(1-2\varepsilon)} \Big( \sum_{i=1}^n \sum_{\ell=1}^k |\boldsymbol{e}_\ell^\top \boldsymbol{\xi}_i|^2 \Big)^{1/2} \\
&\leq \frac{2}{\sqrt{n}(1-2\varepsilon)} \Big( k \sup_{\|\boldsymbol{e}\|_2=1} \sum_{i=1}^n |\boldsymbol{e}^\top \boldsymbol{\xi}_i|^2 \Big)^{1/2} \\
&= \frac{2}{\sqrt{n}(1-2\varepsilon)} \Big( k \sup_{\|\boldsymbol{e}\|_2=1} \big\| [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n]^\top \boldsymbol{e} \big\|_2^2 \Big)^{1/2} \\
&= \frac{2\sqrt{k}}{\sqrt{n}(1-2\varepsilon)} \big\| [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n] \big\|_{\mathrm{op}}.
\end{aligned}$$

By Corollary 5.35 in (Vershynin, 2012), the inequality

$$\big\| [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n] \big\|_{\mathrm{op}} \leq \|\boldsymbol{\Sigma}\|_{\mathrm{op}}^{1/2} \big( \sqrt{n} + \sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}} + \sqrt{2\log(1/\delta)} \big)$$

holds with probability at least $1-\delta$. This yields the desired inequality. □

### B.1.3 Bounding the number of filtered out observations

Throughout this section, we assume without loss of generality that $\|\boldsymbol{\Sigma}\|_{\mathrm{op}} = 1$.

**Lemma 3.** *Let $\tau$ and $\delta$ be two numbers from $(0,1)$. Define*

$$z = 1 + \frac{\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}} + \sqrt{2\log(1/\delta)}}{\sqrt{n\tau}} + \sqrt{2 + 2\log\big(1/\tau\big)}.$$

*With probability at least $1-\delta$, we have*

$$\sup_V \sum_{i=1}^n \mathbb{1}\big( \|\mathtt{P}_V \boldsymbol{\xi}_i\|_2^2 > z^2 \dim(V) \big) \leq n\tau,$$

*where the supremum is over all linear subspaces $V$ of $\mathbb{R}^p$.*

*Proof.* Let us define the random variable

$$T_n = \sup_V \sum_{i=1}^n \mathbb{1}\big( \|\mathtt{P}_V \boldsymbol{\xi}_i\|_2^2 > z^2 \dim(V) \big).$$

In what follows, we write $d_V$ for the dimension of the subspace $V$. We check that

$$\begin{aligned}
\mathbf{P}(T_n > n\tau) &\le \mathbf{P}\Big(\exists V \subset \mathbb{R}^p, \exists J \subset [n] \text{ with } |J| = n\tau, \text{ s.t. } \min_{i\in J} \|\mathtt{P}_V \boldsymbol{\xi}_i\|_2^2 \ge z^2 \dim(V)\Big) \\
&\le \sum_{\substack{J\subseteq[n] \\ |J|=n\tau}} \mathbf{P}\Big(\sup_V \min_{i\in J} \|\mathtt{P}_V \boldsymbol{\xi}_i\|_2^2/d_V \ge z^2\Big) \\
&= \binom{n}{n\tau} \mathbf{P}\Big(\sup_V \min_{i\in\{1,\dots,n\tau\}} \|\mathtt{P}_V \boldsymbol{\xi}_i\|_2^2/d_V \ge z^2\Big) \\
&\le \binom{n}{n\tau} \mathbf{P}\Big(\sup_V \frac{1}{n\tau d_V} \sum_{i=1}^{n\tau} \|\mathtt{P}_V \boldsymbol{\xi}_i\|_2^2 \ge z^2\Big). \qquad \text{(B.5)}
\end{aligned}$$

Given a linear subspace $V \subset \mathbb{R}^p$ of dimension $d_V$, let $\boldsymbol{e}_1^V, \dots, \boldsymbol{e}_{d_V}^V$ be an orthonormal basis of $V$. Using (B.5), we get

$$\begin{aligned}
\mathbf{P}(T_n \ge n\tau) &\le \binom{n}{n\tau} \mathbf{P}\Big(\sup_V \frac{1}{n\tau d_V} \sum_{l=1}^{d_V} \sum_{i=1}^{n\tau} \big|\boldsymbol{\xi}_i^\top \boldsymbol{e}_l^V\big|^2 \ge z^2\Big) \\
&\le \binom{n}{n\tau} \mathbf{P}\Big(\sup_{\|\boldsymbol{e}\|=1} \frac{1}{n\tau} \sum_{i=1}^{n\tau} \big|\boldsymbol{\xi}_i^\top \boldsymbol{e}\big|^2 \ge z^2\Big) \\
&= \binom{n}{n\tau} \mathbf{P}\Big(\|[\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_{n\tau}]\|_{\mathrm{op}} \ge z\sqrt{n\tau}\Big).
\end{aligned}$$

By (Vershynin, 2012, Corollary 5.35), the inequality

$$\big\|[\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_{n\tau}]\big\|_{\mathrm{op}} \le \big(\sqrt{n\tau} + \sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}} + \sqrt{2\log(1/\delta_0)}\big)$$

holds with probability at least $1 - \delta_0$. Taking $\delta_0 = \delta/\binom{n}{n\tau}$ and using the inequality $\log\binom{n}{n\tau} \le n\tau\big(1 + \log(1/\tau)\big)$, we can conclude that $\mathbf{P}(T_n \ge n\tau) \le \delta$. This proves the lemma. □

**Lemma 4.** *Let $\tau \in (0, 1/2)$ and $\delta \in (0, 1/2)$ be arbitrary. Set*

$$t = \frac{3 - 2\varepsilon^*}{1 - 2\varepsilon^*}\left(1 + \frac{\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}} + \sqrt{2\log(2/\delta)}}{\sqrt{n\tau}}\right) + \sqrt{2 + 2\log\big(1/\tau\big)},$$

*where $\varepsilon^* < 1/2$. Assume that $\{\boldsymbol{X}_i\}$ are drawn from GAC$(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon)$ model with $\varepsilon \le \varepsilon^*$. Then, with probability at least $1 - \delta$, for any linear subspace $V$ of $\mathbb{R}^p$, the inequalities*

$$\begin{aligned}
N_V = \sum_{i\in\mathcal{I}} \mathbb{1}\left(\frac{\|\mathtt{P}_V \boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}}\|_2}{\sqrt{\dim(V)}} \le t\right) &\ge n(1 - \varepsilon - \tau), \\
\frac{\|\mathtt{P}_V \boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}}\|_2}{\sqrt{\dim(V)}} &\le \frac{2}{1 - 2\varepsilon}\left(1 + \frac{\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}} + \sqrt{2\log(2/\delta)}}{\sqrt{n}}\right),
\end{aligned}$$

*where $\widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}}$ is the geometric median of $\mathtt{P}_V \boldsymbol{X}_1, \dots, \mathtt{P}_V \boldsymbol{X}_n$.*

*Proof.* To avoid unnecessary technicalities, we assume in this proof that $n\tau$ is an integer. We

also write

$$t_1 = \frac{2}{1-2\varepsilon}\left(1 + \frac{\sqrt{\mathbf{r}_{\mathbf{\Sigma}}} + \sqrt{2\log(2/\delta)}}{\sqrt{n}}\right)$$
$$t_2 = 1 + \frac{\sqrt{\mathbf{r}_{\mathbf{\Sigma}}} + \sqrt{2\log(2/\delta)}}{\sqrt{n\tau}} + \sqrt{2 + 2\log(1/\tau)},$$

so that $t \geq t_1 + t_2$. Simple algebra yields

$$\begin{aligned}
N_V &= \sum_{i\in\mathcal{I}} \mathbb{1}\big(\|\mathtt{P}_V\boldsymbol{\xi}_i + \mathtt{P}_V\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}}\|_2 \leq t\sqrt{\dim(V)}\big) \\
&\geq \sum_{i=1}^{n} \mathbb{1}\big(\|\mathtt{P}_V\boldsymbol{\xi}_i + \mathtt{P}_V\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}}\|_2 \leq t\sqrt{\dim(V)}\big) - n\varepsilon \\
&= n - n\varepsilon - \sum_{i=1}^{n} \mathbb{1}\big(\|\mathtt{P}_V\boldsymbol{\xi}_i + \mathtt{P}_V\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}}\|_2 > t\sqrt{\dim(V)}\big) \\
&\geq n - n\varepsilon - \sum_{i=1}^{n} \mathbb{1}\big(\|\mathtt{P}_V\boldsymbol{\xi}_i\|_2 + \|\mathtt{P}_V\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}}\|_2 > t\sqrt{\dim(V)}\big),
\end{aligned}$$

where in the last step, we used the triangle inequality. According to Lemma 2, on an event $\Omega_1$ of probability at least $1-\delta/2$, we have $\|\mathtt{P}_V\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}}\|_2 \leq t_1\sqrt{\dim(V)}$ for every $V$. This implies that on this event,

$$N_V \geq n - n\varepsilon - \sum_{i=1}^{n} \mathbb{1}\big(\|\mathtt{P}_V\boldsymbol{\xi}_i\|_2 > t_2\sqrt{\dim(V)}\big), \qquad \forall V \subset \mathbb{R}^p.$$

Using Lemma 3, we get that on an event $\Omega_2$ of probability at least $1-\delta/2$, the sum on the right hand side of the last display is less than $n\tau$. Therefore, on the intersection of the events $\Omega_1$ and $\Omega_2$, we have $N_V \geq n(1-\varepsilon-\tau)$ for every linear subspace $V$ of $\mathbb{R}^p$. □

### B.1.4 Estimating the mean from a low-dimensional adversarial projection

In this section, we consider the following problem. We assume that for a $q$ dimensional linear subspace $V$ of $\mathbb{R}^p$, which can depend on the sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, we observe the projected data $\mathtt{P}_V\boldsymbol{X}_1, \ldots, \mathtt{P}_V\boldsymbol{X}_n$. The goal is to estimate the projected mean $\boldsymbol{\mu}_V^* = \mathtt{P}_V\boldsymbol{\mu}^* \in \mathbb{R}^q$. We will estimate $\boldsymbol{\mu}_V^*$ by the mean of filtered data points. More precisely, let $\widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}}$ be the geometric median of $\mathtt{P}_V\boldsymbol{X}_1, \ldots, \mathtt{P}_V\boldsymbol{X}_n$. We set

$$\mathcal{S}_V = \{i \in [n] : \|\mathtt{P}_V\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}}\|_2 \leq t\sqrt{q}\}, \quad N_V = |\mathcal{S}_V|,$$

where $t$ is a real number. We estimate $\boldsymbol{\mu}_V^*$ by

$$\widehat{\boldsymbol{\mu}}_V = \mathtt{P}_V\overline{\boldsymbol{X}}_{\mathcal{S}_V} = \frac{1}{N_V}\sum_{i\in\mathcal{S}_V} \mathtt{P}_V\boldsymbol{X}_i.$$

**Lemma 5.** *For every positive threshold $t > 0$, we have*

$$\|\widehat{\boldsymbol{\mu}}_V - \boldsymbol{\mu}_V^*\|_2 \le \frac{\big\|\mathtt{P}_V\bar{\boldsymbol{\xi}}\big\|_2}{N_V} + \frac{1}{N_V}\Big\|\sum_{i\in\mathcal{S}_V^c\cup\mathcal{O}}\mathtt{P}_V\boldsymbol{\xi}_i\Big\|_2 + \frac{n\varepsilon(t\sqrt{q} + \|\widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}} - \boldsymbol{\mu}_V^*\|_2)}{N_V}.$$

*Proof.* For this estimator, using the triangle inequality and the fact that $\boldsymbol{X}_i = \boldsymbol{\mu}^* + \boldsymbol{\xi}_i$ for every $i \in \mathcal{I}$, we have

$$\begin{aligned}
\|\widehat{\boldsymbol{\mu}}_V - \boldsymbol{\mu}_V^*\|_2 &= \frac{1}{N_V}\Big\|\sum_{i\in\mathcal{S}_V}(\mathtt{P}_V\boldsymbol{X}_i - \mathtt{P}_V\boldsymbol{\mu}^*)\Big\|_2 \\
&\le \frac{1}{N_V}\Big\|\sum_{i\in\mathcal{S}_V\cap\mathcal{I}}\mathtt{P}_V\boldsymbol{\xi}_i\Big\|_2 + \frac{1}{N_V}\Big\|\sum_{i\in\mathcal{S}_V\cap\mathcal{O}}\mathtt{P}_V(\boldsymbol{X}_i - \boldsymbol{\mu}^*)\Big\|_2 \\
&\le \frac{1}{N_V}\Big\|\sum_{i\in\mathcal{S}_V\cap\mathcal{I}}\mathtt{P}_V\boldsymbol{\xi}_i\Big\|_2 + \frac{n\varepsilon(t\sqrt{q} + \|\widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}} - \boldsymbol{\mu}_V^*\|_2)}{N_V}.
\end{aligned}$$

Using once again the triangle inequality, we arrive at

$$\begin{aligned}
\|\widehat{\boldsymbol{\mu}}_V - \boldsymbol{\mu}_V^*\|_2 &\le \frac{\big\|\mathtt{P}_V\bar{\boldsymbol{\xi}}\big\|_2}{N_V} + \frac{1}{N_V}\Big\|\sum_{i\in\mathcal{S}_V^c\cup\mathcal{O}}\mathtt{P}_V\boldsymbol{\xi}_i\Big\|_2 + \frac{n\varepsilon(t\sqrt{q} + \|\widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}} - \boldsymbol{\mu}_V^*\|_2)}{N_V} \\
&\le \frac{\big\|\mathtt{P}_V\bar{\boldsymbol{\xi}}\big\|_2}{N_V} + \frac{1}{N_V}\Big\|\sum_{i\in\mathcal{S}_V^c\cup\mathcal{O}}\mathtt{P}_V\boldsymbol{\xi}_i\Big\|_2 + \frac{n\varepsilon(t\sqrt{q} + \|\widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}} - \boldsymbol{\mu}_V^*\|_2)}{N_V}.
\end{aligned}$$

This completes the proof. □

### B.1.5 Bounding stochastic errors

Throughout this section, without loss of generality, we assume that $\|\boldsymbol{\Sigma}\|_{\mathrm{op}} = 1$.

**Lemma 6.** *For any positive integer $m \le n$ and any $t > 0$, we have*

$$\mathbf{P}\bigg(\max_{|S|\ge n-m}\Big\|\frac{1}{|S|}\sum_{i\in S}\mathtt{P}\,\boldsymbol{\xi}_i\Big\|_2 \le \frac{n\|\mathtt{P}\bar{\boldsymbol{\xi}}_n\|_2}{n-m} + \frac{\sqrt{m(2\mathbf{r}_{\boldsymbol{\Sigma}} + 3t)} + m\sqrt{3\log(2ne/m)}}{n-m}\bigg) \ge 1 - e^{-t}.$$

*Proof.* Using the triangle inequality, one has

$$\begin{aligned}
\frac{1}{|S|}\Big\|\sum_{i\in S}\mathtt{P}\boldsymbol{\xi}_i\Big\|_2 \le \frac{1}{n-m}\Big\|\sum_{i\in S}\mathtt{P}\boldsymbol{\xi}_i\Big\|_2 &= \frac{1}{n-m}\Big\|\sum_{i=1}^n\mathtt{P}\boldsymbol{\xi}_i - \sum_{i\in S^C}\mathtt{P}\boldsymbol{\xi}_i\Big\|_2 \\
&\le \frac{1}{n-m}\Big\|\sum_{i=1}^n\mathtt{P}\boldsymbol{\xi}_i\Big\|_2 + \frac{1}{n-m}\Big\|\sum_{i\in S^C}\mathtt{P}\boldsymbol{\xi}_i\Big\|_2 \\
&\le \frac{n\|\mathtt{P}\bar{\boldsymbol{\xi}}_n\|_2}{n-m} + \frac{1}{n-m}\max_{|J|\le m}\Big\|\sum_{i\in J}\boldsymbol{\xi}_i\Big\|_2. \qquad \text{(B.6)}
\end{aligned}$$

For $s \in [1, m]$, we choose $t_s$ by

$$t_s = 3s \log\Big(\frac{2ne}{s}\Big) + 3t,$$

so that $t_s \le t_m = 3m \log\big(\frac{2ne}{m}\big) + 3t$ and

$$\Big(\frac{ne}{s}\Big)^s e^{-t_s/3} \le 2^{-s} e^{-t}.$$

For every $J$ of cardinality $m$, the random variable $\|\sum_{j\in J} \boldsymbol{\xi}_j\|^2$ has the same distribution as $m\sum_{j=1}^p \lambda_j(\boldsymbol{\Sigma})\alpha_j^2$, where $\alpha_1, \dots, \alpha_p$ are i.i.d. standard Gaussian. Hence, using the union bound, the well-known upper bound on the binomial coefficients and (Comminges and Dalalyan, 2012, Lemma 8), we have

$$\begin{aligned}
\mathbf{P}\Big(\max_{|J|\le m} \Big\|\sum_{i\in J} \boldsymbol{\xi}_i\Big\|_2^2 \ge m(2\mathbf{r}_{\boldsymbol{\Sigma}} + t_m)\Big) &\le \sum_{s=1}^m \binom{n}{s} \mathbf{P}\Big(\Big\|\sum_{i=1}^s \boldsymbol{\xi}_i\Big\|_2^2 \ge m(2\mathbf{r}_{\boldsymbol{\Sigma}} + t_m)\Big) \\
&\le \sum_{s=1}^m \Big(\frac{ne}{s}\Big)^s \mathbf{P}\Big(\Big\|\sum_{i=1}^s \boldsymbol{\xi}_i\Big\|_2^2 \ge s(2\mathbf{r}_{\boldsymbol{\Sigma}} + t_s)\Big) \\
&\le \sum_{s=1}^m \Big(\frac{ne}{s}\Big)^s e^{-t_s/3} \le e^{-t}.
\end{aligned}$$

This entails that with probability at least $1 - e^{-t}$, we have

$$\max_{|J|\le m} \Big\|\sum_{i\in J} \boldsymbol{\xi}_i\Big\|_2 \le \sqrt{m(2\mathbf{r}_{\boldsymbol{\Sigma}} + t_m)} \le \sqrt{m(2\mathbf{r}_{\boldsymbol{\Sigma}} + 3t)} + m\sqrt{3\log(2ne/m)}.$$

Combining this inequality with (B.6), we get the claim of the lemma. □

In the two next lemmas, given a set $\mathcal{S} \subset [n]$, we look at the sample average and sample covariance matrix of the subsample $\{\boldsymbol{X}_i : i \in \mathcal{S}\}$,

$$\overline{\boldsymbol{X}}_{\mathcal{S}} = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \boldsymbol{X}_i, \qquad \widehat{\boldsymbol{\Sigma}}_{\mathcal{S}} = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \boldsymbol{X}_i\boldsymbol{X}_i^\top - \overline{\boldsymbol{X}}_{\mathcal{S}}\overline{\boldsymbol{X}}_{\mathcal{S}}^\top.$$

**Lemma 7.** *There exists a positive constant $A$ such that, for any positive integer $m \le n$ and any $t \ge 1$, with probability at least $1 - 2e^{-t}$, the inequality*

$$\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}} - \boldsymbol{\Sigma}\|_{\mathrm{op}} \le A\frac{\sqrt{n\mathbf{r}_{\boldsymbol{\Sigma}}} + \mathbf{r}_{\boldsymbol{\Sigma}} + m\log(2ne/m) + 2t}{n-m} + \big\|\overline{\boldsymbol{\xi}}_{\mathcal{S}}\big\|_2^2$$

*is satisfied for every $\mathcal{S} \subset [n]$ of cardinality $\ge n - m$.*

*Proof.* The triangle inequality implies

$$\begin{aligned}\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}} - \boldsymbol{\Sigma}\|_{\mathrm{op}} &\le \Big\| \frac{1}{|\mathcal{S}|} \sum_{i\in\mathcal{S}} (\boldsymbol{X}_i - \boldsymbol{\mu})^{\otimes 2} - \boldsymbol{\Sigma} \Big\|_{\mathrm{op}} + \big\| (\boldsymbol{\mu} - \overline{\boldsymbol{X}}_{\mathcal{S}})^{\otimes 2} \big\|_{\mathrm{op}} \\ &= \Big\| \frac{1}{|\mathcal{S}|} \sum_{i\in\mathcal{S}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top - \boldsymbol{\Sigma} \Big\|_{\mathrm{op}} + \|\overline{\boldsymbol{\xi}}_{\mathcal{S}}\|_2^2. \end{aligned} \tag{B.7}$$

Using again the triangle inequality, one gets

$$\begin{aligned}\Big\| \sum_{i\in\mathcal{S}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top - \boldsymbol{\Sigma} \Big\|_{\mathrm{op}} &\le \Big\| \sum_{i\in\mathcal{S}} (\boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top - \boldsymbol{\Sigma}) \Big\|_{\mathrm{op}} \\ &\le \Big\| \sum_{i=1}^{n} (\boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top - \boldsymbol{\Sigma}) \Big\|_{\mathrm{op}} + \Big\| \sum_{i\in\mathcal{S}^C} (\boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top - \boldsymbol{\Sigma}) \Big\|_{\mathrm{op}} \\ &\le \Big\| \sum_{i=1}^{n} (\boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top - \boldsymbol{\Sigma}) \Big\|_{\mathrm{op}} + \max_{|J|\le m} \Big\| \sum_{i\in J} (\boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top - \boldsymbol{\Sigma}) \Big\|_{\mathrm{op}}. \end{aligned} \tag{B.8}$$

In view of (Koltchinskii and Lounici, 2017, Theorems 4 and 5), one can show that there exists a positive universal constant $A_1$ such that for every $t \ge 1$ and every set $J$ of cardinality $s$, the inequality

$$\Big\| \sum_{j\in J} \boldsymbol{\xi}_j \boldsymbol{\xi}_j^\top - s\boldsymbol{\Sigma} \Big\|_{\mathrm{op}} \le A_1 (\sqrt{m \mathbf{r}_{\boldsymbol{\Sigma}}} + \mathbf{r}_{\boldsymbol{\Sigma}} + t)$$

is satisfied with probability at least $1 - e^{-t}$. We define $t_s$ by

$$t_s = s \log\Big(\frac{2ne}{s}\Big) + t,$$

so that $t_s \le t_m = m \log\big(\frac{2ne}{m}\big) + t$ and

$$\Big(\frac{ne}{s}\Big)^s e^{-t_s} \le 2^{-s} e^{-t}.$$

Applying the union bound and the well-known upper bound on the binomial coefficients, this yields

$$\begin{aligned}\mathbf{P}\Big( \max_{|J|\le m} \Big\| \sum_{j\in J} (\boldsymbol{\xi}_j \boldsymbol{\xi}_j^\top - \boldsymbol{\Sigma}) \Big\|_{\mathrm{op}} &\ge A_1 (\sqrt{m \mathbf{r}_{\boldsymbol{\Sigma}}} + \mathbf{r}_{\boldsymbol{\Sigma}} + t_m) \Big) \\ &\le \sum_{s=1}^{m} \binom{n}{s} \mathbf{P}\Big( \Big\| \sum_{j=1}^{s} \boldsymbol{\xi}_j \boldsymbol{\xi}_j^\top - s\boldsymbol{\Sigma} \Big\|_{\mathrm{op}} \ge A_1 (\sqrt{m \mathbf{r}_{\boldsymbol{\Sigma}}} + \mathbf{r}_{\boldsymbol{\Sigma}} + t_s) \Big) \\ &\le \sum_{s=1}^{m} \Big(\frac{ne}{s}\Big)^s e^{-t_s} \le e^{-t}. \end{aligned}$$

One deduces from (B.8) that, with probability at least $1-2e^{-t}$,

$$\Big\|\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top-\boldsymbol{\Sigma}\Big\|_{\mathrm{op}}\le A_1\frac{(\sqrt{n}+\sqrt{m})\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}}+2\mathbf{r}_{\boldsymbol{\Sigma}}+m\log(2ne/m)+2t}{n-m}$$
$$\le A\frac{\sqrt{n\mathbf{r}_{\boldsymbol{\Sigma}}}+\mathbf{r}_{\boldsymbol{\Sigma}}+m\log(2ne/m)+2t}{n-m}.$$

Combining this with (B.7), one gets the claim of the lemma. □

**Lemma 8.** *For any positive integer $m\le n$ and any $t>0$, with probability at least $1-4e^{-t}$, the inequality*

$$\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}}-\boldsymbol{\Sigma}\|_{\mathrm{op}}\le\frac{5p+\big(8\log(2ne/m)+2\big)m+7t}{n-m}+2\frac{\sqrt{p}+\sqrt{t}}{\sqrt{n}-\sqrt{m}}+\|\overline{\boldsymbol{\xi}}_{\mathcal{S}}\|_2^2$$

*is satisfied for every $\mathcal{S}\subset[n]$ of cardinality $\ge n-m$.*

*Proof.* In this proof, without loss of generality, we assume that the matrix $\boldsymbol{\Sigma}$ is invertible. In view of (B.7) and (B.8), we have

$$\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}}-\boldsymbol{\Sigma}\|_{\mathrm{op}}\le\Big\|\frac{1}{|\mathcal{S}|}\sum_{i=1}^n\big(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top-\boldsymbol{\Sigma}\big)\Big\|_{\mathrm{op}}+\max_{|J|\le m}\Big\|\frac{1}{|\mathcal{S}|}\sum_{i\in J}\big(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top-\boldsymbol{\Sigma}\big)\Big\|_{\mathrm{op}}+\|\overline{\boldsymbol{\xi}}_{\mathcal{S}}\|_2^2. \tag{B.9}$$

Let us define $\boldsymbol{\zeta}_i:=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\xi}_i$ for all $i\in[n]$. For every set $J$ of cardinality $s$, it holds that

$$\Big\|\sum_{i\in J}\big(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top-\boldsymbol{\Sigma}\big)\Big\|_{\mathrm{op}}\le\|\boldsymbol{\Sigma}\|_{\mathrm{op}}\Big\|\sum_{i\in J}\big(\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top-\mathbf{I}_p\big)\Big\|_{\mathrm{op}}$$
$$=\max\Big(\lambda_{\max}\big(\sum_{i\in J}\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top\big)-s,s-\lambda_{\min}\big(\sum_{i\in J}\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top\big)\Big)$$
$$=\max\Big(\sigma^2_{\max}(\boldsymbol{\zeta}_J)-s,s-\sigma^2_{\min}(\boldsymbol{\zeta}_J)\Big),$$

where $\boldsymbol{\zeta}_J$ is the $s\times n$ random matrix obtained by concatenating the vectors $\boldsymbol{\zeta}_i$ with $i\in J$. By (Vershynin, 2012, Corollary 5.35), we know that for every $x>0$

$$\sqrt{s}-\sqrt{p}-x\le\sigma_{\min}(\boldsymbol{\zeta}_J)\le\sigma_{\max}(\boldsymbol{\zeta}_J)\le\sqrt{s}+\sqrt{p}+x$$

is true with probability at least $1-2e^{-x^2/2}$. This yields[1]

$$\Big\|\sum_{i\in J}\big(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top-\boldsymbol{\Sigma}\big)\Big\|_{\mathrm{op}}\le\max\big((\sqrt{p}+x)(2\sqrt{s}+\sqrt{p}+x),(\sqrt{p}+x)(2\sqrt{s}-\sqrt{p}-x)\big)$$
$$\le p+x^2+2\sqrt{ps}+2x\sqrt{p}+2x\sqrt{s}$$

with probability at least $1-2e^{-x^2/2}$. By applying the same technique as in the proof of

[1] We provide the argument only in the case $\sqrt{s}\ge\sqrt{p}+x$, but the conclusion is true for every value $s$.

Lemma 7, we can set

$$t_s = 2\sqrt{s\log\Big(\frac{2ne}{s}\Big)} + t,$$

and obtain

$$\mathbf{P}\Big(\max_{|J|\le m}\Big\|\sum_{j\in J}(\boldsymbol{\xi}_j\boldsymbol{\xi}_j^\top - \boldsymbol{\Sigma})\Big\|_{\mathrm{op}} \ge p + t_m^2 + 2\sqrt{pm} + 2t_m\sqrt{p} + 2t_m\sqrt{m}\Big) \le 2e^{-t}.$$

Hence, going back to (B.9), we can show that the inequalities

$$\begin{aligned}\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}} - \boldsymbol{\Sigma}\|_{\mathrm{op}} &\le \frac{p + t + 2\sqrt{pn} + 2\sqrt{tp} + 2\sqrt{tn}}{n-m} + \frac{p + 4t + 4m\log(2ne/m) + 2\sqrt{pm}}{n-m} \\ &\quad + \frac{4(\sqrt{p}+\sqrt{m})\sqrt{m\log(2ne/m)+t}}{n-m} + \|\overline{\boldsymbol{\xi}}_{\mathcal{S}}\|_2^2 \\ &\le \frac{5p + 8m\log(2ne/m) + 2m + 7t}{n-m} + \frac{2(\sqrt{p}+\sqrt{t})(\sqrt{n}+\sqrt{m})}{n-m} + \|\overline{\boldsymbol{\xi}}_{\mathcal{S}}\|_2^2\end{aligned}$$

hold with probability at least $1 - 4e^{-t}$, and this proves the lemma. □

### B.1.6 Putting all the pieces together

All the ingredients provided, we can now compile the complete proof of Theorem 11.

Taking $\mathbf{U}_L := \mathbf{V}_L$, the algorithm detailed in (3.2) returns $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}} = \sum_{\ell=0}^{L}\widehat{\boldsymbol{\mu}}^{(\ell)}$ with $\widehat{\boldsymbol{\mu}}^{(\ell)} \in \mathscr{U}_\ell = \mathrm{Im}(\mathbf{V}_\ell\mathbf{U}_\ell^\top)$ for every $\ell \in \{0,\dots,L\}$ where the two-by-two orthogonal subspaces $\mathscr{U}_0,\dots,\mathscr{U}_L$ span the whole space $\mathbb{R}^p$. This allows us to decompose the risk:

$$\begin{aligned}\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}} - \boldsymbol{\mu}^*\|_2^2 &= \sum_{\ell=0}^{L}\|\widehat{\boldsymbol{\mu}}^{(\ell)} - \mathtt{P}_{\mathscr{U}_\ell}\boldsymbol{\mu}^*\|_2^2 = \sum_{\ell=0}^{L}\|\mathtt{P}_{\mathscr{U}_\ell}(\overline{\boldsymbol{X}}_\ell - \boldsymbol{\mu}^*)\|_2^2 \\ &= \sum_{\ell=0}^{L}\|\mathtt{P}_\ell(\mathbf{V}_\ell^\top\overline{\boldsymbol{X}}_\ell - \mathbf{V}_\ell^\top\boldsymbol{\mu}^*)\|_2^2,\end{aligned}$$

where $\mathtt{P}_\ell := \mathbf{U}_\ell^\top\mathbf{U}_\ell$ is the projection matrix projecting onto the subspace of $\mathbb{R}^{p_\ell}$ spanned by the bottom $p_\ell - p_{\ell+1}$ eigenvectors of $\mathbf{V}_\ell^\top(\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma})\mathbf{V}_\ell$ for $\ell = 0,\dots,L$ with the convention that $p_{L+1} = 0$.

For $\ell \in \{0,\dots,L-1\}$, we intend to apply Proposition 6 to $\boldsymbol{Z}_i = \mathbf{V}_\ell^\top\boldsymbol{X}_i$ and $\boldsymbol{\mu}^Z = \mathbf{V}_\ell^\top\boldsymbol{\mu}^*$ in order to upper bound the error term $\mathrm{Err}_\ell := \|\mathtt{P}_\ell(\mathbf{V}_\ell^\top\overline{\boldsymbol{X}}_\ell - \mathbf{V}_\ell^\top\boldsymbol{\mu}^*)\|_2$. Using the inequalities

$$\|\mathbf{V}^\top(\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma})\mathbf{V}\|_{\mathrm{op}} \le \|\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma}\|_{\mathrm{op}}, \quad \lambda_{p_\ell}(\mathbf{V}^\top\boldsymbol{\Sigma}\mathbf{V}) \le \lambda_p(\boldsymbol{\Sigma}), \quad \lambda_1(\mathbf{V}^\top\boldsymbol{\Sigma}\mathbf{V}) \ge \lambda_1(\boldsymbol{\Sigma})$$

that are true for every orthogonal matrix $\boldsymbol{V}$, and keeping in mind the definition of $\mathtt{P}_\ell$, we get

$$\mathsf{Err}_\ell \le \left\{ 2\omega_{\mathcal{O}} \|\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma}\|_{\mathrm{op}} + \frac{\omega_{\mathcal{O}}^2}{1-\omega_{\mathcal{O}}} \left( (\lambda_p - \lambda_1)(\boldsymbol{\Sigma}) + \frac{\delta_\ell^2}{p_{\ell+1}} \right) \right\}^{1/2} + \left\| \mathtt{P}_\ell \mathbf{V}_\ell^\top \overline{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(\ell)}} \right\|_2,$$

where we have used the notation

$$\omega_{\mathcal{O}} = \max_\ell \frac{|\mathcal{S}^{(\ell)} \cap \mathcal{O}|}{|\mathcal{S}^{(\ell)}|}, \qquad \boldsymbol{\xi}_i = \boldsymbol{X}_i - \boldsymbol{\mu}^*$$

and $\delta_\ell = \inf_{\boldsymbol{\mu}} \max_{i \in \mathcal{S}^{(\ell)}} \|\mathbf{V}_\ell^\top (\boldsymbol{X}_i - \boldsymbol{\mu})\|_2$. Note that when $\mathcal{O}$ and $(\mathcal{S}^{(\ell)} \cap \mathcal{I})^c$ are of cardinality less than $n\varepsilon$ and $n(\varepsilon + \tau)$, respectively, we have

$$\frac{|\mathcal{S}^{(\ell)}|}{|\mathcal{S}^{(\ell)} \cap \mathcal{O}|} = \frac{|\mathcal{S}^{(\ell)} \cap \mathcal{I}| + |\mathcal{S}^{(\ell)} \cap \mathcal{O}|}{|\mathcal{S}^{(\ell)} \cap \mathcal{O}|} = \frac{|\mathcal{S}^{(\ell)} \cap \mathcal{I}|}{|\mathcal{S}^{(\ell)} \cap \mathcal{O}|} + 1 \ge \frac{n(1-\varepsilon-\tau)}{n\varepsilon} + 1 = \frac{1-\tau}{\varepsilon}$$

and, therefore, $\omega_{\mathcal{O}} \le \varepsilon/(1-\tau)$. We set $\eta := \varepsilon + \tau \le 3/4$ and apply Lemma 4 to infer that $\omega_{\mathcal{O}} \le \omega_{\mathcal{O}}/(1-\omega_{\mathcal{O}}) \le \varepsilon/(1-\eta) \le 4\varepsilon$ is true with probability at least $1-\delta$. Furthermore, we know that $\delta_\ell \le \max_{i \in \mathcal{S}^{(\ell)}} \|\mathbf{V}_\ell^\top \boldsymbol{X}_i - \bar{\boldsymbol{\mu}}^{(\ell)}\|_2 \le t\sqrt{p_\ell}$. This yields

$$\mathsf{Err}_\ell \le \left\{ 8\varepsilon \|\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma}\|_{\mathrm{op}} + 16\varepsilon^2 \left( (\lambda_p - \lambda_1)(\boldsymbol{\Sigma}) + \frac{t^2 p_\ell}{p_{\ell+1}} \right) \right\}^{1/2} + \left\| \mathtt{P}_{\mathscr{U}_\ell} \overline{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(\ell)}} \right\|_2.$$

Let us introduce the shorthand

$$T_1 = \max_{\ell \in [L]} \|\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma}\|_{\mathrm{op}} + \varepsilon(\lambda_p - \lambda_1)(\boldsymbol{\Sigma}).$$

This leads to

$$\mathsf{Err}_\ell \le \left\{ 8\varepsilon T_1 + \frac{16\varepsilon^2 t^2 p_\ell}{p_{\ell+1}} \right\}^{1/2} + \left\| \mathtt{P}_{\mathscr{U}_\ell} \overline{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(\ell)}} \right\|_2. \tag{B.10}$$

For the last error term, since $p_L = 1$ we have by Lemma 5

$$\begin{aligned}
\mathsf{Err}_L &\le \left\| \mathtt{P}_{\mathscr{U}_L} \bar{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(L)}} \right\|_2 + \frac{n\varepsilon(t\sqrt{p_L} + \left\| \mathtt{P}_{\mathscr{U}_L} \boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}_{\mathscr{U}_L}^{\mathsf{GM}} \right\|_2)}{|\mathcal{S}^{(L)}|} \\
&\le \left\| \mathtt{P}_{\mathscr{U}_L} \bar{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(L)}} \right\|_2 + \frac{\varepsilon t + \varepsilon \left\| \mathtt{P}_{\mathscr{U}_L} \boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}_{\mathscr{U}_L}^{\mathsf{GM}} \right\|_2}{1-\eta} \\
&\le \left\| \mathtt{P}_{\mathscr{U}_L} \bar{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(L)}} \right\|_2 + 4\varepsilon t + 4\varepsilon \left\| \mathtt{P}_{\mathscr{U}_L} \boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}_{\mathscr{U}_L}^{\mathsf{GM}} \right\|_2.
\end{aligned} \tag{B.11}$$

Combining (B.10), (B.11), inequality $p_\ell \le e p_{\ell+1}$, as well as the Minkowski inequality, we get

$$\begin{aligned}
\big\|\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}\big\|_2 &= \Big\{\sum_{\ell=0}^{L} \mathsf{Err}_\ell^2\Big\}^{1/2} \\
&\le \Big\{8\varepsilon L(T_1 + e\varepsilon t^2) + 16\varepsilon^2\big(t + \big\|\mathsf{P}_{\mathscr{U}_L}\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}^{\mathsf{GM}}_{\mathscr{U}_L}\big\|_2\big)^2\Big\}^{1/2} + \Big\{\sum_{\ell=0}^{L} \|\mathsf{P}_{\mathscr{U}_\ell}\overline{\boldsymbol{\xi}}_{\mathcal{S}_\mathcal{I}^{(\ell)}}\|_2^2\Big\}^{1/2} \\
&\le 2\sqrt{2\varepsilon L T_1} + 9\varepsilon t\sqrt{L} + 4\varepsilon\big\|\mathsf{P}_{\mathscr{U}_L}\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}^{\mathsf{GM}}_{\mathscr{U}_L}\big\|_2 + \Big\{\sum_{\ell=0}^{L} \|\mathsf{P}_{\mathscr{U}_\ell}\overline{\boldsymbol{\xi}}_{\mathcal{S}_\mathcal{I}^{(\ell)}}\|_2^2\Big\}^{1/2}. \qquad \text{(B.12)}
\end{aligned}$$

To ease notation, let us set

$$\mathsf{r}_n = \Big(\frac{2\mathbf{r}_{\boldsymbol{\Sigma}} + 3\log(2/\delta)}{n}\Big)^{1/2}.$$

In view of Lemma 6, with probability at least $1-\delta$, we have

$$\begin{aligned}
\Big\{\sum_{\ell=0}^{L} \|\mathsf{P}_{\mathscr{U}_\ell}\overline{\boldsymbol{\xi}}_{\mathcal{S}_\mathcal{I}^{(\ell)}}\|_2^2\Big\}^{1/2} &\le \Big\{\sum_{\ell=0}^{L}\Big(\frac{\|\mathsf{P}_{\mathscr{U}_\ell}\bar{\boldsymbol{\xi}}_n\|_2}{1-\eta} + \frac{\mathsf{r}_n\sqrt{\eta} + \eta\sqrt{3\log(2e/\eta)}}{1-\eta}\Big)^2\Big\}^{1/2} \\
&\le \Big\{\sum_{\ell=0}^{L}\big(4\|\mathsf{P}_{\mathscr{U}_\ell}\bar{\boldsymbol{\xi}}_n\|_2 + 4\mathsf{r}_n\sqrt{\eta} + 10\eta\sqrt{\log(2/\eta)}\big)^2\Big\}^{1/2} \\
&\le 4\|\bar{\boldsymbol{\xi}}_n\|_2 + 4\mathsf{r}_n\sqrt{\eta L} + 10\eta\sqrt{L\log(2/\eta)}.
\end{aligned}$$

Since the random variable $\|\overline{\boldsymbol{\xi}}_n\|_2^2$ has the same distribution as $\frac{1}{n}\sum_{j=1}^{p}\lambda_j(\boldsymbol{\Sigma})\gamma_j^2$, where $\gamma_1,\dots,\gamma_p$ are i.i.d. standard Gaussian, by (Comminges and Dalalyan, 2012, Lemma 8) we have

$$\|\overline{\boldsymbol{\xi}}_n\|_2^2 \le \frac{2\mathbf{r}_{\boldsymbol{\Sigma}} + 3\log(2/\delta)}{n} = \mathsf{r}_n^2$$

with probability at least $1-\delta$. Therefore, with probability at least $1-2\delta$,

$$\Big(\sum_{\ell=0}^{L} \|\mathsf{P}_{\mathscr{U}_\ell}\overline{\boldsymbol{\xi}}_{\mathcal{S}_\mathcal{I}^{(\ell)}}\|_2^2\Big)^{1/2} \le 4\mathsf{r}_n\big(1+\sqrt{L\eta}\big) + 10\eta\sqrt{L\log(2/\eta)}. \qquad \text{(B.13)}$$

Next, Lemma 4 and the fact that $p_L = \dim(\mathscr{U}_L) = 1$ imply that with probability at least $1-\delta$

$$\big\|\mathsf{P}_{\mathscr{U}_L}\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}^{\mathsf{GM}}_{\mathscr{U}_L}\big\|_2 \le \frac{2(1+\sqrt{2}\mathsf{r}_n)}{1-2\varepsilon}. \qquad \text{(B.14)}$$

Recall that we have chosen $t$ in such a way that

$$t \le \frac{3(1+\sqrt{2}\mathsf{r}_n/\sqrt{\tau})}{1-2\varepsilon^*} + 1.6\sqrt{\log(2/\tau)}. \qquad \text{(B.15)}$$

Combining (B.12), (B.13), (B.14) and (B.15), we arrive at the inequality

$$\begin{aligned}\big\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}-\boldsymbol{\mu}^*\big\|_2 &\le 2\sqrt{2\varepsilon L T_1}+9\varepsilon t\sqrt{L}+\frac{8\varepsilon(1+\sqrt{2}\mathsf{r}_n)}{1-2\varepsilon}+4\mathsf{r}_n\big(1+\sqrt{L\eta}\big)+10\eta\sqrt{L\log(2/\eta)}\\ &\le 2\sqrt{2\varepsilon L T_1}+\frac{27\varepsilon\sqrt{L}(1+\sqrt{2}\mathsf{r}_n/\sqrt{\tau})}{1-2\varepsilon^*}+14.4\varepsilon\sqrt{L\log(2/\tau)}\\ &\quad+\frac{8\varepsilon(1+\sqrt{2}\mathsf{r}_n)}{1-2\varepsilon}+4\mathsf{r}_n\big(1+\sqrt{L\eta}\big)+10\eta\sqrt{L\log(2/\eta)}\end{aligned}$$

that holds with probability at least $1-3\delta$. In the upper bound obtained above, only the term $T_1$ remains random. We can upper bound this term using Lemma 7. It implies that with probability at least $1-2\delta$, we have

$$\begin{aligned}T_1 &\le A\frac{\sqrt{n\mathbf{r}_{\boldsymbol{\Sigma}}}+\mathbf{r}_{\boldsymbol{\Sigma}}+n\eta\log(2e/\eta)+2\log(1/\delta)}{n(1-\eta)}+\big(4\mathsf{r}_n(1+\sqrt{\eta})+10\eta\sqrt{\log(2/\eta)}\big)^2+\varepsilon\\ &\le 2A\big(\sqrt{2}\mathsf{r}_n+\mathsf{r}_n^2+4\eta\log(2/\eta)\big)+\big(7.5\mathsf{r}_n+10\eta\sqrt{\log(2/\eta)}\big)^2+\varepsilon.\end{aligned}$$

Consequently,

$$\begin{aligned}\sqrt{\varepsilon T_1} &\le \big\{2A\varepsilon\big(\sqrt{2}\mathsf{r}_n+\mathsf{r}_n^2+4\eta\log(2/\eta)\big)\big\}^{1/2}+\big(7.5\mathsf{r}_n+10\eta\sqrt{\log(2/\eta)}\big)\sqrt{\varepsilon}+\varepsilon\\ &\le \big\{2A\varepsilon\big(\sqrt{2}\mathsf{r}_n+\mathsf{r}_n^2+4\eta\log(2/\eta)\big)\big\}^{1/2}+5.4\mathsf{r}_n+7.1\eta\sqrt{\log(2/\eta)}+\varepsilon\\ &\le \big\{2A\varepsilon\big(\sqrt{2}\mathsf{r}_n+4\eta\log(2/\eta)\big)\big\}^{1/2}+(\sqrt{A}+5.4)\mathsf{r}_n+7.1\eta\sqrt{\log(2/\eta)}+\varepsilon\\ &\le \sqrt{2\sqrt{2}A\mathsf{r}_n\varepsilon}+2\sqrt{2A\varepsilon\eta\log(2/\eta)}+(\sqrt{A}+5.4)\mathsf{r}_n+7.1\eta\sqrt{\log(2/\eta)}+\varepsilon\\ &\le \varepsilon+A\mathsf{r}_n/\sqrt{2}+2\eta\sqrt{2A\log(2/\eta)}+(\sqrt{A}+5.4)\mathsf{r}_n+7.1\eta\sqrt{\log(2/\eta)}+\varepsilon\\ &\le (A/\sqrt{2}+\sqrt{A}+5.4)\mathsf{r}_n+(7.1+2\sqrt{2A})\tau\sqrt{\log(2/\tau)}+(9.1+2\sqrt{2A})\varepsilon\sqrt{\log(2/\varepsilon)}.\end{aligned}$$

These inequalities imply that there exists a universal constant $\mathsf{C}$ such that

$$\big\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}-\boldsymbol{\mu}^*\big\|_2 \le \frac{\mathsf{C}\big(\mathsf{r}_n+\tau\sqrt{\log(2/\tau)}+\varepsilon\sqrt{\log(2/\varepsilon)}+\mathsf{r}_n\varepsilon/\sqrt{\tau}\big)\sqrt{L}}{1-2\varepsilon^*}. \tag{B.16}$$

We choose

$$\tau=\frac{1}{4}\bigwedge\frac{\bar{\mathsf{r}}_n}{\sqrt{\log_+(2/\bar{\mathsf{r}}_n)}}, \qquad \text{with} \qquad \bar{\mathsf{r}}_n=\frac{\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}}+\sqrt{2\log(2/\delta)}}{\sqrt{n}}.$$

Note that $\mathsf{r}_n\le\sqrt{2}\bar{\mathsf{r}}_n$ and, furthermore, $\tau=1/4$ whenever $\bar{\mathsf{r}}_n\ge 1/2$. Therefore, $\mathsf{r}_n\varepsilon/\sqrt{\tau}\le\mathsf{r}_n+\varepsilon$. Inserting this value of $\tau$ in (B.16) leads to

$$\big\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}-\boldsymbol{\mu}^*\big\|_2 \le \frac{\mathsf{C}\big(\mathsf{r}_n+\varepsilon\sqrt{\log(2/\varepsilon)}\big)\sqrt{L}}{1-2\varepsilon^*}.$$

where $\mathsf{C}$ is a universal constant, the value of which is not necessarily the same in different places where it appears. Replacing $\mathsf{r}_n$ by its expression, and upper bounding $L$ by $2\log p$, we

arrive at

$$\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}} - \boldsymbol{\mu}^*\right\|_2 \leq \frac{\mathsf{C}\sqrt{\log p}}{1-2\varepsilon^*}\left(\sqrt{\frac{\mathbf{r}_{\boldsymbol{\Sigma}}}{n}} + \varepsilon\sqrt{\log(2/\varepsilon)} + \sqrt{\frac{\log(1/\delta)}{n}}\right).$$

Note that this inequality holds true on an event of probability at least $1-5\delta$.

To prove the second part of the theorem, we use Lemma 8 instead of Lemma 7 in order to bound the term $T_1$. Moreover, in the definitions of $\mathsf{r}_n$ and $\bar{\mathsf{r}}_n$ the effective rank $\mathbf{r}_{\boldsymbol{\Sigma}}$ is replaced by the dimension $p$. Then, with probability at least $1-2\delta$, we have

$$\begin{aligned}
T_1 &\leq \frac{5p + n\eta(8\log(2e/\eta)+2) + 7\log(2/\delta)}{n(1-\eta)} + 2\frac{\sqrt{p}+\sqrt{\log(2/\delta)}}{\sqrt{n}(1-\sqrt{\eta})} + \\
&\quad + \left(4\mathsf{r}_n(1+\sqrt{\eta}) + 10\eta\sqrt{\log(2/\eta)}\right)^2 + \varepsilon \\
&\leq (10\mathsf{r}_n^2 + 15\mathsf{r}_n + 72\eta\log(2/\eta)) + \left(7.5\mathsf{r}_n + 10\eta\sqrt{\log(2/\eta)}\right)^2 + \varepsilon.
\end{aligned}$$

Then, repeating the same steps as for the previous case where the effective rank is used instead of dimension we arrive at the following inequality

$$\begin{aligned}
\sqrt{\varepsilon T_1} &\leq \left\{\varepsilon(10\mathsf{r}_n^2 + 15\mathsf{r}_n + 72\eta\log(2/\eta))\right\}^{1/2} + \left(7.5\mathsf{r}_n + 10\eta\sqrt{\log(2/\eta)}\right)\sqrt{\varepsilon} + \varepsilon \\
&\leq 12\mathsf{r}_n + 16\tau\sqrt{\log(2/\tau)} + 18\varepsilon\sqrt{\log(2/\varepsilon)}.
\end{aligned}$$

Combining the obtained inequalities, plugging in the values of $\tau$ and $\mathsf{r}_n$ and bounding $L$ by $2\log p$ we arrive at a final bound which reads as

$$\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}} - \boldsymbol{\mu}^*\right\|_2 \leq \frac{156\sqrt{2\log p}}{1-2\varepsilon^*}\left(\sqrt{\frac{2p}{n}} + \varepsilon\sqrt{\log(2/\varepsilon)} + \sqrt{\frac{3\log(2/\delta)}{n}}\right),$$

which concludes the proof.

## B.2 Proof of Theorem 13

The proof follows the same steps as that of Theorem 11. The assumption $\|\boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2} - \mathbf{I}_p\|_{\mathrm{op}} \leq \gamma$ gives upper and lower bounds on the effective rank of $\widetilde{\boldsymbol{\Sigma}}$ using that of $\boldsymbol{\Sigma}$, which we formulate as a separate lemma. Therefore, the choice of the threshold parameter $\widetilde{t}_\gamma$ stated in Theorem 13 makes Lemmas 3 and 4 applicable to this case as well. To bound the operator norm of $\|\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \widetilde{\boldsymbol{\Sigma}}\|_{\mathrm{op}}$ we make use of the assumption $\|\boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2} - \mathbf{I}_p\|_{\mathrm{op}} \leq \gamma$ and Lemma 7 using triangle inequality. We provide the full proof for reader's convenience.

Before proceeding with the proof we first formulate and prove an auxiliary lemma for bounding the effective rank of $\widetilde{\boldsymbol{\Sigma}}$ using that of $\boldsymbol{\Sigma}$.

**Lemma 9.** *Let $\boldsymbol{\Sigma}$ and $\widetilde{\boldsymbol{\Sigma}}$ be symmetric positive definite matrices such that $\|\boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2} -$*

$\mathbf{I}_p\|_{\mathsf{op}} \le \gamma$. *Then,*

$$\mathsf{r}_{\mathbf{\Sigma}} \cdot \frac{1-\gamma}{1+\gamma} \le \mathsf{r}_{\widetilde{\mathbf{\Sigma}}} \le \mathsf{r}_{\mathbf{\Sigma}} \cdot \frac{1+\gamma}{1-\gamma}.$$

*Proof of Lemma 9.* We start with upper- and lower-bounding the operator norm of $\widetilde{\mathbf{\Sigma}}$. Using triangle inequality we have

$$\big|\|\widetilde{\mathbf{\Sigma}}\|_{\mathrm{op}} - \|\mathbf{\Sigma}\|_{\mathrm{op}}\big| \le \|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\mathrm{op}} \le \|\mathbf{\Sigma}\|_{\mathrm{op}} \cdot \|\mathbf{\Sigma}^{-1/2}\widetilde{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1/2} - \mathbf{I}_p\|_{\mathrm{op}} \le \gamma\|\mathbf{\Sigma}\|_{\mathrm{op}}.$$

This readily yields

$$(1-\gamma)\|\mathbf{\Sigma}\|_{\mathrm{op}} \le \|\widetilde{\mathbf{\Sigma}}\|_{\mathrm{op}} \le (1+\gamma)\|\mathbf{\Sigma}\|_{\mathrm{op}}. \tag{B.17}$$

Moreover, for any pair of positive definite matrices $\mathbf{A}, \mathbf{B} \succeq 0$ the following holds $\mathsf{Tr}(\mathbf{A})\lambda_1(\mathbf{B}) \le \mathsf{Tr}(\mathbf{AB}) \le \mathsf{Tr}(\mathbf{A}) \cdot \|\mathbf{B}\|_{\mathrm{op}}$. Hence, combining the cyclic property of trace, the trace inequality and the fact that the spectrum of the matrix $\mathbf{\Sigma}^{-1/2}\widetilde{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1/2}$ is between $1-\gamma$ and $1+\gamma$, we get both the upper and the lower bounds for $\mathsf{Tr}(\widetilde{\mathbf{\Sigma}})$. The upper bound reads as

$$\begin{aligned}\mathsf{Tr}(\widetilde{\mathbf{\Sigma}}) &= \mathsf{Tr}((\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{-1/2})\widetilde{\mathbf{\Sigma}}(\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{1/2})) = \mathsf{Tr}(\mathbf{\Sigma}(\mathbf{\Sigma}^{-1/2}\widetilde{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1/2})) \\ &\le \|\mathbf{\Sigma}^{-1/2}\widetilde{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1/2}\|_{\mathrm{op}}\,\mathsf{Tr}(\mathbf{\Sigma}) \le (1+\gamma)\,\mathsf{Tr}(\mathbf{\Sigma}).\end{aligned} \tag{B.18}$$

Similarly, the lower bound can be obtained as follows

$$\mathsf{Tr}(\widetilde{\mathbf{\Sigma}}) \ge \lambda_1(\mathbf{\Sigma}^{-1/2}\widetilde{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1/2})\,\mathsf{Tr}(\mathbf{\Sigma}) \ge (1-\gamma)\,\mathsf{Tr}(\mathbf{\Sigma}). \tag{B.19}$$

Therefore, combining (B.17) and (B.19) we get the lower bound for $\mathsf{r}_{\widetilde{\mathbf{\Sigma}}}$, while combining (B.17) and (B.18) yields the upper bound, concluding the proof of the lemma. □

Taking $\mathbf{U}_L := \mathbf{V}_L$, the Algorithm 2 returns $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}} = \sum_{\ell=0}^{L} \widehat{\boldsymbol{\mu}}^{(\ell)}$ with $\widehat{\boldsymbol{\mu}}^{(\ell)} \in \mathscr{U}_\ell = \mathrm{Im}(\mathbf{V}_\ell\mathbf{U}_\ell^\top)$ for every $\ell \in \{0,\dots,L\}$ where the two-by-two orthogonal subspaces $\mathscr{U}_0,\dots,\mathscr{U}_L$ span the whole space $\mathbb{R}^p$. This allows us to decompose the risk:

$$\begin{aligned}\big\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}} - \boldsymbol{\mu}^*\big\|_2^2 &= \sum_{\ell=0}^{L} \big\|\widehat{\boldsymbol{\mu}}^{(\ell)} - \mathtt{P}_{\mathscr{U}_\ell}\boldsymbol{\mu}^*\big\|_2^2 = \sum_{\ell=0}^{L} \big\|\mathtt{P}_{\mathscr{U}_\ell}(\overline{\boldsymbol{X}}_\ell - \boldsymbol{\mu}^*)\big\|_2^2 \\ &= \sum_{\ell=0}^{L} \big\|\mathtt{P}_\ell(\mathbf{V}_\ell^\top\overline{\boldsymbol{X}}_\ell - \mathbf{V}_\ell^\top\boldsymbol{\mu}^*)\big\|_2^2,\end{aligned}$$

where $\mathtt{P}_\ell := \mathbf{U}_\ell^\top\mathbf{U}_\ell$ is the projection matrix projecting onto the subspace of $\mathbb{R}^{p_\ell}$ spanned by the bottom $p_\ell - p_{\ell+1}$ eigenvectors of $\mathbf{V}_\ell^\top(\widehat{\mathbf{\Sigma}}^{(\ell)} - \widetilde{\mathbf{\Sigma}})\mathbf{V}_\ell$ for $\ell = 0,\dots,L$ with the convention that $p_{L+1} = 0$.

For $\ell \in \{0,\dots,L-1\}$, we intend to apply Proposition 6 to $\boldsymbol{Z}_i = \mathbf{V}_\ell^\top\boldsymbol{X}_i$ and $\boldsymbol{\mu}^Z = \mathbf{V}_\ell^\top\boldsymbol{\mu}^*$ in

order to upper bound the error term $\mathrm{Err}_\ell := \|\mathtt{P}_\ell(\mathbf{V}_\ell^\top \overline{\boldsymbol{X}}_\ell - \mathbf{V}_\ell^\top \boldsymbol{\mu}^*)\|_2$. Using the inequalities

$$\|\mathbf{V}^\top(\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \widetilde{\boldsymbol{\Sigma}})\mathbf{V}\|_{\mathrm{op}} \le \|\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \widetilde{\boldsymbol{\Sigma}}\|_{\mathrm{op}}, \quad \lambda_{p_\ell}(\mathbf{V}^\top\widetilde{\boldsymbol{\Sigma}}\mathbf{V}) \le \lambda_p(\widetilde{\boldsymbol{\Sigma}}), \quad \lambda_1(\mathbf{V}^\top\widetilde{\boldsymbol{\Sigma}}\mathbf{V}) \ge \lambda_1(\widetilde{\boldsymbol{\Sigma}})$$

that are true for every orthogonal matrix $\boldsymbol{V}$, and keeping in mind the definition of $\mathtt{P}_\ell$, we get

$$\mathrm{Err}_\ell \le \left\{2\omega_{\mathcal{O}}\|\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \widetilde{\boldsymbol{\Sigma}}\|_{\mathrm{op}} + \frac{\omega_{\mathcal{O}}^2}{1-\omega_{\mathcal{O}}}\left((\lambda_p - \lambda_1)(\widetilde{\boldsymbol{\Sigma}}) + \frac{\delta_\ell^2}{p_{\ell+1}}\right)\right\}^{1/2} + \big\|\mathtt{P}_\ell \mathbf{V}_\ell^\top \overline{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(\ell)}}\big\|_2,$$

where we have used the notation

$$\omega_{\mathcal{O}} = \max_\ell \frac{|\mathcal{S}^{(\ell)} \cap \mathcal{O}|}{|\mathcal{S}^{(\ell)}|}, \qquad \boldsymbol{\xi}_i = \boldsymbol{X}_i - \boldsymbol{\mu}^*$$

and $\delta_\ell = \inf_{\boldsymbol{\mu}} \max_{i\in\mathcal{S}^{(\ell)}} \|\mathbf{V}_\ell^\top(\boldsymbol{X}_i - \boldsymbol{\mu})\|_2$. Note that when $\mathcal{O}$ and $(\mathcal{S}_{\mathcal{I}}^{(\ell)})^c$ are of cardinality less than $n\varepsilon$ and $n(\varepsilon+\tau)$, respectively, we have $\omega_{\mathcal{O}} \le \varepsilon/(1-\tau)$ and

$$\frac{\omega_{\mathcal{O}}}{1-\omega_{\mathcal{O}}} \le \frac{\varepsilon}{1-\varepsilon-\tau}.$$

Since $\mathsf{C}_\gamma \mathbf{r}_{\widetilde{\boldsymbol{\Sigma}}} \ge \mathbf{r}_{\boldsymbol{\Sigma}}$ (by Lemma 9) then Lemma 4 holds for the new threshold $\widetilde{t}_\gamma$ as well. We set $\eta := \varepsilon + \tau \le 3/4$ and apply Lemma 4 to infer that $\omega_{\mathcal{O}} \le \varepsilon/(1-\eta) \le 4\varepsilon$ is true with probability at least $1-\delta$. Furthermore, we know that $\delta_\ell \le \max_{i\in\mathcal{S}^{(\ell)}} \|\mathbf{V}_\ell^\top \boldsymbol{X}_i - \bar{\boldsymbol{\mu}}^{(\ell)}\|_2 \le \widetilde{t}_\gamma\sqrt{p_\ell}$. This yields

$$\mathrm{Err}_\ell \le \left\{8\varepsilon\|\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \widetilde{\boldsymbol{\Sigma}}\|_{\mathrm{op}} + 16\varepsilon^2\left((\lambda_p - \lambda_1)(\widetilde{\boldsymbol{\Sigma}}) + \frac{\widetilde{t}_\gamma^2 p_\ell}{p_{\ell+1}}\right)\right\}^{1/2} + \big\|\mathtt{P}_{\mathscr{U}_\ell}\overline{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(\ell)}}\big\|_2.$$

Let us introduce the shorthand $\widetilde{T}_1 = \max_{\ell\in[L]} \|\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \widetilde{\boldsymbol{\Sigma}}\|_{\mathrm{op}} + \varepsilon(\lambda_p - \lambda_1)(\widetilde{\boldsymbol{\Sigma}})$. This leads to

$$\mathrm{Err}_\ell \le \left\{8\varepsilon\widetilde{T}_1 + \frac{16\varepsilon^2\widetilde{t}_\gamma^2 p_\ell}{p_{\ell+1}}\right\}^{1/2} + \|\mathtt{P}_{\mathscr{U}_\ell}\overline{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(\ell)}}\|_2. \tag{B.20}$$

For the last error term, since $p_L = 1$ we have by Lemma 5

$$\begin{aligned}
\mathrm{Err}_L &\le \big\|\mathtt{P}_{\mathscr{U}_L}\bar{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(L)}}\big\|_2 + \frac{n\varepsilon(\widetilde{t}_\gamma\sqrt{p_L} + \big\|\mathtt{P}_{\mathscr{U}_L}\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}_{\mathscr{U}_L}^{\mathsf{GM}}\big\|_2)}{|\mathcal{S}^{(L)}|} \\
&\le \big\|\mathtt{P}_{\mathscr{U}_L}\bar{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(L)}}\big\|_2 + \frac{\varepsilon\widetilde{t}_\gamma + \varepsilon\big\|\mathtt{P}_{\mathscr{U}_L}\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}_{\mathscr{U}_L}^{\mathsf{GM}}\big\|_2}{1-\eta} \\
&\le \big\|\mathtt{P}_{\mathscr{U}_L}\bar{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(L)}}\big\|_2 + 4\varepsilon\widetilde{t}_\gamma + 4\varepsilon\big\|\mathtt{P}_{\mathscr{U}_L}\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}_{\mathscr{U}_L}^{\mathsf{GM}}\big\|_2.
\end{aligned} \tag{B.21}$$

Combining (B.20), (B.21), inequality $p_\ell \le e p_{\ell+1}$, as well as the Minkowski inequality, we get

$$\begin{aligned}
\big\|\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}\big\|_2 &= \Big\{\sum_{\ell=0}^{L} \mathsf{Err}_\ell^2\Big\}^{1/2} \\
&\le \Big\{8\varepsilon L(\widetilde{T}_1 + e\varepsilon t^2) + 16\varepsilon^2\big(\widetilde{t}_\gamma + \big\|\mathsf{P}_{\mathscr{U}_L}\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}^{\mathsf{GM}}_{\mathscr{U}_L}\big\|_2\big)^2\Big\}^{1/2} + \Big\{\sum_{\ell=0}^{L} \|\mathsf{P}_{\mathscr{U}_\ell}\overline{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(\ell)}}\|_2^2\Big\}^{1/2} \\
&\le 2\sqrt{2\varepsilon L\widetilde{T}_1} + 9\varepsilon\widetilde{t}_\gamma\sqrt{L} + 4\varepsilon\big\|\mathsf{P}_{\mathscr{U}_L}\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}^{\mathsf{GM}}_{\mathscr{U}_L}\big\|_2 + \Big\{\sum_{\ell=0}^{L} \|\mathsf{P}_{\mathscr{U}_\ell}\overline{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(\ell)}}\|_2^2\Big\}^{1/2}. \qquad \text{(B.22)}
\end{aligned}$$

To ease notation, let us set

$$\mathsf{r}_n = \Big(\frac{2\mathbf{r}_{\boldsymbol{\Sigma}} + 3\log(2/\delta)}{n}\Big)^{1/2}.$$

In view of Lemma 6, with probability at least $1-\delta$, we have

$$\begin{aligned}
\Big\{\sum_{\ell=0}^{L} \|\mathsf{P}_{\mathscr{U}_\ell}\overline{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(\ell)}}\|_2^2\Big\}^{1/2} &\le \Big\{\sum_{\ell=0}^{L}\Big(\frac{\|\mathsf{P}_{\mathscr{U}_\ell}\bar{\boldsymbol{\xi}}_n\|_2}{1-\eta} + \frac{\mathsf{r}_n\sqrt{\eta} + \eta\sqrt{3\log(2e/\eta)}}{1-\eta}\Big)^2\Big\}^{1/2} \\
&\le \Big\{\sum_{\ell=0}^{L}\big(4\|\mathsf{P}_{\mathscr{U}_\ell}\bar{\boldsymbol{\xi}}_n\|_2 + 4\mathsf{r}_n\sqrt{\eta} + 10\eta\sqrt{\log(2/\eta)}\big)^2\Big\}^{1/2} \\
&\le 4\|\bar{\boldsymbol{\xi}}_n\|_2 + 4\mathsf{r}_n\sqrt{\eta L} + 10\eta\sqrt{L\log(2/\eta)}.
\end{aligned}$$

Since $\|\overline{\boldsymbol{\xi}}_n\|_2^2$ has the same distribution as $\frac{1}{n}\sum_{j=1}^{p}\lambda_j(\boldsymbol{\Sigma})\gamma_j^2$, where $\gamma_1,\dots,\gamma_p$ are i.i.d. standard Gaussian, by (Comminges and Dalalyan, 2012, Lemma 8) we have

$$\|\overline{\boldsymbol{\xi}}_n\|_2^2 \le \frac{2\mathbf{r}_{\boldsymbol{\Sigma}} + 3\log(2/\delta)}{n} = \mathsf{r}_n^2$$

with probability at least $1-\delta$. Therefore, with probability at least $1-2\delta$,

$$\Big(\sum_{\ell=0}^{L} \|\mathsf{P}_{\mathscr{U}_\ell}\overline{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(\ell)}}\|_2^2\Big)^{1/2} \le 4\mathsf{r}_n\big(1+\sqrt{L\eta}\big) + 10\eta\sqrt{L\log(2/\eta)}. \qquad \text{(B.23)}$$

Next, Lemma 4 and the fact that $p_L = \dim(\mathscr{U}_L) = 1$ imply that with probability at least $1-\delta$

$$\big\|\mathsf{P}_{\mathscr{U}_L}\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}^{\mathsf{GM}}_{\mathscr{U}_L}\big\|_2 \le \frac{2(1+\sqrt{2}\mathsf{r}_n)}{1-2\varepsilon}. \qquad \text{(B.24)}$$

Recall that we have chosen $\widetilde{t}_\gamma$ in such a way that

$$\widetilde{t}_\gamma \le \frac{3(1+\mathsf{C}_\gamma\sqrt{2}\mathsf{r}_n/\sqrt{\tau})}{1-2\varepsilon^*} + 1.6\sqrt{\log(2/\tau)}. \qquad \text{(B.25)}$$

Combining (B.22), (B.23), (B.24) and (B.25), we arrive at the inequality

$$\begin{aligned}\big\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}-\boldsymbol{\mu}^*\big\|_2 &\le 2\sqrt{2\varepsilon L\widetilde{T}_1}+9\varepsilon\widetilde{t}_\gamma\sqrt{L}+\frac{8\varepsilon(1+\mathsf{C}_\gamma\sqrt{2}\mathsf{r}_n)}{1-2\varepsilon}+4\mathsf{r}_n\big(1+\sqrt{L\eta}\big)+10\eta\sqrt{L\log(2/\eta)}\\ &\le 2\sqrt{2\varepsilon L\widetilde{T}_1}+\frac{27\varepsilon\sqrt{L}(1+\mathsf{C}_\gamma\sqrt{2}\mathsf{r}_n/\sqrt{\tau})}{1-2\varepsilon^*}+14.4\varepsilon\sqrt{L\log(2/\tau)}\\ &\quad+\frac{8\varepsilon(1+\mathsf{C}_\gamma\sqrt{2}\mathsf{r}_n)}{1-2\varepsilon}+4\mathsf{r}_n\big(1+\sqrt{L\eta}\big)+10\eta\sqrt{L\log(2/\eta)}\end{aligned}$$

that holds with probability at least $1-3\delta$. In the upper bound obtained above, only the term $\widetilde{T}_1$ remains random. To bound $\widetilde{T}_1$ we first apply a triangle inequality then use Lemma 7. It implies that with probability at least $1-2\delta$, we have

$$\begin{aligned}\widetilde{T}_1 &\le A\frac{\sqrt{n\mathbf{r}_{\boldsymbol{\Sigma}}}+\mathbf{r}_{\boldsymbol{\Sigma}}+n\eta\log(2e/\eta)+2\log(1/\delta)}{n(1-\eta)}+\big(4\mathsf{r}_n(1+\sqrt{\eta})+10\eta\sqrt{\log(2/\eta)}\big)^2+(1+\gamma)\varepsilon+\gamma\\ &\le 2A\big(\sqrt{2}\mathsf{r}_n+\mathsf{r}_n^2+4\eta\log(2/\eta)\big)+\big(7.5\mathsf{r}_n+10\eta\sqrt{\log(2/\eta)}\big)^2+\varepsilon+2\gamma.\end{aligned}$$

Consequently,

$$\begin{aligned}\sqrt{\varepsilon\widetilde{T}_1} &\le \big\{2A\varepsilon\big(\sqrt{2}\mathsf{r}_n+\mathsf{r}_n^2+4\eta\log(2/\eta)\big)\big\}^{1/2}+\big(7.5\mathsf{r}_n+10\eta\sqrt{\log(2/\eta)}\big)\sqrt{\varepsilon}+\varepsilon+\sqrt{2\varepsilon\gamma}\\ &\le \big\{2A\varepsilon\big(\sqrt{2}\mathsf{r}_n+\mathsf{r}_n^2+4\eta\log(2/\eta)\big)\big\}^{1/2}+5.4\mathsf{r}_n+7.1\eta\sqrt{\log(2/\eta)}+\varepsilon+\sqrt{2\varepsilon\gamma}\\ &\le \big\{2A\varepsilon\big(\sqrt{2}\mathsf{r}_n+4\eta\log(2/\eta)\big)\big\}^{1/2}+(\sqrt{A}+5.4)\mathsf{r}_n+7.1\eta\sqrt{\log(2/\eta)}+\varepsilon+\sqrt{2\varepsilon\gamma}\\ &\le \sqrt{2\sqrt{2}A\mathsf{r}_n\varepsilon}+2\sqrt{2A\varepsilon\eta\log(2/\eta)}+(\sqrt{A}+5.4)\mathsf{r}_n+7.1\eta\sqrt{\log(2/\eta)}+\varepsilon+\sqrt{2\varepsilon\gamma}\\ &\le \varepsilon+A\mathsf{r}_n/\sqrt{2}+2\eta\sqrt{2A\log(2/\eta)}+(\sqrt{A}+5.4)\mathsf{r}_n+7.1\eta\sqrt{\log(2/\eta)}+\varepsilon+\sqrt{2\varepsilon\gamma}\\ &\le (A/\sqrt{2}+\sqrt{A}+5.4)\mathsf{r}_n+(7.1+2\sqrt{2A})\tau\sqrt{\log(2/\tau)}+(9.1+2\sqrt{2A})\varepsilon\sqrt{\log(2/\varepsilon)}+\sqrt{2\varepsilon\gamma}.\end{aligned}$$

These inequalities imply that there exists a universal constant $\mathsf{C}$ such that

$$\big\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}-\boldsymbol{\mu}^*\big\|_2 \le \frac{\mathsf{C}\big(\mathsf{C}_\gamma\mathsf{r}_n+\tau\sqrt{\log(2/\tau)}+\varepsilon\sqrt{\log(2/\varepsilon)}+\mathsf{r}_n\varepsilon/\sqrt{\tau}+\sqrt{\varepsilon\gamma}\big)\sqrt{L}}{1-2\varepsilon^*}. \tag{B.26}$$

Let us denote $\log_+(x)=\max\{0,\log(x)\}$ the positive part of logarithm, then we choose

$$\tau=\frac{1}{4}\bigwedge\frac{\widetilde{\mathsf{r}}_n}{\sqrt{\log_+(2/\widetilde{\mathsf{r}}_n)}}, \qquad \text{with} \qquad \widetilde{\mathsf{r}}_n=\frac{\sqrt{\mathsf{C}_\gamma\mathbf{r}_{\widetilde{\boldsymbol{\Sigma}}}}+\sqrt{2\log(2/\delta)}}{\sqrt{n}}.$$

Note that $\mathsf{r}_n\le\sqrt{2}\widetilde{\mathsf{r}}_n$ and, furthermore, $\tau=1/4$ whenever $\widetilde{\mathsf{r}}_n\ge 1/2$. Therefore, $\mathsf{r}_n\varepsilon/\sqrt{\tau}\le\mathsf{r}_n+\varepsilon$. Inserting this value of $\tau$ in (B.26) leads to

$$\big\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}-\boldsymbol{\mu}^*\big\|_2 \le \frac{\mathsf{C}\big(\mathsf{C}_\gamma\mathsf{r}_n+\varepsilon\sqrt{\log(2/\varepsilon)}+\sqrt{\varepsilon\gamma}\big)\sqrt{L}}{1-2\varepsilon^*}.$$

where $\mathsf{C}$ is a universal constant, the value of which is not necessarily the same in different places where it appears. Replacing $\mathsf{r}_n$ by its expression, upper bounding $L$ by $2\log p$, and

using the fact that $\mathsf{C}_\gamma \le 3$ for $\gamma \in (0, 1/2]$ we arrive at

$$\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}} - \boldsymbol{\mu}^*\|_2 \le \frac{\mathsf{C}\sqrt{\log p}}{1-2\varepsilon^*}\left(\sqrt{\frac{\mathbf{r}_{\boldsymbol{\Sigma}}}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} + \varepsilon\sqrt{\log(2/\varepsilon)} + \sqrt{\varepsilon\gamma}\right).$$

Note that this inequality holds true on an event of probability at least $1 - 5\delta$.

## B.3 Extension to Sub-Gaussian distributions

This section is devoted to the proof of Theorem 12, which is an extension of Theorem 11 to the case when the $1-\varepsilon$ portion of observations are sub-Gaussian. First, we formulate some technical lemmas necessary for the proof of Theorem 12 postponing the full proof to the end of the present section.

Recall that a random vector $\boldsymbol{\zeta}$ with zero mean and identity covariance matrix is sub-Gaussian with variance proxy $\mathfrak{s} > 0$, $\boldsymbol{\zeta} \sim \mathsf{SG}_p(\mathfrak{s})$, if

$$\mathbb{E}[e^{\boldsymbol{v}^\top \boldsymbol{\zeta}}] \le \exp\left\{\frac{\mathfrak{s}}{2}\|\boldsymbol{v}\|^2\right\}, \quad \forall \boldsymbol{v} \in \mathbb{R}^p.$$

The concentration inequality for sub-Gaussian vectors is a well-known fact (see, e.g. (Rigollet and Hütter, 2019), Theorem 1.19) that if $\boldsymbol{\zeta} \sim \mathsf{SG}_p(\mathfrak{s})$ then for all $\delta \in (0,1)$, it holds

$$\mathbf{P}\big(\|\boldsymbol{\zeta}\|_2 \le 4\sqrt{p\mathfrak{s}} + \sqrt{8\mathfrak{s}\log(1/\delta)}\big) \ge 1-\delta. \tag{B.27}$$

In the definition of $\mathsf{SGAC}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \mathfrak{s}, \varepsilon)$ we assume that the $1-\varepsilon$ portion of the data $\{\boldsymbol{X}_i\}_{i=1}^n$ are sub-Gaussian, that is $\boldsymbol{X}_i = \boldsymbol{\mu}^* + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\zeta}_i$ for all $i \in \mathcal{I}$, where the set $\mathcal{I}$ is of cardinality at least $(1-\varepsilon)n$. Denote $\boldsymbol{\xi}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\zeta}_i$ for all $i = 1, \dots, n$ and assume that $\|\boldsymbol{\Sigma}\|_{\mathsf{op}} = 1$.

First, we show that with the choice of threshold parameter the analogous to Lemma 2, Lemma 3, Lemma 4 lemmas hold true. Notice that all three lemmas are using the concentration bound for the operator norm of (sub)-Gaussian vectors. In the case of Gaussian vectors we make use of (Vershynin, 2012, Corollary 5.35), while the analogous result for the sub-Gaussian distributions is also known (Vershynin, 2012, Theorem 5.39). For the readers convenience the latter is formulated in Lemma 1.

**Lemma 10.** *Let $J \subset \{1, \dots, n\}$ be a subset of cardinality $m$. For every $\delta \in (0,1)$, it holds that*

$$\mathbf{P}\left(\left\|\sum_{j\in J}\boldsymbol{\xi}_j\right\|_2 \le 4\sqrt{pm\mathfrak{s}} + \sqrt{8m\mathfrak{s}\log(1/\delta)}\right) \ge 1-\delta.$$

*Proof of Lemma 10.* Without loss of generality, we assume that $J = \{1, \dots, m\}$. On the one hand, $\|\boldsymbol{\Sigma}\|_{\mathsf{op}} = 1$ implies that

$$\left\|\sum_{i=1}^m \boldsymbol{\xi}_i\right\|_2 \le \left\|\sum_{i=1}^m \boldsymbol{\zeta}_i\right\|_2.$$

On the other hand, $\zeta_1+\ldots+\zeta_m \sim \mathrm{SG}_p(m\mathfrak{s})$. Applying inequality (B.27) to this random variable yields the desired result. □

We now state the versions of Lemma 2 and Lemma 4 that are valid in the setting of sub-Gaussian vectors. Notice that the only difference is in the choice of the threshold; thanks to Lemma 1, the threshold now includes the universal constant $\mathsf{C}_0$ and the variance proxy $\mathfrak{s}$. The proofs of these lemmas are omitted, since they are the same as in the Gaussian case presented in Appendix B.1.3 (except for bounding the operator norm of a matrix with sub-Gaussian columns we use Lemma 1).

**Lemma 11.** *With probability at least $1-\delta$, for all linear subspaces $V \subset \mathbb{R}^p$, we have*

$$\frac{\|\widehat{\boldsymbol{\mu}}_V^{\mathsf{GM}} - \mathsf{P}_V \boldsymbol{\mu}^*\|_2}{\sqrt{\dim(V)}} \le \frac{2\sqrt{\|\boldsymbol{\Sigma}\|_{\mathrm{op}}}}{1-2\varepsilon}\left(1 + \frac{\mathsf{C}_0\big(\mathfrak{s}\sqrt{p} + 2\mathfrak{s}\sqrt{\log(1/\delta)}\big)}{\sqrt{n}}\right),$$

*where the constant $\mathsf{C}_0$ is the same constant as in Lemma 1.*

**Lemma 12.** *Let $\tau$ and $\delta$ be two numbers from $(0,1)$. Define*

$$z = 1 + \frac{\mathsf{C}_0\mathfrak{s}(\sqrt{p} + \sqrt{2\log(1/\delta)})}{\sqrt{n\tau}} + \mathsf{C}_0\mathfrak{s}\sqrt{2 + 2\log\big(1/\tau\big)}$$

*with the same constant $\mathsf{C}_0$ as in Lemma 1. Then, with probability at least $1-\delta$, we have*

$$\sup_V \sum_{i=1}^n \mathbb{1}\big(\|\mathsf{P}_V \boldsymbol{\xi}_i\|_2^2 > z^2 \dim(V)\big) \le n\tau,$$

*where the supremum is over all linear subspaces $V$ of $\mathbb{R}^p$.*

**Lemma 13** (Koltchinskii and Lounici (2017), Theorem 9)**.** *There is a constant $A_3 > 0$ depending only on the variance proxy $\tau$ such that for every pair of integers $n \ge 1$ and $p \ge 1$, we have*

$$\mathbf{P}\left(\|\boldsymbol{\zeta}_{1:n}\boldsymbol{\zeta}_{1:n}^\top - n\boldsymbol{\Sigma}\|_{\mathsf{op}} \ge A_3\Big(\sqrt{(p+t)n} + p + t\Big)\right) \le e^{-t}, \quad \forall t \ge 1.$$

**Lemma 14.** *There exists a positive constant $A$ such that, for any positive integer $m \le n$ and any $t \ge 1$, with probability at least $1-2e^{-t}$, the inequality*

$$\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}} - \boldsymbol{\Sigma}\|_{\mathrm{op}} \le A\frac{\sqrt{np} + p + m\log(2ne/m) + 2t}{n-m} + \big\|\overline{\boldsymbol{\xi}}_{\mathcal{S}}\big\|_2^2$$

*is satisfied for every $\mathcal{S} \subset [n]$ of cardinality $\ge n-m$.*

*Proof.* The proof of this lemma is similar to the proof of Lemma 7 with the only difference that instead of Theorems 4 and 5 from (Koltchinskii and Lounici, 2017) we now use Lemma 13. □

**Lemma 15.** *For any positive integer $m \leq n$ and any $t > 0$, with probability at least $1 - e^{-t}$, we have*

$$\max_{|S| \geq n-m} \left\| \frac{1}{|S|} \sum_{i \in S} \mathrm{P}\, \boldsymbol{\xi}_i \right\|_2 \leq \frac{n \|\mathrm{P}\bar{\boldsymbol{\xi}}_n\|_2}{n-m} + \frac{\sqrt{m\mathfrak{s}}(4\sqrt{p} + 2\sqrt{2t}) + 2m\sqrt{2\mathfrak{s}\log(2ne/m)}}{n-m}.$$

*Proof.* The proof of this theorem is similar to that of Lemma 6, with the only exception that now we need to bound the maximum of a norm of a sum of at most $m$ sub-Gaussian vectors, where the maximum is taken over all subsets of $[n]$ of size at most $m$. Since, each sub-Gaussian vector has a variance proxy $\mathfrak{s}$ then using Lemma 10 along with union bound, we have

$$\begin{aligned}
\mathbf{P}\Big( \max_{|J| \leq m} \Big\| \sum_{i \in J} \boldsymbol{\xi}_i \Big\|_2 \geq \sqrt{m\mathfrak{s}}(4\sqrt{p} + t_m) \Big) &\leq \sum_{l=1}^{m} \binom{n}{l} \mathbf{P}\Big( \Big\| \sum_{i=1}^{l} \boldsymbol{\xi}_i \Big\|_2 \geq \sqrt{m\mathfrak{s}}(4\sqrt{p} + t_m) \Big) \\
&\leq \sum_{l=1}^{m} \Big(\frac{ne}{l}\Big)^l \mathbf{P}\Big( \Big\| \sum_{i=1}^{l} \boldsymbol{\xi}_i \Big\|_2 \geq \sqrt{l\mathfrak{s}}(4\sqrt{p} + t_l) \Big) \\
&\leq \sum_{l=1}^{m} \Big(\frac{ne}{l}\Big)^l e^{-t_l/3} \leq e^{-t}.
\end{aligned}$$

Therefore, we obtain that with probability at least $1 - e^{-t}$ the inequality

$$\max_{|J| \leq m} \Big\| \sum_{i \in J} \boldsymbol{\xi}_i \Big\|_2 \leq \sqrt{m\mathfrak{s}}(4\sqrt{p} + 2\sqrt{2t}) + 2m\sqrt{2\mathfrak{s}\log(2ne/m)}$$

holds. Then, combining with

$$\frac{1}{|S|} \Big\| \sum_{i \in S} \mathrm{P}\boldsymbol{\xi}_i \Big\|_2 \leq \frac{n \|\mathrm{P}\bar{\boldsymbol{\xi}}_n\|_2}{n-m} + \frac{1}{n-m} \max_{|J| \leq m} \Big\| \sum_{i \in J} \boldsymbol{\xi}_i \Big\|_2.$$

yields the desired result.

□

### B.3.1 Proof of Theorem 12

All the ingredients provided, we can now compile the complete proof of Theorem 12.

Taking $\mathbf{U}_L := \mathbf{V}_L$, the algorithm detailed in (3.2) returns $\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}} = \sum_{\ell=0}^{L} \widehat{\boldsymbol{\mu}}^{(\ell)}$ with $\widehat{\boldsymbol{\mu}}^{(\ell)} \in \mathscr{U}_\ell = \mathrm{Im}(\mathbf{V}_\ell \mathbf{U}_\ell^\top)$ for every $\ell \in \{0, \dots, L\}$ where the two-by-two orthogonal subspaces $\mathscr{U}_0, \dots, \mathscr{U}_L$ span the whole space $\mathbb{R}^p$. This allows us to decompose the risk:

$$\big\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}} - \boldsymbol{\mu}^*\big\|_2^2 = \sum_{\ell=0}^{L} \big\|\mathrm{P}_{\mathscr{U}_\ell}(\overline{\boldsymbol{X}}_\ell - \boldsymbol{\mu}^*)\big\|_2^2 = \sum_{\ell=0}^{L} \big\|\mathrm{P}_\ell(\mathbf{V}_\ell^\top \overline{\boldsymbol{X}}_\ell - \mathbf{V}_\ell^\top \boldsymbol{\mu}^*)\big\|_2^2,$$

where $\mathrm{P}_\ell := \mathbf{U}_\ell^\top \mathbf{U}_\ell$ is the projection matrix projecting onto the subspace of $\mathbb{R}^{p_\ell}$ spanned by the bottom $p_\ell - p_{\ell+1}$ eigenvectors of $\mathbf{V}_\ell^\top(\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma})\mathbf{V}_\ell$ for $\ell = 0, \dots, L$ with the convention that

$p_{L+1} = 0$.

For $\ell \in \{0, \dots, L-1\}$, we intend to apply Proposition 6 to $\boldsymbol{Z}_i = \mathbf{V}_\ell^\top \boldsymbol{X}_i$ and $\boldsymbol{\mu}^Z = \mathbf{V}_\ell^\top \boldsymbol{\mu}^*$ in order to upper bound the error term $\mathrm{Err}_\ell := \|\mathtt{P}_\ell(\mathbf{V}_\ell^\top \overline{\boldsymbol{X}}_\ell - \mathbf{V}_\ell^\top \boldsymbol{\mu}^*)\|_2$. Using the inequalities

$$\|\mathbf{V}^\top(\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma})\mathbf{V}\|_{\mathrm{op}} \le \|\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma}\|_{\mathrm{op}}, \quad \lambda_{p_\ell}(\mathbf{V}^\top \boldsymbol{\Sigma}\mathbf{V}) \le \lambda_p(\boldsymbol{\Sigma}), \quad \lambda_1(\mathbf{V}^\top \boldsymbol{\Sigma}\mathbf{V}) \ge \lambda_1(\boldsymbol{\Sigma})$$

that are true for every orthogonal matrix $\boldsymbol{V}$, and keeping in mind the definition of $\mathtt{P}_\ell$, we get

$$\mathrm{Err}_\ell \le \left\{ 2\omega_{\mathcal{O}} \|\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma}\|_{\mathrm{op}} + \frac{\omega_{\mathcal{O}}^2}{1-\omega_{\mathcal{O}}} \left( (\lambda_p - \lambda_1)(\boldsymbol{\Sigma}) + \frac{\delta_\ell^2}{p_{\ell+1}} \right) \right\}^{1/2} + \left\| \mathtt{P}_\ell \mathbf{V}_\ell^\top \overline{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(\ell)}} \right\|_2,$$

where we have used the notation

$$\omega_{\mathcal{O}} = \max_\ell \frac{|\mathcal{S}^{(\ell)} \cap \mathcal{O}|}{|\mathcal{S}^{(\ell)}|}, \qquad \boldsymbol{\xi}_i = \boldsymbol{X}_i - \boldsymbol{\mu}^*$$

and $\delta_\ell = \inf_{\boldsymbol{\mu}} \max_{i \in \mathcal{S}^{(\ell)}} \|\mathbf{V}_\ell^\top (\boldsymbol{X}_i - \boldsymbol{\mu})\|_2$. Note that when $\mathcal{O}$ and $(\mathcal{S}_{\mathcal{I}}^{(\ell)})^c$ are of cardinality less than $n\varepsilon$ and $n(\varepsilon + \tau)$, respectively, we have $\omega_{\mathcal{O}} \le \varepsilon/(1-\tau)$ and $\frac{\omega_{\mathcal{O}}}{1-\omega_{\mathcal{O}}} \le \frac{\varepsilon}{1-\varepsilon-\tau}$.

We set $\eta := \varepsilon + \tau \le 3/4$ and apply Lemma 12 to infer that $\omega_{\mathcal{O}} \le \varepsilon/(1-\eta) \le 4\varepsilon$ is true with probability at least $1-\delta$. Furthermore, we know that $\delta_\ell \le \max_{i \in \mathcal{S}^{(\ell)}} \|\mathbf{V}_\ell^\top \boldsymbol{X}_i - \bar{\boldsymbol{\mu}}^{(\ell)}\|_2 \le t\sqrt{p_\ell}$. This yields

$$\mathrm{Err}_\ell \le \left\{ 8\varepsilon \|\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma}\|_{\mathrm{op}} + 16\varepsilon^2 \left( (\lambda_p - \lambda_1)(\boldsymbol{\Sigma}) + \frac{t^2 p_\ell}{p_{\ell+1}} \right) \right\}^{1/2} + \left\| \mathtt{P}_{\mathscr{U}_\ell} \overline{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(\ell)}} \right\|_2.$$

Let us introduce the shorthand $T_1 = \max_{\ell \in [L]} \|\widehat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma}\|_{\mathrm{op}} + \varepsilon(\lambda_p - \lambda_1)(\boldsymbol{\Sigma})$. This leads to

$$\mathrm{Err}_\ell \le \left\{ 8\varepsilon T_1 + \frac{16\varepsilon^2 t^2 p_\ell}{p_{\ell+1}} \right\}^{1/2} + \left\| \mathtt{P}_{\mathscr{U}_\ell} \overline{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(\ell)}} \right\|_2. \tag{B.28}$$

For the last error term, since $p_L = 1$ then, by the combination of Lemma 11 and Lemma 12, we have

$$\begin{aligned}
\mathrm{Err}_L &\le \left\| \mathtt{P}_{\mathscr{U}_L} \bar{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(L)}} \right\|_2 + \frac{n\varepsilon(t\sqrt{p_L} + \left\| \mathtt{P}_{\mathscr{U}_L} \boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}_{\mathscr{U}_L}^{\mathsf{GM}} \right\|_2)}{|\mathcal{S}^{(L)}|} \\
&\le \left\| \mathtt{P}_{\mathscr{U}_L} \bar{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(L)}} \right\|_2 + \frac{\varepsilon t + \varepsilon \left\| \mathtt{P}_{\mathscr{U}_L} \boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}_{\mathscr{U}_L}^{\mathsf{GM}} \right\|_2}{1-\eta} \\
&\le \left\| \mathtt{P}_{\mathscr{U}_L} \bar{\boldsymbol{\xi}}_{\mathcal{S}_{\mathcal{I}}^{(L)}} \right\|_2 + 4\varepsilon t + 4\varepsilon \left\| \mathtt{P}_{\mathscr{U}_L} \boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}_{\mathscr{U}_L}^{\mathsf{GM}} \right\|_2.
\end{aligned} \tag{B.29}$$

Combining (B.28), (B.29), inequality $p_\ell \le e p_{\ell+1}$, as well as the Minkowski inequality, we get

$$\begin{aligned}
\big\|\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}\big\|_2 &= \bigg\{\sum_{\ell=0}^{L} \mathsf{Err}_\ell^2\bigg\}^{1/2} \\
&\le \bigg\{8\varepsilon L(T_1 + e\varepsilon t^2) + 16\varepsilon^2\big(t + \big\|\mathsf{P}_{\mathscr{U}_L}\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}^{\mathsf{GM}}_{\mathscr{U}_L}\big\|_2\big)^2\bigg\}^{1/2} + \bigg\{\sum_{\ell=0}^{L} \|\mathsf{P}_{\mathscr{U}_\ell}\overline{\boldsymbol{\xi}}_{\mathcal{S}_\mathcal{I}^{(\ell)}}\|_2^2\bigg\}^{1/2} \\
&\le 2\sqrt{2\varepsilon L T_1} + 9\varepsilon t\sqrt{L} + 4\varepsilon\big\|\mathsf{P}_{\mathscr{U}_L}\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}^{\mathsf{GM}}_{\mathscr{U}_L}\big\|_2 + \bigg\{\sum_{\ell=0}^{L} \|\mathsf{P}_{\mathscr{U}_\ell}\overline{\boldsymbol{\xi}}_{\mathcal{S}_\mathcal{I}^{(\ell)}}\|_2^2\bigg\}^{1/2}. \qquad \text{(B.30)}
\end{aligned}$$

To ease notation, let us set

$$\mathsf{r}_{n,\mathfrak{s}} = \frac{4\sqrt{\mathfrak{s}}\big(\sqrt{p} + 2\sqrt{\log(2/\delta)}\big)}{\sqrt{n}}.$$

In view of Lemma 15, with probability at least $1-\delta$, we have

$$\begin{aligned}
\bigg\{\sum_{\ell=0}^{L} \|\mathsf{P}_{\mathscr{U}_\ell}\overline{\boldsymbol{\xi}}_{\mathcal{S}_\mathcal{I}^{(\ell)}}\|_2^2\bigg\}^{1/2} &\le \bigg\{\sum_{\ell=0}^{L}\bigg(\frac{\|\mathsf{P}_{\mathscr{U}_\ell}\bar{\boldsymbol{\xi}}_n\|_2}{1-\eta} + \frac{\mathsf{r}_{n,\mathfrak{s}}\sqrt{\eta} + 2\eta\sqrt{2\log(2e/\eta)}}{1-\eta}\bigg)^2\bigg\}^{1/2} \\
&\le \bigg\{\sum_{\ell=0}^{L}\big(4\|\mathsf{P}_{\mathscr{U}_\ell}\bar{\boldsymbol{\xi}}_n\|_2 + 4\mathsf{r}_{n,\mathfrak{s}}\sqrt{\eta} + 10\eta\sqrt{\log(2/\eta)}\big)^2\bigg\}^{1/2} \\
&\le 4\|\overline{\boldsymbol{\xi}}_n\|_2 + 4\mathsf{r}_{n,\mathfrak{s}}\sqrt{\eta L} + 10\eta\sqrt{L\log(2/\eta)}.
\end{aligned}$$

Moreover, since the random variable $\overline{\boldsymbol{\xi}}_n$ is sub-Gaussian with variance proxy $\mathfrak{s}/n$, hence by Lemma 8 we have

$$\|\overline{\boldsymbol{\xi}}_n\|_2^2 \le \frac{16\mathfrak{s}\big(\sqrt{p} + 2\sqrt{\log(2/\delta)}\big)^2}{n} = \mathsf{r}_{n,\mathfrak{s}}^2$$

with probability at least $1-\delta$. Therefore, with probability at least $1-2\delta$,

$$\bigg(\sum_{\ell=0}^{L} \|\mathsf{P}_{\mathscr{U}_\ell}\overline{\boldsymbol{\xi}}_{\mathcal{S}_\mathcal{I}^{(\ell)}}\|_2^2\bigg)^{1/2} \le 4\mathsf{r}_{n,\mathfrak{s}}\big(1 + \sqrt{L\eta}\big) + 10\eta\sqrt{L\log(2/\eta)}. \qquad \text{(B.31)}$$

Next, the combination of Lemma 11 and Lemma 12 and the fact that $p_L = \dim(\mathscr{U}_L) = 1$ imply that with probability at least $1-\delta$

$$\big\|\mathsf{P}_{\mathscr{U}_L}\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}^{\mathsf{GM}}_{\mathscr{U}_L}\big\|_2 \le \frac{2(1 + \mathsf{C}\mathsf{r}_{n,\mathfrak{s}}\sqrt{\mathfrak{s}}/4)}{1-2\varepsilon}, \qquad \text{(B.32)}$$

where $\mathsf{C}$ is the same universal constant as in Lemma 1. Recall that we have chosen $t$ in such a way that

$$t \le \frac{3(1 + \mathsf{C}_0\mathsf{r}_{n,\mathfrak{s}}\sqrt{\mathfrak{s}}/4\sqrt{\tau})}{1-2\varepsilon^*} + 1.6\mathsf{C}_0\mathfrak{s}\sqrt{\log(2/\tau)}. \qquad \text{(B.33)}$$

Combining (B.30), (B.31), (B.32) and (B.33), we arrive at the inequality

$$\begin{aligned}\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}-\boldsymbol{\mu}^*\right\|_2 &\le 2\sqrt{2\varepsilon L T_1}+9\varepsilon t\sqrt{L}+\frac{8\varepsilon(1+\mathsf{C}\mathsf{r}_{n,\mathfrak{s}}\sqrt{\mathfrak{s}}/4)}{1-2\varepsilon}+4\mathsf{r}_{n,\mathfrak{s}}\big(1+\sqrt{L\eta}\big)+10\eta\sqrt{L\log(2/\eta)}\\ &\le 2\sqrt{2\varepsilon L T_1}+\frac{27\varepsilon\sqrt{L}(1+\mathsf{C}\mathsf{r}_{n,\mathfrak{s}}\sqrt{\mathfrak{s}}/4\sqrt{\tau})}{1-2\varepsilon^*}+14.4\mathsf{C}\mathfrak{s}\varepsilon\sqrt{L\log(2/\tau)}\\ &\quad+\frac{8\varepsilon(1+\mathsf{C}\mathsf{r}_{n,\mathfrak{s}}\sqrt{\mathfrak{s}}/4)}{1-2\varepsilon}+4\mathsf{r}_{n,\mathfrak{s}}\big(1+\sqrt{L\eta}\big)+10\eta\sqrt{L\log(2/\eta)}\end{aligned}$$

that holds with probability at least $1-3\delta$. In the upper bound obtained above, only the term $T_1$ remains random. We can upper bound this term using Lemma 14. It implies that with probability at least $1-2\delta$, we have

$$\begin{aligned}T_1 &\le A\frac{\sqrt{np}+p+n\eta\log(2e/\eta)+2\log(1/\delta)}{n(1-\eta)}+\big(4\mathsf{r}_{n,\mathfrak{s}}(1+\sqrt{\eta})+10\eta\sqrt{\log(2/\eta)}\big)^2+\varepsilon\\ &\le A_{\mathfrak{s}}\big(\mathsf{r}_{n,\mathfrak{s}}+\mathsf{r}_{n,\mathfrak{s}}^2+8\eta\log(2/\eta)\big)+\big(7.5\mathsf{r}_{n,\mathfrak{s}}+10\eta\sqrt{\log(2/\eta)}\big)^2+\varepsilon,\end{aligned}$$

where $A_{\mathfrak{s}}$ is a constant depending only on the variance proxy $\mathfrak{s}$, the value of which is not necessarily the same in further simplifications of the expression from the last display. Then, using the triangle inequality several times we arrive at the following expression

$$\begin{aligned}\sqrt{\varepsilon T_1} &\le \big\{A_{\mathfrak{s}}\varepsilon\big(\mathsf{r}_{n,\mathfrak{s}}+\mathsf{r}_{n,\mathfrak{s}}^2+8\eta\log(2/\eta)\big)\big\}^{1/2}+\big(7.5\mathsf{r}_n+10\eta\sqrt{\log(2/\eta)}\big)\sqrt{\varepsilon}+\varepsilon\\ &\le (A_{\mathfrak{s}}+\sqrt{A_{\mathfrak{s}}/2}+5.4)\mathsf{r}_{n,\mathfrak{s}}+(7.1+2\sqrt{2A_{\mathfrak{s}}})\tau\sqrt{\log(2/\tau)}+(9.1+2\sqrt{2A_{\mathfrak{s}}})\varepsilon\sqrt{\log(2/\varepsilon)}.\end{aligned}$$

These inequalities imply that there is a universal constant $\mathsf{C}$ such that

$$\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}-\boldsymbol{\mu}^*\right\|_2 \le \frac{\mathsf{C}\mathfrak{s}\big(A_{\mathfrak{s}}\mathsf{r}_{n,\mathfrak{s}}/\sqrt{\mathfrak{s}}+\tau\sqrt{\log(2/\tau)}+\varepsilon\sqrt{\log(2/\varepsilon)}+\mathsf{r}_{n,\mathfrak{s}}\varepsilon/\sqrt{\mathfrak{s}\tau}\big)\sqrt{L}}{1-2\varepsilon^*}. \tag{B.34}$$

Let us denote $\log_+(x)=\max\{0,\log(x)\}$ the positive part of logarithm, then we choose

$$\tau=\frac{1}{4}\bigwedge\frac{\bar{\mathsf{r}}_{n,\mathfrak{s}}}{\sqrt{\log_+(2/\bar{\mathsf{r}}_{n,\mathfrak{s}})}},\qquad\text{with}\qquad\bar{\mathsf{r}}_{n,\mathfrak{s}}=\frac{3\sqrt{\mathfrak{s}}\big(\sqrt{p}+2\sqrt{\log(2/\delta)}\big)}{\sqrt{n}}.$$

Note that $\mathsf{r}_{n,\mathfrak{s}}\le\sqrt{2}\bar{\mathsf{r}}_{n,\mathfrak{s}}$ and, furthermore, $\tau=1/4$ whenever $\bar{\mathsf{r}}_{n,\mathfrak{s}}\ge 1/2$. Therefore, $\mathsf{r}_{n,\mathfrak{s}}\varepsilon/\sqrt{\tau}\le\mathsf{r}_{n,\mathfrak{s}}+\varepsilon$. Inserting this value of $\tau$ in (B.34) leads to

$$\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}-\boldsymbol{\mu}^*\right\|_2 \le \frac{\mathsf{C}\,\mathfrak{s}\big(A_{\mathfrak{s}}\mathsf{r}_{n,\mathfrak{s}}/\sqrt{\mathfrak{s}}+\varepsilon\sqrt{\log(2/\varepsilon)}\big)\sqrt{L}}{1-2\varepsilon^*}.$$

where $\mathsf{C}$ is a universal constant. Replacing $\mathsf{r}_{n,\mathfrak{s}}$ by its expression, and upper bounding $L$ by $2\log p$, we arrive at

$$\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{SDR}}-\boldsymbol{\mu}^*\right\|_2 \le \frac{\mathsf{C}\,\mathfrak{s}\sqrt{\log p}}{1-2\varepsilon^*}\left(A_{\mathfrak{s}}\sqrt{\frac{p}{n}}+\varepsilon\sqrt{\log(2/\varepsilon)}+A_{\mathfrak{s}}\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

Note that this inequality holds true on an event of probability at least $1 - 5\delta$.

# Bibliography


Alon, N., Matias, Y., and Szegedy, M. (1996). The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '96, page 20–29, New York, NY, USA. Association for Computing Machinery.

Alquier, P. and Gerber, M. (2020). Universal robust regression via maximum mean discrepancy. *arXiv: Statistics Theory*.

Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794.

Bahmani, S. (2021). Nearly optimal robust mean estimation via empirical characteristic function. *Bernoulli*, 27(3):2139 – 2158.

Bakshi, A. and Prasad, A. (2021). *Robust Linear Regression: Optimal Rates in Polynomial Time*, page 102–115. Association for Computing Machinery, New York, NY, USA.

Balakrishnan, S., Du, S. S., Li, J., and Singh, A. (2017). Computationally efficient robust sparse estimation in high dimensions. In *COLT 2017*, pages 169–212.

Bateni, A.-H. and Dalalyan, A. S. (2020). Confidence regions and minimax rates in outlier-robust estimation on the probability simplex. *Electron. J. Statist.*, 14(2):2653–2677.

Bateni, A.-H., Minasyan, A., and Dalalyan, A. S. (2022). Nearly minimax robust estimator of the mean vector by iterative spectral dimension reduction.

Bellec, P. C. (2016). Adaptive confidence sets in shape restricted regression. *arXiv preprint arXiv:1601.05766*.

Bellec, P. C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression. *Ann. Statist.*, 46(2):745–780.

Berend, D. and Kontorovich, A. (2013). A sharp estimate of the binomial mean absolute deviation with applications. *Statistics & Probability Letters*, 83(4):1254 – 1259.

Bhatia, K., Jain, P., Kamalaruban, P., and Kar, P. (2017). Consistent robust regression. In *Advances in Neural Information Processing Systems*, pages 2110–2119.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford.

Braess, D. and Sauer, T. (2004). Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128(2):187 – 206.

Brunel, V.-E. (2019). Concentration of the empirical level sets of Tukey's halfspace depth. *Probability Theory and Related Fields*, 173(3-4):1165–1196.

Cai, T. T. and Low, M. G. (2006). Adaptive confidence balls. *Ann. Statist.*, 34(1):202–228.

Cardot, H., Cenac, P., and Zitt, P.-A. (2013). Efficient and fast estimation of the geometric median in hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19:18–43.

Carpentier, A., Delattre, S., Roquain, E., and Verzelen, N. (2018). Estimating minimum effect with outlier selection. *arXiv e-prints*, page arXiv:1809.08330.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148 – 1185.

Catoni, O. and Giulini, I. (2017). Dimension-free pac-bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv: Statistics Theory*.

Catoni, O. and Giulini, I. (2018). Dimension-free pac-bayesian bounds for the estimation of the mean of a random vector. *arXiv: Statistics Theory*.

Chen, M., Gao, C., and Ren, Z. (2016). A general decision theory for Huber's $\varepsilon$-contamination model. *Electronic Journal of Statistics*, 10(2):3752–3774.

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

Chen, S., Li, J., and Moitra, A. (2020). Efficiently learning structured distributions from untrusted batches. In *Proccedings of STOC 2020, June 22-26*, pages 960–973. ACM.

Chen, Y., Caramanis, C., and Mannor, S. (2013). Robust sparse regression under adversarial corruption. In *International Conference on Machine Learning*, pages 774–782.

Cheng, Y., Diakonikolas, I., and Ge, R. (2019a). High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of SODA 2019*, pages 2755–2771.

Cheng, Y., Diakonikolas, I., Ge, R., and Soltanolkotabi, M. (2020). High-dimensional robust mean estimation via gradient descent. *CoRR*, abs/2005.01378.

Cheng, Y., Diakonikolas, I., Ge, R., and Woodruff, D. P. (2019b). Faster algorithms for high-dimensional robust covariance estimation. In Beygelzimer, A. and Hsu, D., editors, *COLT 2019*, volume 99 of *Proceedings of Machine Learning Research*, pages 727–757. PMLR.

Cheng, Y., Diakonikolas, I., Kane, D. M., Ge, R., Gupta, S., and Soltanolkotabi, M. (2021). Outlier-robust sparse estimation via non-convex optimization. *CoRR*, abs/2109.11515.

Cherapanamjeri, Y., Aras, E., Tripuraneni, N., Jordan, M. I., Flammarion, N., and Bartlett, P. L. (2020). Optimal robust linear regression in nearly linear time. *ArXiv*, abs/2007.08137.

Cherapanamjeri, Y., Flammarion, N., and Bartlett, P. L. (2019). Fast mean estimation with sub-Gaussian rates. In *COLT 2019*, volume 99 of *Proceedings of Machine Learning Research*, pages 786–806.

Chinot, G. (2020). Erm and rerm are optimal estimators for regression problems when malicious outliers corrupt the labels. *Electron. J. Statist.*, 14(2):3563–3605.

Chinot, G., Lecué, G., and Lerasle, M. (2018). Statistical learning with Lipschitz and convex loss functions. *arXiv e-prints*, page arXiv:1810.01090.

Chinot, G., Lecué, G., and Lerasle, M. (2020). Robust high dimensional learning for Lipschitz and convex losses. *J. Mach. Learn. Res.*, 21:Paper No. 233, 47.

Cohen, D., Kontorovich, A., and Wolfer, G. (2020). Learning discrete distributions with infinite support.

Collier, O. and Dalalyan, A. S. (2019). Multidimensional linear functional estimation in sparse gaussian models and robust estimation of the mean. *Electron. J. Statist.*, 13(2):2830–2864.

Comminges, L., Collier, O., Ndaoud, M., and Tsybakov, A. B. (2021). Adaptive robust estimation in sparse vector model. *Ann. Statist.*, 49(3):1347–1377.

Comminges, L. and Dalalyan, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistics*, 40(5):2667 – 2696.

Dalalyan, A. and Chen, Y. (2012). Fused sparsity and robust estimation for linear models with unknown variance. In *Advances in Neural Information Processing Systems 25*, pages 1259–1267. Curran Associates, Inc.

Dalalyan, A. and Thompson, P. (2019). Outlier-robust estimation of a sparse linear model using $\ell_1$-penalized Huber's M-estimator. In *NeurIPS 32*, pages 13188–13198.

Dalalyan, A. S. and Minasyan, A. (2020). All-in-one robust estimator of the gaussian mean. *math.ST*, arXiv:2002.01432.

Dalalyan, A. S. and Sebbar, M. (2018). Optimal Kullback-Leibler aggregation in mixture density estimation by maximum likelihood. *Math. Stat. Learn.*, 1(1):1–35.

Depersin, J. (2020). A spectral algorithm for robust regression with subgaussian rates. *ArXiv*, abs/2007.06072.

Depersin, J. and Lecué, G. (2021). On the robustness to adversarial corruption and to heavy-tailed data of the stahel-donoho median of means. *arXiv preprint arXiv:2101.09117*.

Depersin, J. and Lecué, G. (2022). Robust sub-Gaussian estimation of a mean vector in nearly linear time. *Ann. Statist.*, 50(1):511–536.

Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. (2016). Sub-gaussian mean estimators. *Ann. Statist.*, 44(6):2695–2725.

Devroye, L. and Lugosi, G. (2000). *Combinatorial methods in density estimation*. Springer Series in Statistics.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2016a). Robust estimators in high dimensions without the computational intractability. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016*, pages 655–664.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2017). Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 999–1008.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2018a). Robustly learning a gaussian: Getting optimal error, efficiently. In Czumaj, A., editor, *Proceedings of SODA 2018*, pages 2683–2702. SIAM.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Steinhardt, J., and Stewart, A. (2019a). Sever: A robust meta-algorithm for stochastic optimization. *ArXiv*, abs/1803.02815.

Diakonikolas, I. and Kane, D. M. (2019). Recent advances in algorithmic high-dimensional robust statistics. *CoRR*, abs/1911.05911.

Diakonikolas, I., Kane, D. M., and Pensia, A. (2020). Outlier robust mean estimation with subgaussian rates via stability. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1830–1840. Curran Associates, Inc.

Diakonikolas, I., Kane, D. M., and Stewart, A. (2016b). Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. *Electron. Colloquium Comput. Complex.*, 23:177.

Diakonikolas, I., Kong, W., and Stewart, A. (2019b). Efficient algorithms and lower bounds for robust linear regression. SODA '19, pages 2745–2754, USA. Society for Industrial and Applied Mathematics.

Diakonikolas, I., Li, J., and Schmidt, L. (2018b). Fast and sample near-optimal algorithms for learning multidimensional histograms. In *COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, pages 819–842.

Dong, Y., Hopkins, S. B., and Li, J. (2019). Quantum entropy scoring for fast robust mean estimation and improved outlier detection. In *NeurIPS 2019*, pages 6065–6075.

Donoho, D. L. and Gasko, M. (1992). Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness. *The Annals of Statistics*, 20(4):1803 – 1827.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. Festschr. for Erich L. Lehmann, 157-184 (1983).

Donoho, D. L. and Montanari, A. (2016). High dimensional robust m-estimation: asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3):935–969.

Feng, J., Xu, H., Mannor, S., and Yan, S. (2014). Robust logistic regression and classification. In *Advances in Neural Information Processing Systems 27*, pages 253–261. Curran Associates, Inc.

Gao, C. (2020). Robust regression via mutivariate regression depth. *Bernoulli*, 26(2):1139 – 1170.

Gao, C., Liu, J., Yao, Y., and Zhu, W. (2018). Robust estimation and generative adversarial nets.

Gao, C., Yao, Y., and Zhu, W. (2020). Generative adversarial nets for robust scatter estimation: A proper scoring rule perspective. *J. Mach. Learn. Res.*, 21:160:1–160:48.

Goes, J., Lerman, G., and Nadler, B. (2020). Robust sparse covariance estimation by thresholding Tyler's M-estimator. *Ann. Statist.*, 48(1):86–110.

Guntuboyina, A. and Sen, B. (2018). Nonparametric shape-restricted regression. *Statist. Sci.*, 33(4):568–594.

Halko, N., Martinsson, P. G., and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288.

Hoffmann, M. and Nickl, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.*, 39(5):2383–2409.

Hopkins, S. B. (2020). Mean estimation with sub-Gaussian rates in polynomial time. *The Annals of Statistics*, 48(2):1193 – 1213.

Hsu, D. and Sabato, S. (2016). Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.*, 17(1):543–582.

Huber, P. and Ronchetti, E. (2011). *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101.

Jain, A. and Orlitsky, A. (2019). Robust learning of discrete distributions from batches. *CoRR*, arXiv:1911.08532.

Jain, A. and Orlitsky, A. (2021). Robust density estimation from batches: The best things in life are (nearly) free. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4698–4708. PMLR.

Jerrum, M. R., Valiant, L. G., and Vazirani, V. V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188.

Joly, E., Lugosi, G., and Imbuzeiro Oliveira, R. (2017). On the estimation of the mean of a random vector. *Electron. J. Statist.*, 11(1):440 – 451.

Kamath, S., Orlitsky, A., Pichapati, D., and Suresh, A. T. (2015). On learning distributions from their samples. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 1066–1100. JMLR.org.

Kearns, M. (1998). Efficient noise-tolerant learning from statistical queries. *J. ACM*, 45(6):983–1006.

Koltchinskii, V. and Lounici, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110 – 133.

Lai, K. A., Rao, A. B., and Vempala, S. S. (2016). Agnostic estimation of mean and covariance. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016*, pages 665–674.

Lecué, G. and Lerasle, M. (2017). Learning from MOM's principles: Le Cam's approach. *Stoch. Proc. App., to appear*.

Lecué, G., Lerasle, M., and Mathieu, T. (2020). Robust classification via mom minimization. *Machine Learning*, pages 1–31.

Lei, Z., Luh, K., Venkat, P., and Zhang, F. (2019). A fast spectral algorithm for mean estimation with sub-gaussian rates. *ArXiv*, abs/1908.04468.

Liu, H. and Gao, C. (2019). Density estimation with contamination: minimax rates and theory of adaptation. *Electron. J. Stat.*, 13(2):3613–3653.

Liu, J., Cosman, P. C., and Rao, B. D. (2018). Robust linear regression via $\ell_0$ regularization. *IEEE Transactions on Signal Processing*, 66(3):698–713.

Liu, J., Deshmukh, A., and Veeravalli, V. V. (2020a). Robust mean estimation in high dimensions via $\ell_0$ minimization.

Liu, L., Shen, Y., Li, T., and Caramanis, C. (2020b). High dimensional robust sparse regression. In Chiappa, S. and Calandra, R., editors, *AISTATS 2020*, volume 108 of *Proceedings of Machine Learning Research*, pages 411–421. PMLR.

Liu, X., Mosler, K., and Mozharovskyi, P. (2019). Fast computation of tukey trimmed regions and median in dimension $p > 2$. *Journal of Computational and Graphical Statistics*, 28:682–697.

Liu, X., Zuo, Y., and Wang, Q. (2017). Finite sample breakdown point of tukey's halfspace median. *Sci. China Math*, 60:861–874.

Lopuhaa, H. P. and Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 19(1):229–248.

Lugosi, G. and Mendelson, S. (2019a). Mean estimation and regression under heavy-tailed distributions: A survey. *Found Comput Math*, 19:1145–1190.

Lugosi, G. and Mendelson, S. (2019b). Sub-gaussian estimators of the mean of a random vector. *Ann. Statist.*, 47(2):783–794.

Lugosi, G. and Mendelson, S. (2021). Robust multivariate mean estimation: The optimality of trimmed mean. *The Annals of Statistics*, 49(1):393 – 410.

Maronna, R., Martin, D., and Yohai, V. (2006). *Robust Statistics: Theory and Methods*. Wiley Series in Probability and Statistics. Wiley.

Minasyan, A. G. (2020). Excess-risk consistency of group-hard thresholding estimator in robust estimation of Gaussian mean. *Journal of Contemporary Mathematical Analysis*, 55(3):208–212.

Minsker, S. (2015). Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335.

Minsker, S. (2018a). Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.*, 46(6A):2871–2903.

Minsker, S. (2018b). Uniform bounds for robust mean estimators.

Minsker, S. and Ndaoud, M. (2021). Robust and efficient mean estimation: an approach based on the properties of self-normalized sums. *Electronic Journal of Statistics*, 15(2):6036 – 6070.

Nemirovsky, A. and Yudin, D. (1983). Problem complexity and method efficiency in optimization.

Nguyen, N. H. and Tran, T. D. (2013). Robust lasso with missing and grossly corrupted observations. *IEEE Trans. Inform. Theory*, 59(4):2036–2058.

Pensia, A., Jog, V., and Loh, P.-L. (2020). Robust regression with covariate filtering: Heavy tails and adversarial contamination. *preprint arXiv:2009.12976*.

Prasad, A., Balakrishnan, S., and Ravikumar, P. (2019). A unified approach to robust mean estimation.

Prasad, A., Balakrishnan, S., and Ravikumar, P. (2020). A robust univariate mean estimator is all you need. In *AISTATS*.

Qiao, M. and Valiant, G. (2018). Learning discrete distributions from untrusted batches. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, volume 94 of *LIPIcs*, pages 47:1–47:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Rigollet, P. and Hütter, J.-C. (2019). High dimensional statistics. *Lecture notes (MIT)*.

Rousseeuw, P., Hampel, F., Ronchetti, E., and Stahel, W. (2011). *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Statistics. Wiley.

Rousseeuw, P. J. and Hubert, M. (1999). Regression depth. *Journal of the American Statistical Association*, 94(446):388–402.

Sasai, T. and Fujisawa, H. (2021). Adversarial robust weighted huber regression. *arXiv preprint arXiv:2102.11120*.

Steinhardt, J., Charikar, M., and Valiant, G. (2017). Resilience: A criterion for learning in the presence of arbitrary outliers.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York.

Tukey, J. W. (1975). Mathematics and the picturing of data. Proc. int. Congr. Math., Vancouver 1974, Vol. 2, 523-531.

Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices*, page 210–268. Cambridge University Press.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wang, Z. and Tan, Z. (2021). Tractable and near-optimal adversarial algorithms for robust estimation in contaminated gaussian models.

Xia, D. and Koltchinskii, V. (2016). Estimation of low rank density matrices: Bounds in schatten norms and other distances. *Electron. J. Statist.*, 10(2):2717–2745.

Zhu, B., Jiao, J., and Jordan, M. I. (2022). Robust estimation for nonparametric families via generative adversarial networks. *CoRR*, abs/2202.01269.

Zhu, B., Jiao, J., and Steinhardt, J. (2019). Generalized resilience and robust statistics. *ArXiv*, abs/1909.08755.

Zhu, B., Jiao, J., and Steinhardt, J. (2020). When does the tukey median work? *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 1201–1206.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Ann. Statist*, 28:461–482.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

**Titre:** Contributions à l'estimation robuste : optimalité minimax *vs.* efficacité calculatoire

**Mots clés:** robustesse, approche minimax, efficacité calculatoire, statistique en grande dimension

**Résumé:** En statistique et en théorie de l'apprentissage statistique, on suppose souvent que les échantillons sont distribués indépendamment et identiquement selon une distribution de probabilité de référence. Une approche plus réaliste pourrait consister à relaxer cette hypothèse en permettant à une fraction des échantillons de ne pas nécessairement suivre la distribution de référence. Ces échantillons désobéissants, appelés données aberrantes, peuvent considérablement détériorer la performance des estimateurs classiques. Dans ce travail, nous cherchons à estimer la moyenne des distributions de référence par des estimateurs robustes aux données aberrantes. Nous nous intéressons au comportement non-asymptotique des estimateurs.

Dans un premier temps, nous décrivons divers modèles de contamination qui déterminent la nature des données aberrantes parmi nos observations. Puis, nous considérons le problème de l'estimation de la moyenne d'une distribution dont le support est le simplexe de probabilité de dimension $k$ dans le cas où une fraction $\varepsilon$ d'observations sont des données aberrantes générées par un adversaire. Un exemple particulier simple est le problème de l'estimation de la distribution d'une variable aléatoire discrète. Dans un deuxième temps, nous étudions le problème de l'estimation robuste de la moyenne d'une distribution gaussienne. Les estimateurs minimax-optimaux connus pour ce problème ne sont pas calculables en temps polynomial. Nous introduisons un estimateur efficace basé sur la réduction spectrale de dimension et établissons une borne supérieure sur son erreur qui est minimax-optimale modulo des facteurs logarithmiques.

**Title:** Contributions to robust estimation: minimax optimality *vs.* computational tractability

**Keywords:** robustness, minimax approach, computational efficiency, high-dimensional data

**Abstract:** In statistics and learning theory, it is common to assume that samples are independently and identically distributed according to a reference probability distribution. A more realistic approach could be to relax this assumption by allowing a fraction of samples to not necessarily follow the reference distribution. These disobeying samples, called outliers, may drastically skew the classical estimators. In this work, we aim to estimate the mean of reference distributions by estimators robust to outliers. We are interested in the non-asymptotic behavior of the estimators.

In the first stage, we describe various contamination models which determine the nature of the outliers among our observations. Then, we consider the problem of estimating the mean of a distribution supported by the $k$-dimensional probability simplex in the setting where an $\varepsilon$ fraction of observations are outliers generated by an adversary. A simple particular example is the problem of estimating the distribution of a discrete random variable. In the second stage, we study the problem of robust estimation of the mean of a Gaussian distribution. The known minimax-optimal estimators for this problem are not computationally tractable. We introduce a computationally efficient estimator based on spectral dimension reduction and establish a finite sample upper bound on its error that is minimax-optimal up to logarithmic factors.

**Institut Polytechnique de Paris**
91120 Palaiseau, France