Department of Mathematics
TUM School of Computation, Information and Technology
Technical University of Munich

Robustifying score matching for graphical models

Robust Score matching Methode für Grafische Modelle

Yuki Suzuki

Thesis for the attainment of the academic degree

Master of Science

at the TUM School of Computation, Information and Technology of the Technical University of Munich

Supervisor:
Prof. Mathias Drton

Advisors:
Dr. Oleksandr Zadorozhnyi

Submitted:
Munich, 15th. Nobember 2023

I hereby declare that this thesis is entirely the result of my own work except where otherwise indicated. I have only used the resources given in the list of references.

Munich, 15th. Nobember 2023 Yuki Suzuki 

Abstract

A score matching estimator is defined as the minimizer of the score loss function. In the case of an exponential family with a density that satisfies boundary conditions, the score matching estimator can be explicitly written down. In this thesis, our primary focus is on applying score matching method to estimate the parameters of a Gaussian distribution $\mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$, with the goal of constructing a robust version of it. To achieve this, we start by deriving the explicit form of the score matching estimator for the unknown parameters $\boldsymbol{\mu}$ and Σ of the Gaussian distribution. We see that the form can be expressed as a composition of empirical mean and empirical covariance matrix. Then, replacing them to robust alternatives, we obtain the robust score matching estimator for the parameter $\boldsymbol{\mu}$ and Σ. To observe the behavior of the estimators, we derive the concentration inequalities and conduct numerical simulations.

Contents

1 Introduction

When conducting statistical experiments, one of our interests is how the estimated value is differ from the true parameter of interest. This is crucial in practice, because we have a restriction of budget for collecting data, and the experiment organizer need to know the minimum sample size required to obtain results within the error that the researcher can admit. In order to describe this topic, we start from considering a statistical model $(S, \mathcal{F}, \mathcal{P})$. Statistical model is characterized as a pair of a sample space $(S, \mathcal{F})$ and a set of probability distributions $\mathcal{P}$ on $(S, \mathcal{F})$. We suppose that $\mathcal{P}$ is parametrized by population parameter $\boldsymbol{\theta} \in \Theta : \mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m\}$, where Θ is the set of all the possible parameters. We denote the density function of $P_{\boldsymbol{\theta}}$ by $p_{\boldsymbol{\theta}}$ for each $\boldsymbol{\theta} \in \Theta$. In a statistical experiment one typically considers a sample of size n, $(\boldsymbol{X}_i)_{i\in[n]} \in S$, and the goal is to estimate a parameter $\boldsymbol{\theta}$ of the statistical model $(S, \mathcal{F}, \mathcal{P})$. Then, our problems is paraphrased as the problem of designing data-dependent decision rules which describe behavior of the unknown population parameter $\boldsymbol{\theta}$ based on the sample of the fixed size. In particular, when we have a population parameter $\boldsymbol{\theta}$ and the value of its point estimate $\hat{\boldsymbol{\theta}}$, we can consider the probability with which the underlying estimator gather around the population value with error ϵ : For the acceptable error $\epsilon > 0$, we are interested in the value $\delta(n, \epsilon) \in [0, 1]$ which satisfies

$$P_{\theta_0}\left(\left|\hat{\boldsymbol{\theta}}((X_i)_{i\in[n]}) - \boldsymbol{\theta}_0\right| \leq \epsilon\right) \geq 1 - \delta(n, \epsilon). \tag{1.1}$$

(1.1) is called *Concentration inequality* for parameter $\boldsymbol{\theta}$. In Chapter 3 in this thesis, we review several important concentration inequalities.

Score matching method was firstly developed in Hyvärinen (2005) in [7]. Suppose that for each $\boldsymbol{\theta} \in \Theta$ the density function $p_{\boldsymbol{\theta}}$ of the distribution $P_{\boldsymbol{\theta}}$ is supported on $\mathbb{R}^d$ and twice continuously differentiable. The score matching estimator $\hat{\boldsymbol{\theta}}_{\mathrm{SM}}$ [1] is defined as the minimizer with respect to $\boldsymbol{\theta} \in \Theta$ of the expected squared ℓ_2 distance between the gradients of $\log p_0$ and $\log p$. That is,

$$\hat{\boldsymbol{\theta}}_{\mathrm{SM}} \equiv \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \int_{\mathbb{R}^d} p_0(\boldsymbol{x}) \left|\nabla \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \nabla \log p_{\boldsymbol{\theta}_0}(\boldsymbol{x})\right|^2 \mathbf{d}\boldsymbol{x}. \tag{1.2}$$

As shown in [7], when $p_{\boldsymbol{\theta}}$ satisfies several moderate conditions, (1.2) can be rewritten without any terms involving $p(x)$,

$$\hat{\boldsymbol{\theta}}_{\mathrm{SM}} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \int_{\mathbb{R}^d} p_{\boldsymbol{\theta}_0}(\boldsymbol{x}) \sum_{j=1}^{d} \left[\partial_{jj} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) + \frac{(\partial_j \log p_{\boldsymbol{\theta}}(\boldsymbol{x}))^2}{2}\right] \mathbf{d}\boldsymbol{x}. \tag{1.3}$$

Moreover for an exponential family of distributions with the density function satisfying $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{t}(\boldsymbol{x}) - a(\boldsymbol{\theta}) + b(\boldsymbol{x})$ with a sufficient statistics $\boldsymbol{t}$ and real-valued functions $a : \mathbb{R}^m \mapsto \mathbb{R}$ and $b : \mathbb{R}^d \mapsto \mathbb{R}$, we can explicitly write down (1.3) by

$$\hat{\boldsymbol{\theta}}_{\mathrm{SM}} = \boldsymbol{\Gamma}^{-1}\boldsymbol{g}, \quad \text{with} \quad \begin{array}{rcl} \boldsymbol{\Gamma} & \equiv & \mathbb{E}_{p_{\boldsymbol{\theta}_0}} \sum_{j=1}^{d} \boldsymbol{t}_j'(X)\boldsymbol{t}_j'(X)^\top \\ \boldsymbol{g} & \equiv & -\mathbb{E}_{p_{\boldsymbol{\theta}_0}} \sum_{j=1}^{d} [b_j'(X)\boldsymbol{t}_j'(X) + \boldsymbol{t}_j''(X)]. \end{array} \tag{1.4}$$

The score matching estimator introduce by Hyvärinen is defined as the empirical version of (1.4), which we call "the original score matching estimator". Yu et al.(2019) [8] generalized the idea of score matching method to broad ranges of distributions. Their works enable us to consider distributions whose densities

[1] In [7], Score matching estimator is denoted by the natation $\hat{\boldsymbol{\theta}}$, but in this thesis we use the notation $\hat{\boldsymbol{\theta}}_{\mathrm{SM}}$ in order to distinguish it from other estimators.

are supported on a subset S of $\mathbb{R}^d$. The idea is to introduce a function $\boldsymbol{h} : S \to \mathbb{R}$which is enough small around the boundary of S. Instead of (1.2), generalized score matching method consider the following optimization problem.

$$\hat{\boldsymbol{\theta}}_{\mathrm{SM}} \equiv \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \int_S p_{\theta_0}(\boldsymbol{x}) \left| \nabla \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \circ \boldsymbol{h}(\boldsymbol{x})^{1/2} - \nabla \log p_{\theta_0}(\boldsymbol{x}) \circ \boldsymbol{h}(\boldsymbol{x})^{1/2} \right|^2 \mathbf{d}\boldsymbol{x} \tag{1.5}$$

Similar transformation like (1.3) is possible under moderate conditions with respect to p_{θ_0} and $\boldsymbol{h}$. And for exponential families we can obtain an analogical result with (1.4). We call emprical version of it the *generalized score matching estimator*.

In this thesis, one of the problems we consider is the application of the score matching method for estimating the unknown parameters $(\boldsymbol{\mu}, \Sigma)$ of Gaussian distribution. We consider a statistical model with $S = \mathbb{R}^d$ which is characterized by parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$, and the density of each distribution $F_{\boldsymbol{\theta}}$ is given by

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) \equiv \frac{1}{\sqrt{2\pi|\Omega|^{-1}}} \exp\left(-\frac{1}{2}(x-\boldsymbol{\mu})^\top \Omega (x-\boldsymbol{\mu})\right).$$

For each $\boldsymbol{\mu}$ and Σ, F_θ is called *Gaussian distribution*, which is one of exponential families. As introduced in many books such as [15], the classical estimator for the unkown parameter $\boldsymbol{\mu}$ and Σ of Gaussian distribution, the empirical mean and covariance

$$\hat{\boldsymbol{\mu}} \equiv \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i \qquad \hat{\Sigma} \equiv \frac{1}{n}\sum_{i=1}^n (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})^\top$$

works well. However, in practice observed data sometimes contains outliers, in other words the observed data may be contaminated by several data whose distribution is different from the distribution that occupies a large portion of the sample. In order to ease the effect of contamination, we consider *robust* estimator instead of classical estimator. As introduced in literatures such as [3], [5],[14], and [16], a lot of robust estimators have been developed, and we can write down the concentration inequality for them. We will see some of them in Chapter 4 in this thesis.

In Chapter 5, we obtain the explicit form of $\hat{\boldsymbol{\theta}}_{\mathrm{SM}}$ for the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma)$ of Gaussian distribution by applying (1.5). And we robustify the score matching estimator by replacing the components of the estimator to robust ones. Then, we check the behaviors of the robust score matching estimators from the view point of concentration inequality and numerical simulations.

This thesis is organized as follows. In Chapter 3, we review several properties of statistical inequality and show the inequalities for sample mean and sample covariance. In Chapter 4, we review the properties of previously well-known robust estimators (Median of means, Trimmed mean, Minsker's covariance estimator, and so on.) for mean and covariance. In Chapter 5, we introduce *score matching loss function* and define the *score matching estimator*, then derive the explicit form for Gaussian distribution and Pareto distribution. Moreover, we combine the idea of score matching method and robust statistics to create robust version of $\hat{\boldsymbol{\theta}}_{\mathrm{SM}}$. Then, analyze the error through applying concentration inequalities concentration inequality and investigate the behavior of the proposed estimator by means of numerical simulations.

2 Notation

In this thesis, we use the following notations unless otherwise stated.

We use regular font for scalars, boldface for vectors. We use characters in lower-case for deterministic elements, in upper-case for random elements, e.g.

x	: deterministic variable,	$\boldsymbol{x}$	: deterministic vector,
X	: random variable,	$\boldsymbol{X}$	: random vector.

Following alphabets are used as natural numbers for specific meanings.

n	: the number of samples,	i	: the index of the sample
d	: the dimension the sample space,	j	: the coordinate,
m	: the dimension of the parameter space.		

We defines the following symbols by the right sides.

$\mathbb{R}^d$: real valued $d-$dimensional Euclidean space

$\mathbb{R}^{d_1 \times d_2}$: the whole set of real valued $d_1 \times d_2$ matrices

$\mathrm{S}_d \equiv \{M \in \mathbb{R}^{d \times d} : M = M^\top\}$

$\mathrm{PD}_d \equiv \{M \in \mathbb{R}^{d \times d} : M = M^\top \text{ and } M \text{ is positive semidefinite}\}$

$|\boldsymbol{x}|$: Euclidean norm for $\boldsymbol{x} \in \mathbb{R}^d$

$\|\boldsymbol{M}\|$: the spectral norm for the matrix $\boldsymbol{M} \in \mathbb{R}^{d_1 \times d_2}$ (in case that $d_1 = d_2$, the operator norm)

$$\mathrm{vec} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 d_2} \quad \begin{pmatrix} a_{11} & \dots & a_{1d_2} \\ \vdots & \ddots & \vdots \\ a_{d_1 1} & \dots & a_{d_1 d_2} \end{pmatrix} \mapsto (a_{11}, \dots a_{d_1 1}, \dots, a_{1d_2}, \dots a_{d_1 d_2})^\top$$

$[n] \equiv \{1, \dots, n\} \in \mathbb{N}$ for $n \in \mathbb{N}$

$$\mathrm{Sign}(x) \equiv \begin{cases} 1 & x > 0 \\ 0 & \text{when } x = 0 \\ -1 & x < 0 \end{cases}$$

$\mathbf{e}_{j,d} \in \mathbb{R}^d$ is a a vector the j-th entry 1 and otherwise 0, that is, $\mathbf{e}_{j,d} \equiv (0, ..., 0, \underset{\text{j th}}{1}, 0, ..., 0)^\top$.

$\int_{\mathbb{R}^{d-1}} \cdot \, \boldsymbol{x}_{-\mathrm{j}}$ means the integration with respect to all the coordinates except for x_j

3 Concentration inequalities and their usage in statistics

In an estimation problem, one of our interest is estimating the population parameter $\boldsymbol{\theta}_0 \in \Theta \subset \mathbb{R}^m$ by using a set of observations $(x_i)_{i\in[n]}$. Here, Θ is the parameter space, and family of probability measures $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\}$are parameterized by the elements $\boldsymbol{\theta} \in \Theta$. P_{θ_0} is the probability distribution which corresponds to the unknown parameter θ_0, and $(x_i)_{i\in[n]}$ comes from P_{θ_0}. Let $(X_i)_{i\in[n]}$ be i.i.d. random vectors which distribute from P_{θ_0}. We consider a function f of these random vectors, and we call the random vector $\hat{\theta} \equiv f(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ *estimator* and the certain vector $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ *estimate* of θ_0 based on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. In order to measure the error of the estimator from the true paremeter θ_0, one one usually considers the *mean squared error*, which is defined as

$$\mathbb{E}_{P_{\theta_0}{}^{\otimes n}}[(\hat{\theta} - \theta_0)^2]$$

Another tool to quantify how a random variable deviates from some value is *concentration inequality*, which is the main topic of this chapter. Concentration inequalities quantify random fluctuations of functions of independent random variables by bounding the probability with which the function differs from its expected value by more than a certain amount.

3.1 Concentration inequalities for mean of random vector

Starting from Markov's inequality, we can get following two concentration inequalities. One is *Chebyshev's inequality*, and the other is *Chernoff Bound*, which is the key equality in this section.

Theorem 3.1.1 (Markov's inequality) *If X is a nonnegative random variable with $\mathbb{E}[X] < \infty$ and $\epsilon > 0$, then*

$$\mathbb{P}(X \geq \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon}. \tag{3.1}$$

Proof.
Define a new random variable $Y \equiv \epsilon 1_{\{X\geq\epsilon\}}$, i.e.

$$Y(\omega) \equiv \begin{cases} \epsilon & X(\omega) \geq \epsilon \\ 0 & X(\omega) < \epsilon \end{cases}$$

since $Y \leq X$,

$$\frac{\mathbb{E}[X]}{\epsilon} \geq \frac{\mathbb{E}[Y]}{\epsilon} = \mathbb{E}\left[1_{\{X\geq\epsilon\}}\right] = \mathbb{P}(X \geq \epsilon).$$

□

By applying Markov's inequality to $(X - \mathbb{E}[X])^2$ and e^X,which are always positive, we get the following two inequalities.

Theorem 3.1.2 (Chebyshev's inequality) *If X is a random variable with $\mathbb{E}[X^2] < \infty$ and $\epsilon > 0$, then*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\mathrm{Var}(X)}{\epsilon^2}. \tag{3.2}$$

Proof.

$$\begin{aligned}\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) &= \mathbb{P}\left((X - \mathbb{E}[X])^2 \geq \epsilon^2\right) \\ &\underset{Markov}{\leq} \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}{\epsilon^2} \\ &= \frac{\mathrm{Var}(X)}{\epsilon^2}\end{aligned}$$

□

Theorem 3.1.3 (Chernoff bound, see in [13]) *Suppose X is a random variable with $\mathbb{E}\left[e^{tX}\right] < \infty$ for all $t \in \mathbb{R}_+$. Then,*

$$\forall \epsilon \in \mathbb{R} \quad \forall t \in \mathbb{R}_+ \qquad \mathbb{P}(X \geq \epsilon) \leq e^{-t\epsilon} \boldsymbol{M}_X(t), \tag{3.3}$$

where $M_X(t) \equiv \mathbb{E}\left[e^{tX}\right]$ is the moment-generating function of X.
Proof.

$$\mathbb{P}(X \geq \epsilon) = \mathbb{P}\left(e^{tX} \geq e^{t\epsilon}\right) \underset{Markov}{\leq} \frac{\mathbb{E}\left[e^{tX}\right]}{e^{t\epsilon}} = e^{-t\epsilon}\mathbf{M}_X(t)$$

□

Note1 Markov's inequality and Chebyshev's inequality assume $\mathbb{E}[X] < \infty$ and $\mathbb{E}[X^2] < \infty$ respectively, and Chernoff bound assumes the existence of the moment generating function for $t > 0$. From Jensen's inequality, $\mathbb{E}[X^2] < \infty \rightarrow \mathbb{E}[X] < \infty$, but there are no order of the strongness between $\mathbb{E}[X] < \infty$ and $\mathbb{E}[e^{tX}] < \infty$ for all $t \in \mathbb{R}_+$ and between $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[e^{tX}] < \infty$ for all $t \in \mathbb{R}_+$. In fact, when X distributes from Log-normal distribution, $E[X^2]$ is finite, but $E[e^{tX}]$ can not be defined for any t > 0. On the other hand, when we define the distribution of X_1 and X_2 by

$$\begin{aligned}\mathbb{P}(X_1 = x) &\equiv \begin{cases} 2^{-k} & (when\ x = -2^k,\ k \in \mathbb{N}) \\ 0 & (others) \end{cases} \\ \text{and} \qquad \mathbb{P}(X_2 = x) &\equiv \begin{cases} 2^{-k} & (when\ x = -2^{\frac{k}{2}},\ k \in \mathbb{N}) \\ 0 & (others) \end{cases},\end{aligned}$$

$\mathbb{E}[e^{tX_1}] < \infty$, but $\mathbb{E}[X_1]$ doesn't exist. $\mathbb{E}[e^{tX_2}] < \infty$ and $\mathbb{E}[X_2] < \infty$, but $\mathbb{E}[X_2]$ doesn't exist. So, the inclusion relationships are described by the following figure.

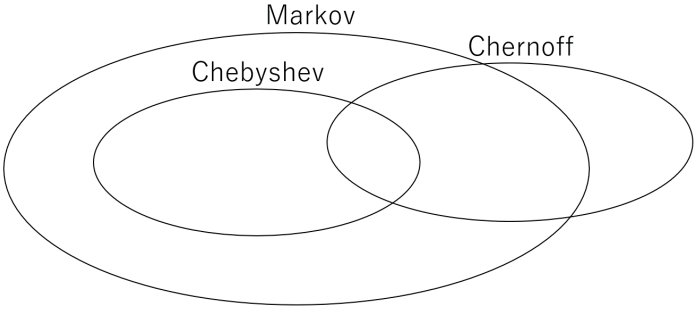


Figure 3.1 The order of the strongness of each assumptions

Note2 (3.3) holds for all $t > 0$, so we have

$$\mathbb{P}(X \geq \epsilon) \leq \inf_{t>0} e^{-t\epsilon} M_X(t) \tag{3.4}$$

And the following corollary is used later.

Corollary 3.1.4 *For $S_n \equiv \sum_{i=1}^n X_i$ with $(X_i)_{i=1}^n$ independent and for any $\epsilon \in \mathbb{R}$*

$$\begin{aligned}\mathbb{P}(S_n - E(S_n) \geq \epsilon) &\underset{(3.3)}{\leq} e^{-t\epsilon}\mathbb{E}\left[\exp\left(t\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right)\right] \\ &= e^{-t\epsilon}\left(\mathbb{E}\left[e^{t(X_1-\mathbb{E}[X_1])}\right]\right)^n \\ &= e^{-t\epsilon}\boldsymbol{M}_{X_1-\mathbb{E}[X_1]}(t)^n.\end{aligned}$$

Now, the problem of finding an upper bound for the probability can be paraphrased by the problem of finding an upper bound for the moment generating function.

Concentration inequality for sub-Gaussian Distribution Next let's consider a specific class of distribution. One of the most famous distribution is Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. In this case $\boldsymbol{M}_{X-\mathbb{E}[X]}(t) = \exp\left(\frac{\sigma^2 t^2}{2}\right)$. We define sub-Gaussian distribution as an extension of Gaussian distribution.

Definition 3.1.5 (sub-Gaussian distribution, introduced in [12]) *Let X to be a random variable with finite mean. X is called sub-Gaussian with variance proxy σ^2 if*

$$\forall t \in \mathbb{R} \qquad \boldsymbol{M}_{X-\mathbb{E}[X]}(t) = \mathbb{E}\left[e^{t(X-\mathbb{E}[X])}\right] \leq \exp\left(\frac{\sigma^2 t^2}{2}\right). \tag{3.5}$$

The following lemma states that Sub-Gaussian property is closed under algebraic operations.

Lemma 3.1.6 (Algebraic properties of sub-Gaussian variables) *If X_1 and X_2 are independent sub-Gaussian random variables with variance proxies σ_1^2 and σ_2^2, then*

- *$X_1 + X_2$ is sub-Gaussian with variance* proxy $\sigma_1^2 + \sigma_2^2$.
- *cX_1 is sub-Gaussian with variance proxy $c^2\sigma^2$ for any constant $c \neq 0$.*

Proof.
For any $t \in \mathbb{R}$

$$\begin{aligned}\mathbb{E}\left[e^{t(X_1+X_2-\mathbb{E}[X_1+X_2])}\right] &= \mathbb{E}\left[e^{t(X_1-\mathbb{E}[X_1])}e^{t(X_2-\mathbb{E}[X_2])}\right] \\ &= \mathbb{E}\left[e^{t(X_1-\mathbb{E}[X_1])}\right]\mathbb{E}\left[e^{t(X_2-\mathbb{E}[X_2])}\right] \quad \text{independenceness} \\ &\leq \exp\left(\frac{\sigma_1^2 t^2}{2}\right)\exp\left(\frac{\sigma_2^2 t^2}{2}\right) \\ &= \exp\left(\frac{(\sigma_1^2+\sigma_2^2)t^2}{2}\right),\end{aligned}$$

so, $X_1 + X_2$ is sub-Gaussian with variance proxy $\sigma_1^2 + \sigma_2^2$. For any $t \in \mathbb{R}$

$$\mathbb{E}\left[e^{t(cX-\mathbb{E}[cX])}\right] = \mathbb{E}\left[e^{(tc)(X-\mathbb{E}[X])}\right] \leq \exp\left(\frac{\sigma^2(tc)^2}{2}\right) = \exp\left(\frac{\left(c^2\sigma^2\right)t^2}{2}\right),$$

so cX is sub-Gaussian with variance proxy $c^2\sigma^2$.

□

Theorem 3.1.7 (Concentration inequality for sub-Gaussian distribution) *Let X be sub-Gaussian random variable with the variance proxy σ^2. Then for any $\epsilon > 0$*

$$\mathbb{P}(X - \mathbb{E}[X] \geq \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

$$\mathbb{P}(\mathbb{E}[X] - X \geq \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

Proof.
Applying Markov's inequality and using sub-Gaussian property,

$$\begin{aligned}\mathbb{P}(X - \mathbb{E}[X] \geq \epsilon) &\leq \frac{\mathbb{E}\left[e^{t(X-\mathbb{E}[X])}\right]}{e^{t\epsilon}} \leq \exp\left(\frac{\sigma^2 t^2}{2} - t\epsilon\right) \\ \mathbb{P}(\mathbb{E}[X] - X \geq \epsilon) &= \mathbb{P}\left(e^{-t(X-\mathbb{E}[X])} \geq e^{t\epsilon}\right) \leq \frac{\mathbb{E}\left[e^{-t(X-\mathbb{E}[X])}\right]}{e^{t\epsilon}} \leq \exp\left(\frac{\sigma^2 t^2}{2} - t\epsilon\right)\end{aligned}$$

This holds for every $t > 0$, and $\exp\left(\frac{\sigma^2 t^2}{2} - t\epsilon\right)$ take the minimum value $\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$ at $t = \epsilon/\sigma^2$.

□

The following two-side bound can be derived directly as a consequence of theorem 3.1.7.

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \mathbb{P}(X - \mathbb{E}[X] \geq \epsilon) + \mathbb{P}(\mathbb{E}[X] - X \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right) \tag{3.6}$$

Corollary 3.1.8 *Suppose $X_1, \ldots, X_n$ are independent sub-Gaussian random variables with variance proxies $\sigma_1^2, \ldots, \sigma_n^2$. Let $S_n = \sum_{i=1}^n X_i$ and $\bar{X}_n = \frac{1}{n} S_n$. Then for any $\epsilon \geq 0$*

$$\mathbb{P}\left(|S_n - \mathbb{E}\left[S_n\right]| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{2\sum_{i=1}^n \sigma_i^2}\right), \tag{3.7}$$

and

$$\mathbb{P}\left(\left|\bar{X}_n - \mathbb{E}\left[\bar{X}_n\right]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{n^2\epsilon^2}{2\sum_{i=1}^n \sigma_i^2}\right). \tag{3.8}$$

Proof.
Lemma 3.1.6 implies that S_n and $\bar{X}_n$ are also sub-Gaussian, and the variance proxies of S_n and $\bar{X}_n$ are $\sum_{i=1}^n \sigma_i^2$ and $\frac{1}{n^2}\sum_{i=1}^n \sigma_i^2$. Then, applying (3.6), we get the statement.

□

Note one-sided versions of the inequalities above also hold without the leading factor of 2. By taking a common σ for $\sigma_1^2, \ldots, \sigma_n^2$, we get the following.

Corollary 3.1.9 *Let X be a sub-Gaussian random variable with variance proxy σ^2 and $X_1, \ldots, X_n$ be independent copies of X. Let $S_n = \sum_{i=1}^n X_i$ and $\bar{X}_n = \frac{1}{n} S_n$. Then for any $\epsilon > 0$*

$$\mathbb{P}\left(|S_n - \mathbb{E}\left[S_n\right]| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{2n\sigma^2}\right)$$

and

$$\mathbb{P}\left(\left|\bar{X}_n - \mathbb{E}\left[\bar{X}_n\right]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)$$

By taking the compliment, we get the following equation which are equivalent to (3.7) and (3.8)

$$\mathbb{P}\left(|S_n - \mathbb{E}\left[S_n\right]| \leq \epsilon\right) \geq 1 - 2\exp\left(-\frac{\epsilon^2}{2\sum_{i=1}^n \sigma_i^2}\right) \tag{3.9}$$

and

$$\mathbb{P}\left(\left|\bar{X}_n - \mathbb{E}\left[\bar{X}_n\right]\right| \leq \epsilon\right) \geq 1 - 2\exp\left(-\frac{n^2\epsilon^2}{2\sum_{i=1}^n \sigma_i^2}\right) \tag{3.10}$$

These equations show the probability with which the deviances of S_n and $\bar{X}_n$ from the mean are smaller than the given ϵ. Now, by reparametrizing ϵ to t by $\epsilon \equiv \sqrt{2t\sum_{i=1}^n \sigma_i^2}$, we get

$$\forall t > 0 \quad \mathbb{P}\left(|S_n - \mathbb{E}[S_n]| \leq \sqrt{2t\sum_{i=1}^n \sigma_i^2}\right) \geq 1 - 2\exp(-t) \tag{3.11}$$

In case that $(X_i)_{i\in[n]}$ are i.i.d.,

$$\forall t > 0 \quad \mathbb{P}\left(|S_n - \mathbb{E}[S_n]| \leq \sqrt{2tn\sigma^2}\right) \geq 1 - 2\exp(-t) \tag{3.12}$$

Note This inequality implies that S_n have a deviation of order $O\left(\sqrt{n}\right)$ with exponential decay in probability.

Next we consider equivalent condition of sub-Gaussian property. In fact, (3.6) is not only an necessary condition but also an equivalent condition to sub-Gaussian. Furthermore, the following statement of sub-Gaussian holds, which provides us with other definitions of sub-Gaussian random variable.

Lemma 3.1.10 (Equivalent conditions for variable to be sub-Gaussian) *For a random variable X with $\mathbb{E}[X] < \infty$, the following statements are equivalent.*

(a) $\exists\sigma^2 > 0 \quad \forall t \in \mathbb{R} \qquad \boldsymbol{M}_{X-\mathbb{E}[X]}(t) = \mathbb{E}\left[e^{t(X-\mathbb{E}[X])}\right] \leq \exp\left(\frac{\sigma^2 t^2}{2}\right)$

(b) $\forall\epsilon \geq 0 \qquad \mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$

(c) $\sup_{p\geq 1} p^{-1/2}\left(\mathbb{E}|X - \mathbb{E}[X]|^p\right)^{1/p} < \infty$

Proof.

(a) $\Rightarrow$ (b) : From Theorem 3.1.7.

(b) $\Rightarrow$ (c) : It is enough to show case when $\mathbb{E}[X] = 0, \sigma^2 = 1$

$$\begin{aligned}
\mathbb{E}[|X|^p] &= \int_0^\infty \mathbb{P}(|X|^p \geq u)du \\
&= \int_0^\infty \mathbb{P}(|X| \geq t)pt^{p-1}dt && \text{by taking } u \equiv t^p \\
&\leq \int_0^\infty 2p\exp(-\frac{t^2}{2})t^{p-1}dt && \text{from the assumption of (b)} \\
&= 2p\int_0^\infty \exp(-s)s^{\frac{p}{2}-1}ds && \text{by taking s} \equiv \tfrac{t^2}{2} \\
&= 2p\Gamma\left(\frac{p}{2}\right) \\
&\leq 2pe\sqrt{\frac{p}{2}}\left(\frac{p}{2e}\right)^{\frac{p}{2}}. && \text{Stirling's approximation}
\end{aligned}$$

Taking $p^{-1/2}(\cdot)^{\frac{1}{p}}$ for both sides,

$$p^{-1/2}\left(\mathbb{E}|X|^p\right)^{1/p} \leq \left(2ep\sqrt{\frac{p}{2}}\right)^{\frac{1}{p}}\left(\frac{1}{2e}\right)^{\frac{1}{2}} < \infty.$$

(c) $\Rightarrow$ (a) : From the assumption of (c), $\mathbb{E}[X] < \infty$, and $\gamma \equiv \|X\|_{\psi_2} < \infty$. It is enough to show (a) in case that $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = 1$. For any $t \in \mathbb{R}$

$$\begin{aligned}
\mathbb{E}[\exp(tX)] &= 1 + t\underbrace{\mathbb{E}[X]}_{=0} + \sum_{p=2}^{\infty} \frac{t^p \mathbb{E}X^p}{p!} \\
&\leq 1 + \sum_{p=2}^{\infty} \frac{\gamma t^p p^{p/2}}{p!} && \text{from the assumption of (c)} \\
&\leq 1 + \sum_{p=2}^{\infty} \left(\frac{\gamma e\,|t|}{\sqrt{p}}\right)^p && \text{from} \quad p! \geq (p/e)^p \\
&\leq 1 + \sum_{p \in 2\mathbb{N}} \left(\frac{C\,|t|}{\sqrt{p/2}}\right)^2 && \text{taking } C \text{ enough large} \\
&= 1 + \sum_{k=1}^{\infty} \left(\frac{C\,|t|}{\sqrt{k}}\right)^{2k} \\
&\leq 1 + \sum_{k=1}^{\infty} \frac{(C\,|t|)^{2k}}{k!} \\
&= \exp(C^2 t^2)
\end{aligned}$$

□

We introduce the ψ_2 norm by $\|X\|_{\psi_2} \equiv \sup_{p \geq 1} p^{-1/2} \left(\mathbb{E}|X|^p\right)^{1/p} < \infty$. Then (c) in Lemma 3.1.10 can be written as

$$\|\mathbb{E}|X - \mathbb{E}[X]\|_{\psi_2} < \infty$$

Next, We define a similar norm $\|\cdot\|_{\psi_1}$ and a new category of random variables, which is called *sub-exponential*.

Definition 3.1.11 (ψ_1 norm and Sub-exponential random variables) *For a random variable X, we define $\|\cdot\|_{\psi_1}$ as*

$$\|X\|_{\psi_1} \equiv \sup_{p \geq 1} p^{-1} \left(\mathbb{E}|X|^p\right)^{1/p}$$

We call X sub-exponential random variable if

$$\|X - \mathbb{E}[X]\|_{\psi_1} < \infty$$

We use the following lemmas related to sub-exponential distribution in the subsequent chapter.

Lemma 3.1.12 *A random variable X is sub-exponential if and only if there exist $C > 0$ and $c > 0$ such that*

$$|t| \leq c/\|X - \mathbb{E}[X]\|_{\psi_1} \;\Rightarrow\; \mathbb{E}\exp(t(X - \mathbb{E}[X])) \leq \exp\left(Ct^2 \|X - \mathbb{E}[X]\|_{\psi_1}^2\right).$$

where $C, c > 0$ are absolute constants.

Outline of the proof
It's enough to show the statement in case that the distribution is centered, and We can assume that $\|X\|_{\psi_1} = 1$ since we can show general cases by replacing X with $X/\|X\|_{\psi_1}$ and t with $t\|X\|_{\psi_1}$. From the definition of $\|\cdot\|_{\psi_1}$, $\mathbb{E}|X|^p \leq p^p$, and Taylor expansion provides us with $\mathbb{E}\exp(tX) \leq 1 + \sum_{p=2}^{\infty} \frac{t^p p^{p/2}}{p!}$. And we can check this is bounded by $1 + \sum_{p=2}^{\infty} (e|t|)^p$. Moreover, this is bounded by $1 + 2e^2t^2 \leq \exp\left(2e^2t^2\right)$ when $|t| \leq 1/2e$. For the details, see Lemma 5.15 in [18]

Lemma 3.1.13 (Equivalence between sub-exponential and sub-Gaussian) *A random variable X is sub-Gaussian if and only if X^2 is sub-exponential. Moreover,*

$$\|X\|_{\psi_2}^2 \leq \|X^2\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2$$

Proof.
The first inequality is derived from the definitions of $\|\cdot\|_{\psi_1}$ and $\|\cdot\|_{\psi_2}$ amd Jensen's inequality. The second is from $\|\cdot\|_{\psi_2} = \sup_{p\geq 1} p^{\frac{1}{2}} \left(\mathbb{E}|X|^p\right)^{1/p}$

Hoeffding's inequality What is the condition for X to be sub-Gaussian? A sufficient condition is that X should be bounded.

Lemma 3.1.14 (Hoeffding's lemma, see [13]) *When a random variable X is bounded i.e. $a \leq X \leq b$ a.s. Then X is sub-Gaussian with variance proxy $\frac{(b-a)^2}{4}$, i.e. for any $t \in \mathbb{R}$*

$$\mathbb{E}\left[e^{t(X-\mathbb{E}[X])}\right] \leq \exp\left(\frac{(b-a)^2t^2}{8}\right).$$

Proof.
First, we prove for the case that $\mathbb{E}[X] = 0$, note that $a \leq 0 \leq b$. From the convexity of $x \mapsto e^{tx}$ with respect to x

$$\forall t \in \mathbb{R} \quad \forall x \in [a,b] \qquad e^{tx} \leq \frac{b-x}{b-a}e^{ta} + \frac{x-a}{b-a}e^{tb}.$$

By taking the expectation,

$$\begin{aligned}
\mathbb{E}\left[e^{tX}\right] &\leq \tfrac{b}{b-a}e^{ta} + \tfrac{-a}{b-a}e^{tb} \\
&= (1-\theta)e^{ta} + \theta e^{tb} && \theta \equiv -a/(b-a) \leq 0 \\
&= \left(1-\theta+\theta e^{t(b-a)}\right)e^{ta} \\
&= \left(1-\theta+\theta e^{u}\right)e^{-\theta u} && u \equiv t(b-a) \\
&= e^{\phi(u)} && \phi(u) \equiv \log\left(\left(1-\theta+\theta e^{u}\right)e^{-\theta u}\right) = \log\left(1-\theta+\theta e^{u}\right) - \theta u
\end{aligned}$$

From the arbitrariness of t, we can also arbitrarily choose u. And $\psi(u)$ is well-defined. This is because $\theta > 0$, so $1-\theta+\theta e^u = \theta\left(\frac{1}{\theta} - 1 + e^u\right) = \theta\left(-\frac{b}{a} + e^u\right) > 0$. Thus, all we need to do is to find the bound of $\phi(u)$. We can show the bound as followings

$$\forall u \in \mathbb{R}\ \exists v \in [0,u] \quad \text{s.t.} \quad \phi(u) = \phi(0) + u\phi'(0) + \frac{1}{2}u^2\phi''(v) \qquad \text{from Taylor's theorem}$$

$$\begin{aligned}
\phi(0) &= \log(1-\theta+\theta) - 0 = 0 \\
\phi'(0) &= \frac{\theta e^u}{1-\theta+\theta e^u} - \theta\big|_{u=0} = \theta - \theta = 0 \\
\phi''(v) &= \frac{\theta e^v\left(1-\theta+\theta e^v\right) - \theta^2 e^{2v}}{\left(1-\theta+\theta e^v\right)^2} = \frac{\theta e^v}{1-\theta+\theta e^v}\left(1 - \frac{\theta e^v}{1-\theta+\theta e^v}\right) \leq \frac{1}{4}
\end{aligned}$$

The last inequality is derived from the inequality that $a(1-a) \leq \frac{1}{4}$ for $a \in [0,1]$. Then, we get

$$\phi(u) \leq 0 + u\cdot 0 + \frac{1}{2}u^2\cdot\frac{1}{4} = \frac{u^2}{8} = \frac{(b-a)^2t^2}{8}.$$

Combining this bound and $\mathbb{E}\left[e^{tX}\right] \leq e^{\phi(u)}$,

$$\mathbb{E}\left[e^{tX}\right] \leq e^{\phi(u)} \leq \exp\left(\frac{(b-a)^2t^2}{8}\right)$$

In the case of general X (in which the mean may not equal 0), we apply the aforementioned argument to $Y \equiv X - \mathbb{E}[x]$. Since $\mathbb{E}[Y] = 0$ and $a - \mathbb{E}[X] \leq Y \leq b - \mathbb{E}[X]$ a.s.

$$\mathbb{E}\left[e^{t(X-\mathbb{E}[X])}\right] = \mathbb{E}\left[e^{tY}\right] \leq \exp\left(\frac{(b-\mathbb{E}[X]-(a-\mathbb{E}[X]))^2t^2}{8}\right) = \exp\left(\frac{(b-a)^2t^2}{8}\right).$$

□

From the Hoeffding's Lemma and (3.1.8), we immediately get the following theorem.

Theorem 3.1.15 (Hoeffding's inequality, see [13]) *Suppose $X_1, \ldots, X_n$ are independent random variables such that $a_i \leq X_i \leq b_i$ almost surely for each $i = 1, \ldots, n$. Let $S_n = \sum_{i=1}^n X_i$ and $\bar{X}_n = \frac{1}{n} S_n$. Then*

$$\mathbb{P}\left(|S_n - \mathbb{E}\left[S_n\right]| \geq \epsilon\right) \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n \left(b_i - a_i\right)^2}\right)$$

and

$$\mathbb{P}\left(\left|\bar{X}_n - \mathbb{E}\left[\bar{X}_n\right]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n \left(b_i - a_i\right)^2}\right)$$

Bernstein's inequality Hoeffding's inequality is a well-known concentration inequality, and another famous concentration inequality is the following.

Theorem 3.1.16 (Bernstein's inequality, see [10]) *Let $X_1, \ldots, X_n$ be independent real-valued random variables with $\mathbb{E}[X_i] = 0$ and $X_i \leq 1$ a.s. for all $i \in [n]$. Define*

$$\sigma^2 \equiv \frac{1}{n}\sum_{i=1}^n \mathrm{Var}\left\{X_i\right\}.$$

Then,

$$\forall \epsilon \geq 0 \qquad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i > \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2\left(\sigma^2 + \epsilon/3\right)}\right).$$

Moreover, if $|X_i| \leq 1$, then

$$\forall \epsilon \geq 0 \qquad \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| > \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2\left(\sigma^2 + \epsilon/3\right)}\right).$$

To prove (3.1.16), we use the following lemma.

Lemma 3.1.17 (Bennett's inequality, see [11]) *Let $X_1, \ldots, X_n$ be independent real-valued random variables s.t. $\mathbb{E}[X_i] = 0$ and $X_i \leq 1$ a.s. for all $i \in [n]$ Let*

$$\sigma^2 \equiv \frac{1}{n}\sum_{i=1}^n \mathrm{Var}\left\{X_i\right\}.$$

Then,

$$\forall t \geq 0 \qquad \mathbb{P}\left\{\sum_{i=1}^n X_i > t\right\} \leq \exp\left(-n\sigma^2 h\left(\frac{t}{n\sigma^2}\right)\right).$$

where $h(u) \equiv (1+x)\log(1+x) - x$ for $x \geq 0$. Moreover, if $|X_i| \leq 1$, then

$$\forall t \geq 0 \qquad \mathbb{P}\left\{\left|\sum_{i=1}^n X_i > t\right|\right\} \leq 2\exp\left(-n\sigma^2 h\left(\frac{t}{n\sigma^2}\right)\right).$$

Proof of Bennett's inequality

Let $\psi(x) \equiv \exp(x) - x - 1$, then ψ satisfy the following.

$$\psi(x) \geq x^2/2 \quad for\ x \geq 0 \tag{3.13}$$

$$\psi(x) \leq x^2/2 \quad for\ x \leq 0 \tag{3.14}$$

$$\psi(sx) \leq x^2\psi(s) \quad for\ s \geq 0\ and\ x \in [0,1] \tag{3.15}$$

Then, for $s \geq 0$

$$\begin{aligned}
\mathbb{E}\left[e^{sX_i}\right] &= 1 + s\underbrace{\mathbb{E}\left[X_i\right]}_{=0} + \mathbb{E}\left[\psi\left(sX_i\right)\right] \\
&= 1 + \mathbb{E}\left[\psi\left(s\left(X_i\right)_+\right) + \psi\left(-s\left(X_i\right)_-\right)\right] \quad (\text{ where } x_+ = \max(0,x) \text{ and } x_- = \max(0,-x)) \\
&\leq 1 + \mathbb{E}\left[\psi\left(s\left(X_i\right)_+\right) + \frac{s^2}{2}\left(X_i\right)_-^2\right] \qquad \text{from (3.14)} \\
&\leq 1 + \mathbb{E}\left[\psi(s)\left(X_i\right)_+^2 + \frac{s^2}{2}\left(X_i\right)_-^2\right] \qquad \text{from (3.21) and } 0 \leq \left(X_i\right)_+ \leq 1 \\
&= 1 + \mathbb{E}\left[\psi(s)\left(X_i\right)^2 - \psi(s)\left(X_i\right)_-^2 + \frac{s^2}{2}\left(X_i\right)_-^2\right] \\
&\leq 1 + \psi(s)\mathbb{E}\left[X_i^2\right] - \underbrace{\mathbb{E}\left[\left\{\psi(s) - \frac{s^2}{2}\right\}\left(X_i\right)_-^2\right]}_{\geq 0 \text{ from } 3.14} \\
&\leq \exp\left(\psi(s)\mathbb{E}\left[X_i^2\right]\right)
\end{aligned}$$

Applying corollary 3.1.4 to $(X_i)_{i=1}^n$,

$$\forall t \in \mathbb{R} \qquad \mathbb{P}\left\{\sum_{i=1}^n X_i > t\right\} \leq \exp\left(\sum_{i=1}^n \mathbb{E}\left[X_i^2\right]\psi(s) - st\right) \leq e^{n\sigma^2\psi(s)-st}.$$

This holds for arbitrary $s \in \mathbb{R}_+$. The upper bound is minimized at

$$s = \log\left(1 + \frac{t}{n\sigma^2}\right),$$

and the minimum value is the left hand side of the statement. Repeating this argument for $-X_i$ instead of X_i, we obtain the same upper bound for $\mathbb{P}\left\{-\sum_{i=1}^n X_i > t\right\}$, so

$$\mathbb{P}\left\{\left|\sum_{i=1}^n X_i > t\right|\right\} = \mathbb{P}\left\{\sum_{i=1}^n X_i > t\right\} + \mathbb{P}\left\{-\sum_{i=1}^n X_i > t\right\} \leq 2\exp\left(-n\sigma^2 h\left(\frac{t}{n\sigma^2}\right)\right)$$

□

Proof of Bernstein's inequality

For $x \geq 0$, $h(x) \geq x^2/(2 + 2x/3)$ since $F(x) \equiv h(x) - x^2/(2 + 2x/3)$ takes 0 at $x = 0$ and $\frac{dF}{dx} > 0$ for $x \geq 0$. By applying Bennett's inequality to
$\sum_{i=1}^n X_i$, we get

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i > t\right) = \mathbb{P}\left(\sum_{i=1}^n X_i > tn\right) \underset{Bennett}{\leq} \exp\left(-n\sigma^2 h\left(\frac{t}{n\sigma^2}\right)\right) \leq \exp\left(-\frac{nt^2}{2\left(\sigma^2 + t/3\right)}\right).$$

□

Two sided inequality is proved in the same way as Bennett's inequality. By reparametrization , we obtain the following corollary.

Corollary 3.1.18 *For i.i.d. random variables* $X_1, \ldots, X_n$ *with* $\mathbb{E}[X_i] = 0$ *and* $|X_i| \leq 1$ *for all* $i \in [n]$,

$$\forall \delta > 0 \qquad \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| < \frac{2\ln\frac{2}{\delta}}{3n} + \sqrt{\frac{2\sigma^2\ln\frac{2}{\delta}}{n}}\right) \geq 1 - \delta. \tag{3.16}$$

Proof.
Take

$$t \equiv \frac{\frac{2}{3}\log(\frac{2}{\delta}) + \sqrt{\left(\frac{2}{3}\log(\frac{2}{\delta})\right)^2 + 8n\delta^2\log(\frac{2}{\delta})}}{2n}.$$

Note that

$$2\exp\left(-\frac{nt^2}{2\left(\sigma^2 + t/3\right)}\right) = \delta.$$

since t is the positive solution of $nt^2 - \frac{2}{3}\ln\left(\frac{2}{\delta}\right)t - 2\sigma^2\ln\left(\frac{2}{\delta}\right) = 0$. And the following inequality holds.

$$t \leq \frac{2\ln\frac{2}{\delta}}{3n} + \sqrt{\frac{2\sigma^2\ln\frac{2}{\delta}}{n}}$$

From this inequality

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i\right| < \frac{2\ln\frac{2}{\delta}}{3n} + \sqrt{\frac{2\sigma^2\ln\frac{2}{\delta}}{n}}\right) &\geq \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i\right| \leq t\right) \\
&\geq 1 - 2\exp\left(-\frac{nt^2}{2\left(\sigma^2 + t/3\right)}\right) \qquad \text{from the theorem 3.1.16} \\
&= 1 - \delta
\end{aligned}$$

□

Theorem 3.1.19 (Bernstein-type inequality) *Let $X_1, \ldots, X_n$ be i.i.d. centered sub-exponential random variables, and $K \equiv \max_i \|X_i\|_{\psi_1}$. Then,*

$$\exists c_1 > 0 \quad \forall t \geq 0 \qquad \mathbb{P}\left(\left|\sum_{i=1}^{n}X_i\right| \geq t\right) \leq 2\exp\left[-c_1\min\left(\frac{t^2}{K^2n}, \frac{t}{K}\right)\right] \tag{3.17}$$

Proof.
At first we prove the case in which $\|X_1\|_{\psi_1} = K = 1$ Define $S \equiv \sum_{i=1}^{n}X_i$. By taking c > 0 satisfying Lemma 3.1.12,

$$\begin{aligned}
\mathbb{P}\left(S \geq t\right) &\leq e^{-\lambda t}(\mathbb{E}[e^{\lambda S}])^n \qquad \textit{Corollary } 3.1.4 \\
&\leq e^{-\lambda t}\exp\left(nC\lambda^2\right) \qquad \textit{Lemma } 3.1.12 \\
&= \exp\left(-\lambda t + Cn\lambda^2\right).
\end{aligned}$$

Choosing $\lambda = \min\left(t/2nC, c\right)$, we obtain

$$\begin{aligned}
\mathbb{P}(S \geq t) &\leq \begin{cases} \exp\left(-\frac{t^2}{4nC}\right) & \text{if } \frac{t}{2nC} \leq c \\ \exp\left(-ct + Cc^2n\right) & \text{else} \end{cases} \\
&\leq \begin{cases} \exp\left(-\frac{t^2}{4nC}\right) & \text{if } \frac{t}{2nC} \leq c \\ \exp\left(-\frac{ct}{2}\right) & \text{else} \end{cases} \\
&\leq \exp\left[-\min\left(\frac{t^2}{4nC}, \frac{ct}{2}\right)\right].
\end{aligned}$$

Note that in the second inequality we use that $\frac{t}{2nC} > c \Leftrightarrow \frac{ct}{2} > Cc^2n$, and under this condition $-ct + Cc^2n < -\frac{ct}{2}$. Repeating the above argument for $-S$ instead of S, we obtain the same bound for $\mathbb{P}(-S \geq t)$. Then,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) = \mathbb{P}(S \geq t) + \mathbb{P}(-S \geq t) \leq 2\exp\left[-\min\left(\frac{t^2}{4nC}, \frac{ct}{2}\right)\right]$$

Taking $c_1 \equiv \left(\frac{1}{4C} \wedge \frac{c}{2}\right)$, we get

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2\exp\left[-c_1 \min\left(\frac{t^2}{n}, t\right)\right]. \tag{3.18}$$

For general K,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) = \mathbb{P}\left(\left|\sum_{i=1}^{n} \frac{X_i}{K}\right| \geq \frac{t}{K}\right) \underset{(3.18)}{\leq} 2\exp\left[-c_1 \min\left(\frac{t^2}{K^2 n}, \frac{t}{K}\right)\right].$$

□

By replacing t to tn, we obtain the following inequality regarding sample average.

Corollary 3.1.20 (Bernstein-type inequality) *Let $X_1, \ldots, X_n$ be i.i.d. centered sub-exponential random variables, and $K \equiv \max_i \|X_i\|_{\psi_1}$. Then,*

$$\exists c_1 > 0 \quad \forall t \geq 0 \qquad \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2\exp\left[-c_1 \min\left(\frac{t^2}{K^2}, \frac{t}{K}\right) n\right] \tag{3.19}$$

Next, we consider reformulating (3.19) to another representation like (3.1.18).

$$\begin{aligned}
2\exp\left[-c_1 \min\left(\tfrac{t^2n^2}{K^2}, \tfrac{tn}{K}\right)\right] = \epsilon \quad &\Leftrightarrow \quad c_1 \min\left(\tfrac{t^2n^2}{K^2}, \tfrac{tn}{K}\right) &&= \log\tfrac{2}{\epsilon} \\
&\Leftrightarrow \quad \min\left(\tfrac{t^2n^2}{K^2}, \tfrac{tn}{K}\right) &&= \tfrac{1}{c_1}\log\tfrac{2}{\epsilon} \\
&\Leftrightarrow \quad \tfrac{tn}{K} &&= \max\left(\tfrac{1}{c_1}\log\tfrac{2}{\epsilon}, \sqrt{\tfrac{1}{c_1}\log\tfrac{2}{\epsilon}}\right) \\
&\Leftrightarrow \quad t &&= \tfrac{K}{n}\max\left(\tfrac{1}{c_1}\log\tfrac{2}{\epsilon}, \sqrt{\tfrac{1}{c_1}\log\tfrac{2}{\epsilon}}\right)
\end{aligned}$$

Note that since for $\epsilon \in [0,1]$, $\frac{2}{\epsilon} > 1$ and $\log\frac{2}{\epsilon} > 0$ and we use that for $x \geq 0, \min(x, x^2) = y \Leftrightarrow x = \max(y, \sqrt{y})$ in the second line. Thus, by re-parametrization and taking the compliment, (3.19) can be written as

$$\exists c_1 > 0 \quad \forall 1\epsilon \in [0,1] \qquad \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \leq \frac{K}{n}\max\left(\frac{1}{c_1}\log\frac{2}{\epsilon}, \sqrt{\frac{1}{c_1}\log\frac{2}{\epsilon}}\right)\right) \geq 1 - \epsilon$$

Theorem 3.1.21 (See Corollary 1 in [2]) *Let $(\boldsymbol{X}_i)_{i\in[n]}$ be a sequence of independent random vectors in $\mathbb{R}^d$, define $S_n \equiv \sum_{i=1}^n \boldsymbol{X}_i$. Assume*

$$\begin{aligned}
&\forall i \in [n] \qquad \mathbb{E}\boldsymbol{X}_i = 0, \\
\exists B, L > 0 \quad 2 \leq \forall m \in \mathbb{N} \qquad &\sum_{i=1}^{n} \mathbb{E}\left|\boldsymbol{X}_i\right|^m \leq \frac{m! B^2 L^{m-2}}{2}.
\end{aligned} \tag{3.20}$$

Then,

$$\forall r \geq 0 \qquad \mathbb{P}_\theta(|S_n| \geq r) \leq 2\exp\left\{\frac{-r^2}{2B^2 + 2rL}\right\}. \tag{3.21}$$

Moreover,

$$\forall \epsilon \geq 0 \qquad \mathbb{P}_{\theta}(\left|\bar{X}\right| \geq \epsilon) \leq 2\exp\left\{\frac{-n^2\epsilon^2}{2B^2 + 2n\epsilon L}\right\}. \tag{3.22}$$

Outline of proof

$$\mathbb{P}_{\theta}(|S_n| \geq r) \leq e^{-\lambda r}\mathbb{E}(e^{r|S_n|}) \leq 2e^{-\lambda r}\mathbb{E}\cosh\lambda\,|S_n| \leq 2e^{-\lambda r}\prod_{j=1}^{n}\mathbb{E}(e^{\lambda|X_j|} - \lambda\left|X_j\right|)$$

From Taylor's expansion and the assumption, the last equals to $2e^{-\lambda r}\prod_{j=1}^{n}\mathbb{E}(\sum_{k=2}^{\infty}\frac{B^2L^{m-2}}{2})$. Calculating the sum of the geometric series and the λ minimizing it, we obtain (3.21). For the details, see Corollary 1 in [2].

Corollary 3.1.22 *Under the same assumption of (3.20)*

$$\forall \delta \geq 0 \qquad \mathbb{P}\left(\|S_n\| \leq 2L\log\frac{2}{\delta} + \sqrt{2B^2\log\frac{2}{\delta}}\right) \geq 1 - \delta. \tag{3.23}$$

Proof.
Let

$$t \equiv L\log\frac{2}{\delta} + \sqrt{\left(L\log\frac{2}{\delta}\right)^2 + 2B^2\log\frac{2}{\delta}}. \tag{3.24}$$

Note that

$$2\exp\left\{\frac{-r^2}{2B^2 + 2rL}\right\} = \delta \tag{3.25}$$

since r is the positive solution of $t^2 - 2L\log\frac{2}{\delta}\,t + 2B^2\log\frac{2}{\delta} = 0$. And the following inequality holds.

$$t \leq L\log\frac{2}{\delta} + \sqrt{\left(L\log\frac{2}{\delta}\right)^2} + \sqrt{2B^2\log\frac{2}{\delta}} = 2L\log\frac{2}{\delta} + \sqrt{2B^2\log\frac{2}{\delta}} \tag{3.26}$$

From this inequality

$$\begin{aligned}
\text{L.H.S of (3.27)} &\geq \mathbb{P}\left(\|S_n\| \leq t\right) && \text{from (3.26)}\\
&\geq 1 - 2\exp\left\{\frac{-t^2}{2B^2 + 2rL}\right\} && \text{from Theorem 3.1.21}\\
&= 1 - \delta && \text{from (3.25)}
\end{aligned}$$

□

Two specific case which satisfies the assumption of theorem 3.1.21

Case 1 : ($|X_i|$) are i.i.d bounded (and $\mathbb{E}X_1 = 0$)
When ($|X_1|$) are bounded : $L \equiv \sup|X_1| < \infty$, ($|X_i|$) also has a finite variance. So, take

$$B^2 \equiv \frac{n}{2}\mathbb{E}|X_1|^2 = \frac{n}{2}\mathbb{E}\left(\mathrm{Tr}X_1X_1^{\top}\right) = \frac{n}{2}\mathrm{Tr}\mathbb{E}\left(X_1X_1^{\top}\right) < \infty$$

X_1 fulfills (3.20). By Substituting $\mathrm{Tr}\mathbb{E}[X_1X_1^{\top}]$ for B^2, we obtain the next corollary.

Corollary 3.1.23 *Let $(\boldsymbol{X}_i)_{1\leq i\leq n} \in \mathbb{R}^d$ be a sequence of i.i.d random vectors. Define $S_n \equiv \sum_{i=1}^{n} X_i$. Assume that* $\sup|X_1| \equiv L < \infty$. *Then,*

$$\forall \delta > 0 \qquad \mathbb{P}\left(|S_n| \leq \sqrt{\mathrm{Tr}\mathbb{E}[X_1X_1^{\top}]}\sqrt{\log\frac{2}{\delta}} + 2L\log\frac{2}{\delta}\right) \geq 1 - \delta$$

Case 2 : ($|X_i|$) are i.i.d Gaussian (and $\mathbb{E}X_1 = 0$) The following lemma states that (Multivariate) normal distribution satisfies the assumption of Theorem 3.1.21 for $L = D$ and $B^2 = nD^2$.

Lemma 3.1.24 (Lemma : Case of Gaussian random vector) *Let a random vector $(X_i)_{i\in[n]}$ distribute from $\mathcal{N}(\mathbf{0}, \Sigma)$. Then*

$$\forall m \in \mathbb{N} \qquad \sum_{i=1}^{n} \mathbb{E}\,|X_i|^m \le nD^m \frac{m!}{2} \quad = \quad \frac{m! n D^2 D^{m-2}}{2}$$

where $D \equiv \mathrm{Tr}\Sigma$.

Proof.

When $X_i \sim X \sim \mathcal{N}(0, \Sigma)$, and Σ is represented as

$$\Sigma = \boldsymbol{Q}^\top \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)\boldsymbol{Q},$$

here $\sigma_j \in \mathbb{R}$ for $j \in [d]$ and $\boldsymbol{Q} \in \mathbb{R}^{d\times d}$ is an orthogonal matrix. Nete that $D \equiv \sqrt{\sum {\sigma_j}^2}$. Define $\tilde{\boldsymbol{e}}_i \equiv \boldsymbol{Q}^\top \boldsymbol{e}_i$ for all $i \in [d]$, where $(\boldsymbol{e}_i)_{i\in[d]}$ is the standard basis for $\mathbb{R}^d$. And we can write $\boldsymbol{X}$ as

$$X = \Sigma^{\frac{1}{2}} Y,$$

where $Y = \sum_{j=1}^{d} \xi_j \boldsymbol{e}_j$ and $\xi_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}_1(0,1)$ for all $j \in [d]$. Then,

$$\begin{array}{rcccccl} X & = & \sum_{j=1}^{d} \xi_j \Sigma^{\frac{1}{2}} \boldsymbol{e}_j & = & \sum_{j=1}^{d} \xi_j \boldsymbol{Q}^\top \mathrm{diag}(\sigma_1, \ldots, \sigma_d)\boldsymbol{e}_j & & \\ & = & \sum_{j=1}^{d} \xi_j \boldsymbol{Q}^\top \sigma_j \boldsymbol{e}_j & = & \sum_{j=1}^{d} \xi_j \sigma_j (\boldsymbol{Q}^\top \boldsymbol{e}_j) & = & \sum_{j=1}^{d} \sigma_j \xi_j \tilde{\boldsymbol{e}}_j \end{array}$$

Thus,

$$\begin{aligned} \sum_{i=1}^{n} E\,|X_i|^m &= nE\,|X|^m && (X_i)_{i\in[n]} \underset{i.i.d.}{\sim} X \\ &\le n\sqrt{\mathbb{E}\,|X|^{2m}} && \textit{Jensen's inequality} \\ &= n\sqrt{\mathbb{E}\left(\sum_{j=1}^{d} {\sigma_j}^2 {\xi_j}^2\right)^m} \\ &\le n\sqrt{\mathbb{E}\sum_{j=1}^{d} \frac{{\sigma_j}^2}{D^2}\left(D^2 {\xi_j}^2\right)^m} \\ &= n\sqrt{\sum_{j=1}^{d} \frac{{\sigma_j}^2}{D^2}\mathbb{E}\left(D^2 {\xi_j}^2\right)^m} && \textit{linearity of expectation} \\ &= n\sqrt{\mathbb{E}\left(D^2\xi_1^2\right)^m} && (\xi_j)\textit{'s i.i.d and the definition of } D^2 \\ &= nD^m\sqrt{\mathbb{E}\left(\xi_1^{2m}\right)} \\ &= nD^m\sqrt{2^m \prod_{k=1}^{m} \frac{2k-1}{2}} \\ &\le nD^m \frac{m!}{2} \quad = \quad \frac{m! n D^2 D^{m-2}}{2}. \end{aligned}$$

□

On the fourth line, we apply Jensen's inequality : $\left(\sum_{j=1}^d \frac{\sigma_j^2}{D^2} a_j\right)^m \leq \sum_{j=1}^d \frac{\sigma_j^2}{D^2} (a_j)^m$ for non negative values $(a_j)_{j\in[n]} \equiv (D^2\xi_j^2)_{j\in[n]}$. Thus, by applying Corollary 3.1.22, we get the following.

Corollary 3.1.25 *Let random vectors $(X_i)_{i\in[n]}$ distribute from $\mathcal{N}(\mathbf{0}, \Sigma)$. Then,*

$$\forall \delta \geq 0 \qquad \mathbb{P}\left(|S_n| \leq 2\mathrm{Tr}(\Sigma)\log\frac{2}{\delta} + \sqrt{2n\mathrm{Tr}(\Sigma)\log\frac{2}{\delta}}\right) \geq 1-\delta. \tag{3.27}$$

Moreover,

$$\forall \delta \geq 0 \qquad \mathbb{P}\left(\left|\bar{X}\right| \leq \frac{2}{n}\mathrm{Tr}(\Sigma)\log\frac{2}{\delta} + \sqrt{\frac{2}{n}\mathrm{Tr}(\Sigma)\log\frac{2}{\delta}}\right) \geq 1-\delta. \tag{3.28}$$

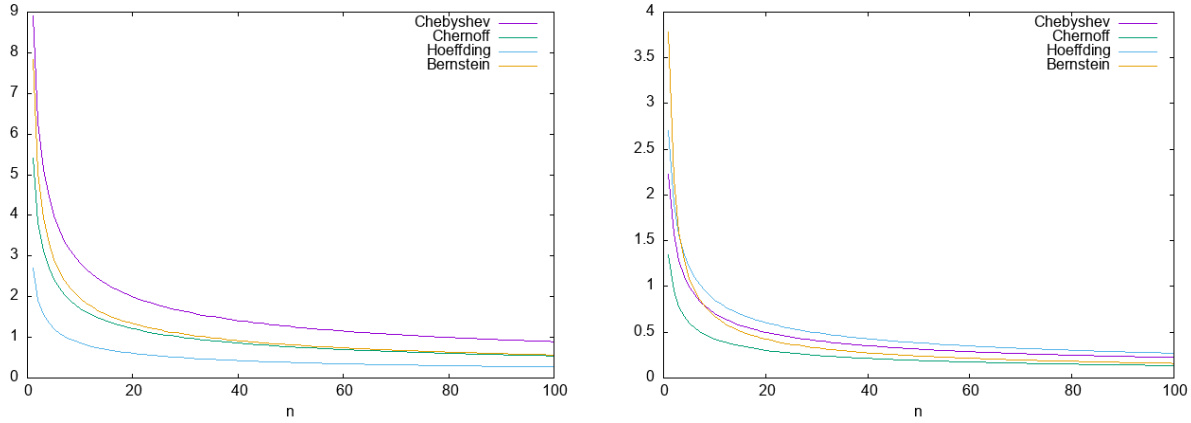
At the end of this subsection, we compare the above concentration inequalities : Chebyshev inequality, Chernoff bound (under sub-Gaussian assumption), Hoeffding inequality, and Bernstein inequiality. By re-parametrization, each inequality can be written as

$$\mathbb{P}\left(\left|\bar{X}_n - \mathbb{E}\left[\bar{X}_n\right]\right| \geq h(\delta, n)\right) \leq \delta,$$

where

$$\begin{aligned}
h(\delta,n) &\equiv \frac{\sigma}{\sqrt{n\delta}} && \text{for Chebyshev inequiality,}\\
h(\delta,n) &\equiv \sqrt{\frac{2\sigma^2\ln\frac{2}{\delta}}{n}} && \text{for Chernoff bound,}\\
h(\delta,n) &\equiv \sqrt{\frac{(b-a)^2\log(2/\delta)}{2n}} && \text{for Hoeffding inequiality,}\\
h(\delta,n) &\equiv \frac{2\ln\frac{2}{\delta}}{3n} + \sqrt{\frac{2\sigma^2\ln\frac{2}{\delta}}{n}}. && \text{for Bernstein inequiality,}
\end{aligned}$$

The following is the relation between the number of samples $(= n)$ and $h(\delta, n)$.



here, we draw the line of Bernstein inequality by setting $a_i = -1$ and $b_i = 1$ in order to compare it with Hoeffding inequality.

3.2 Concentration inequalities for covariance matrix of random vector

Next, we construct concentration inequalities for a random matrix. When the operator norm of the random matrix is bounded, the following probability inequality hold.

Theorem 3.2.1 (Bernstein's inequality due to Tropp(2015), see Theorem 6.6.1 in [4]) *Let* $(\boldsymbol{X}_i)_{1\leq i\leq n} \in \mathbb{R}^{d\times d}$ *be a sequence of independent, random, positive semidefinite Hermitian matrices. Assume*

$$\exists R > 0 \quad \forall k \in \mathbb{N}_0 \qquad \mathbb{E}\boldsymbol{X}_k = \boldsymbol{0} \quad \textit{and} \quad \|\boldsymbol{X}_k\| \leq R$$

Then, then it holds that

$$\forall t \geq 0 \qquad \mathbb{P}\left(\left\|\sum_k \boldsymbol{X}_k\right\| \geq t\right) \leq d \cdot \exp\left(\frac{-t^2/2}{\left\|\sum_k \mathbb{E}\boldsymbol{X}_k^2\right\| + Rt/3}\right).$$

The outline of the proof (For the details, see Theorem 6.6.1 in [4].)
At first, we consider a function f_θ on $\mathbb{R}$ satisfying $f_\theta(x) = \frac{e^{\theta x}-\theta x-1}{x^2}$ (for $x = 0$, define $f_\theta(0) = \frac{\theta^2}{2}$). Performing calculations provides us with $f_\theta(R) \leq \frac{\theta^2/2}{1-\theta R/3}$. From the assumption $\|\boldsymbol{X}_k\| \leq R$, we get $f(\boldsymbol{X}_k) \preccurlyeq f(R)\cdot I_d$. When, we obtain

$$\mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq I_d + \theta\boldsymbol{X} + \boldsymbol{X}(f_\theta(R)\cdot I_d)\boldsymbol{X} = I_d + \theta\boldsymbol{X} + f_\theta(R)\cdot\boldsymbol{X}^2 \preccurlyeq I_d + \theta\boldsymbol{X} + \frac{\theta^2/2}{1-\theta R/3}\cdot\boldsymbol{X}^2. \tag{3.29}$$

Taking the expectation for the both sides, from the assumption of $\mathbb{E}(\boldsymbol{X}_k) = 0$

$$\mathbb{E}\mathrm{e}^{\theta X} \preccurlyeq \mathbf{I} + \frac{\theta^2/2}{1-\theta R/3}\cdot\mathbb{E}X^2 \preccurlyeq \exp\left(\frac{\theta^2/2}{1-\theta R/3}\cdot\mathbb{E}X^2\right) \tag{3.30}$$

Then,

$$\begin{aligned}\mathbb{P}\left(\left\|\textstyle\sum_{i=1}^n X_i\right\| \geq t\right) \quad &\leq \quad \inf_{0<\theta} \quad \mathrm{e}^{-\theta t}\,\mathrm{tr}\exp\left(\textstyle\sum_i \log \mathbb{E}\mathrm{e}^{\theta \boldsymbol{X}_i}\right)\\ &\underset{\substack{\text{logarithm}\\ \text{of (3.30)}}}{\leq} \quad \inf_{0<\theta<3/R} \quad \mathrm{e}^{-\theta t}\,\mathrm{tr}\exp\left(\frac{\theta^2/2}{1-\theta R/3}\textstyle\sum_i \mathbb{E}X_i^2\right)\end{aligned}$$

□By taking $\theta \equiv \frac{t}{\sum_i \mathbb{E}X_i^2 + Rt/3}$, we obtain the statement.

Corollary 3.2.2 *Consider a sequence* $(X_k)_{1\leq k\leq n} \in \mathbb{R}^{d\times d}$ *of independent, random, Hermitian matrices. Assume that* $\mathbb{E}[X_k] = 0$ *and* $\forall k \in [n]$ $\|X_k\| \leq R$. *Then,*

$$\forall \delta \geq 0 \qquad \mathbb{P}\left(\left\|\sum_{k=1}^n X_k\right\| \leq \sqrt{2\left\|\sum_{k=1}^n \mathbb{E}\left[X_k^2\right]\right\|\log\left(\frac{d}{\delta}\right)} + \frac{2R}{3}\log\left(\frac{d}{\delta}\right)\right) \geq 1-\delta \tag{3.31}$$

Proof.
Let

$$t \equiv \frac{R}{3}\log\frac{d}{\delta} + \sqrt{\frac{R^2}{9}(\log\frac{d}{\delta})^2 + 2\left\|\sum_{k=1}^n \mathbb{E}X_k^2\right\|\log\frac{d}{\delta}} \tag{3.32}$$

Note that

$$d\exp\left(\frac{-t^2/2}{\left\|\sum_{k=1}^n \mathbb{E}X_k^2\right\| + Rt/3}\right) = \delta \tag{3.33}$$

since t is the positive solution of $\frac{t^2}{2} - \frac{R}{3}(\log\frac{d}{\delta})t - \left\|\sum_{k=1}^n \mathbb{E}X_k^2\right\| \log\frac{d}{\delta} = 0$. And the following inequality holds.

$$t \leq \frac{R}{3}\log\frac{d}{\delta} + \sqrt{\frac{R^2}{9}(\log\frac{d}{\delta})^2} + \sqrt{2\left\|\sum_{k=1}^n \mathbb{E}X_k^2\right\|\log\frac{d}{\delta}} = \frac{2L}{3}\log\frac{d}{\delta} + \sqrt{2\left\|\sum_{k=1}^n \mathbb{E}X_k^2\right\|\log\frac{d}{\delta}} \tag{3.34}$$

From this inequality

$$\begin{aligned}
\text{L.H.S of (3.31)} &\geq \mathbb{P}\left(\left\|\sum_{k=1}^n X_k\right\| \leq t\right) && \text{from (3.34)} \\
&\geq 1 - d\cdot\exp\left(\frac{-t^2/2}{\left\|\sum_{k=1}^n \mathbb{E}X_k^2\right\| + Rt/3}\right) && \text{from Theorem 3.2.1} \\
&= 1-\delta && \text{from (3.33)}
\end{aligned}$$

Taking $1/n$ for the both sides of (3.31), it holds with probability at least $1-\delta$ that

$$\left\|\frac{1}{n}\sum_{t=1}^n X_t\right\| \leq \sqrt{2\left\|\frac{1}{n^2}\sum_{t=1}^n \mathbb{E}{X_t}^2\right\|\log\left(\frac{d}{\delta}\right)} + \frac{2R}{3n}\log\left(\frac{d}{\delta}\right)$$

□Our interest in this thesis is the case in which we have i.i.d. sample $(X_i)_i \in [n] \in \mathbb{R}^d$, and the random matrix is represented as the sample covariance of $(X_i)_i \in [n]$ In order to discuss this topic, we introduce *Covariance operator.*

Definition 3.2.3 *Let $(E, \|\cdot\|)$ be a Hilbert space with the dual space E^*, $\boldsymbol{X}$ be a centered random vector in E with $\mathbb{E}|\boldsymbol{X}|^2 < \infty$, and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. copies of $\boldsymbol{X}$. The covariance operator $\Sigma : E^* \to E$ is defined as*

$$\Sigma u \equiv \mathbb{E}\left[\langle \boldsymbol{X}, u\rangle \boldsymbol{X}\right] \qquad u \in E^*.$$

And, the sample covariance operator $\hat{\Sigma} : E^* \to E$ is defined as

$$\hat{\Sigma} u \equiv n^{-1}\sum_{j=1}^n \left\langle \boldsymbol{X}_j, u\right\rangle \boldsymbol{X}_j, \quad u \in E^*$$

here, $\langle x, u\rangle$ denotes the value of $u \in E^*$ at $x \in E$.

Note Since E is a hilbert space, E^* is regarded as the same space as E by the canonical isomorphism. In the case when $E = \mathbb{R}^d$, the representation matrix of Σ with respect to the standard basis equals to $\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]$, that of $\hat{\Sigma}$ equals to the empirical covariance, $\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top$.
The goal of this section is creating upper bounds for the deviation $\hat{\Sigma}$ from Σ with some probability. In order to measure how for $\hat{\Sigma}$ deviates from Σ, we introduce *operator norm*

Definition 3.2.4 *For operator $A : E^* \to E$,*

$$\|A\| \equiv \sup_{u\in E^*, \|u\|\leq 1} \|Au\|$$

In case that $E = \mathbb{R}^d$, the covariance operator Σ w.r.t. X, $\|\Sigma\|$ is the maximum length of the eigen vectors of $Cov(X)$. For the random variable $\|\hat{\Sigma} - \Sigma\|$, the following inequality is known.

Theorem 3.2.5 *Let $\boldsymbol{X}$ in $\mathbb{R}^d$ to be Gaussian with $\mathbb{E}[\boldsymbol{X}] = 0$, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ to be i.i.d. copies of X. Then,*

$$\exists C > 0\ \forall t > 1 \qquad P\left(\|\hat{\Sigma} - \Sigma\| \leq C\|\Sigma\|\left(\sqrt{\frac{d}{n}} \bigvee \frac{d}{n} \bigvee \sqrt{\frac{t}{n}} \bigvee \frac{t}{n}\right)\right) \geq 1 - e^{-t} \tag{3.35}$$

Proof.

Step 1 : Simplifying the situation

Define $n \times d$ matrix $\tilde{X} \equiv (X_1, \ldots, X_n)^\top$, where $X_1, \ldots, X_n$ are independent copies of X. Then, $\hat{\Sigma} = \frac{1}{n}\tilde{X}^\top\tilde{X}$. Using this notation, (3.35) is rewritten as

$$\exists C > 0 \ \forall t > 1 \qquad P\left(\|\frac{1}{n}\tilde{X}^\top\tilde{X} - \Sigma\| \le C\|\Sigma\|\left(\sqrt{\frac{d}{n}} \bigvee \frac{d}{n} \bigvee \sqrt{\frac{t}{n}} \bigvee \frac{t}{n}\right)\right) \ge 1 - e^{-t} \tag{3.36}$$

To show (3.2), it is enough to show (3.35) only in case of $\Sigma \equiv I_d$, i.e. in case that X is isortopic. This can be checked as the followings.

$$\begin{aligned} \|\frac{1}{n}\tilde{X}^\top\tilde{X} - \Sigma\| &= \|\Sigma^{\frac{1}{2}\top}(\frac{1}{n}\Sigma^{-\frac{1}{2}\top}\tilde{X}^\top\tilde{X}\Sigma^{-\frac{1}{2}} - I_d)\Sigma^{\frac{1}{2}}\| \\ &\le \|\Sigma\|\|\frac{1}{n}\Sigma^{-\frac{1}{2}\top}\tilde{X}^\top\tilde{X}\Sigma^{-\frac{1}{2}} - I_d\|. \end{aligned}$$

Thus,

$$\begin{aligned} & \mathbb{P}\left(\|\frac{1}{n}\tilde{X}^\top\tilde{X} - \Sigma\| \le C\|\Sigma\|\left(\sqrt{\frac{d}{n}} \bigvee \frac{d}{n} \bigvee \sqrt{\frac{t}{n}} \bigvee \frac{t}{n}\right)\right) \\ \ge\ & \mathbb{P}\left(\|\Sigma\|\|\frac{1}{n}\Sigma^{-\frac{1}{2}\top}\tilde{X}^\top\tilde{X}\Sigma^{-\frac{1}{2}} - I_d\| \le C\|\Sigma\|\left(\sqrt{\frac{d}{n}} \bigvee \frac{d}{n} \bigvee \sqrt{\frac{t}{n}} \bigvee \frac{t}{n}\right)\right) \\ =\ & \mathbb{P}\left(\|\frac{1}{n}\Sigma^{-\frac{1}{2}\top}\tilde{X}^\top\tilde{X}\Sigma^{-\frac{1}{2}} - I_d\| \le C\left(\sqrt{\frac{d}{n}} \bigvee \frac{d}{n} \bigvee \sqrt{\frac{t}{n}} \bigvee \frac{t}{n}\right)\right). \end{aligned}$$

The transpose of each rows of $\tilde{X}\Sigma^{-\frac{1}{2}}$, which is represented as $\boldsymbol{e}_{j,d}^\top\tilde{X}\Sigma^{-\frac{1}{2}}$, are isotropic[1] This is shown by the following. Note that the j-th row of $\tilde{X}\Sigma^{-\frac{1}{2}}$ can be written by $\boldsymbol{e}_{j,d}^\top\tilde{\boldsymbol{X}}\Sigma^{-\frac{1}{2}} = \boldsymbol{X}_j{}^\top\Sigma^{-\frac{1}{2}}$(For the definition of $e_{j,d}$, see the notation chapter.).

$$\begin{aligned} \mathbb{E}\left(\left(\boldsymbol{X}_j{}^\top\Sigma^{-\frac{1}{2}}\right)^\top \boldsymbol{X}_j{}^\top\Sigma^{-\frac{1}{2}}\right) &= \mathbb{E}\left(\Sigma^{-\frac{1}{2}\top}\boldsymbol{X}_j\boldsymbol{X}_j{}^\top\Sigma^{-\frac{1}{2}}\right) \\ &= \Sigma^{-\frac{1}{2}\top}\mathbb{E}\left(\boldsymbol{X}_j\boldsymbol{X}_j{}^\top\right)\Sigma^{-\frac{1}{2}} \\ &= \Sigma^{-\frac{1}{2}\top}\Sigma\Sigma^{-\frac{1}{2}} \\ &= I_d \end{aligned}$$

So, for the proof of , it's suffice to show it in case of $(X_i)_{i\in[n]}$'s are isotropic, that is our problem boils down to show the following.

Isotropic case of (3.35)

Let $A_1, \ldots, A_n \in \mathbb{R}^d$ to be i.i.d. isotropic sub-Gaussian with $\mathbb{E}[A_i] = 0$ *for all i. Define* $A \equiv (A_1, \ldots, A_n)^\top \in \mathbb{R}^{n\times d}$. *Then,*

$$\exists C > 0 \ \forall t > 1 \qquad \mathbb{P}\left(\left\|\frac{1}{n}A^\top A - I_d\right\| \le C\left(\sqrt{\frac{d}{n}} \bigvee \frac{d}{n} \bigvee \sqrt{\frac{t}{n}} \bigvee \frac{t}{n}\right)\right) \ge 1 - e^{-t}. \tag{3.37}$$

[1] A random vector $\boldsymbol{X} \in \mathbb{R}^d$ is called *isotropic*, when $\mathbb{E}[XX^T] = I_d$.

Note that since $\left(\sqrt{\frac{d}{n}} \bigvee \sqrt{\frac{t}{n}}\right) \leq \sqrt{\frac{d}{n}} + \sqrt{\frac{t}{n}} \leq 2\left(\sqrt{\frac{d}{n}} \bigvee \sqrt{\frac{t}{n}}\right)$, (3.37) is equivalent to

$$\exists C > 0 \; \forall t > 1 \quad \mathbb{P}\left(\left\|\frac{1}{n}A^\top A - I_d\right\| \leq C\left(\delta \bigvee \delta^2\right)\right) \geq 1 - e^{-t} \quad \textit{where } \delta \equiv \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{t}{n}}\right). \tag{3.38}$$

Moreover, (3.38) is equivalent to

$$\exists C > 0 \; \forall t > 1 \quad \mathbb{P}\left(\left\|\frac{1}{n}A^\top A - I_d\right\| \leq \left(\delta \bigvee \delta^2\right)\right) \geq 1 - e^{-t} \quad \textit{where } \delta \equiv C\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{t}{n}}\right). \tag{3.39}$$

This is because replacing C to $(C \bigvee 1)^2$, we obtain (3.38) $\Rightarrow$ (3.39), and replacing C to $\left(C \bigvee C^2\right)$, we obtain (3.39) $\Rightarrow$ (3.38)

. **Step 2: approximation by ϵ-net**
Let $\mathcal{N}$ to be a $\frac{1}{4}$-net of the unit sphere S^{d-1}. We can choose the net $\mathcal{N} \subset S^{d-1}$ s.t.

$$|\mathcal{N}| \leq 9^d \qquad \text{and} \qquad \left\|\frac{1}{n}A^\top A - I_d\right\| \leq 2\max_{x \in \mathcal{N}}\left|\left\langle\left(\frac{1}{n}A^\top A - I_d\right)x, x\right\rangle\right| = 2\max_{x \in \mathcal{N}}\left|\frac{1}{n}|Ax|^2 - 1\right|.$$

So, to show (3.39), it suffices to show

$$\exists C > 0 \quad \forall t > 1 \quad \mathbb{P}\left(\max_{x \in \mathcal{N}}\left|\frac{1}{n}|Ax|^2 - 1\right| \leq \frac{\left(\delta \bigvee \delta^2\right)}{2}\right) \geq 1 - e^{-t} \tag{3.40}$$

,where $\delta \equiv C\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{t}{n}}\right)$. This is equivalent to

$$\exists C > 0 \quad \forall t > 1 \quad \mathbb{P}\left(\max_{x \in \mathcal{N}}\left|\frac{1}{n}|Ax|^2 - 1\right| \geq \frac{\left(\delta \bigvee \delta^2\right)}{2}\right) \leq e^{-t}$$

By taking the union bound,

$$\mathbb{P}\left(\max_{x \in \mathcal{N}}\left|\frac{1}{n}|Ax|^2 - 1\right| \geq \frac{\left(\delta \bigvee \delta^2\right)}{2}\right) \leq 9^d \max_{x \in \mathcal{N}} \mathbb{P}\left(\left|\frac{1}{n}\|Ax\|_2^2 - 1\right| \geq \frac{\left(\delta \bigvee \delta^2\right)}{2}\right)$$

for any given x. Therefore, to show (3.40), it suffices to show

$$\forall x \in S^{d-1} \quad \exists C > 0 \quad \forall t > 1 \quad \mathbb{P}\left(\left|\frac{1}{n}|Ax|^2 - 1\right| \geq \frac{\left(\delta \bigvee \delta^2\right)}{2}\right) \leq e^{-t}9^{-d} \quad \textit{, where} \qquad \delta \equiv C\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{t}{n}}\right) \tag{3.41}$$

Step 3 : Concentration
Take $Z_i \equiv \langle A_i, x\rangle$ for any $x \in S^{n-1}$, then, $(Z_i)_{i=1}^n$ are independent, $(Z_i)_{i=1}^n$ are also sub-Gaussian from the algebraic property of sub-Gaussian distribution, and

$$\mathbb{E}\left[Z_i^2\right] \;=\; \mathbb{E}\left[(A_i x)^\top A_i x\right] \;=\; \mathbb{E}\left[x^\top A_i^\top A_i x\right] \underset{A_i \textit{ is isotropic}}{=} \mathbb{E}\left[x^\top I_d x\right] \;= 1$$

$(Z_i^2 - 1)_{i=1}^n$ are also independent, and

$$\begin{aligned}
\left\|Z_i^2 - 1\right\|_{\psi_1} &\leq 2\left\|Z_i^2\right\|_{\psi_1} \\
&\leq 4\left\|Z_i\right\|_{\psi_2}^2 \qquad \text{Lemma 3.1.13} \\
&\leq 4K^2 \qquad \text{from the defnition of } \|\cdot\|_{\psi_2}
\end{aligned}$$

The first inequality is derived by

$$\left\|Z_i^2-1\right\|_{\psi_1}=\left\|Z_i^2-\mathbb{E}(Z_i^2)\right\|_{\psi_1}\underset{\text{Triangle inequality}}{\leq}\left\|Z_i^2\right\|_{\psi_1}+\underset{\leq \left\|(Z_i^2)\right\|_{\psi_1}}{\underset{=\mathbb{E}(Z_i^2)}{\left\|\mathbb{E}(Z_i^2)\right\|_{\psi_1}}}\leq 2\left\|Z_i^2\right\|_{\psi_1}$$

So, $(Z_i^2-1)_{i=1}^n$ are centered sub-exponential distributions. Thus, we can apply Theorem 3.1.19 to $(Z_i^2-1)_{i=1}^n$, and we obtain

$$\begin{aligned}\exists c_1>0\quad\forall\epsilon>0\qquad\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i^2-1\right|\geq\frac{\epsilon}{2}\right)&\leq 2\exp\left[-c_1\left(\frac{\epsilon^2}{16K^4}\bigwedge\frac{\epsilon}{4K^2}\right)n\right]\\&\leq 2\exp\left[-\frac{c_1}{16K^4}\left(\epsilon^2\bigwedge\epsilon\right)n\right]\end{aligned}$$

The second inequality holds because $K\geq\|Z_i\|_{\psi_2}\geq\frac{1}{\sqrt{2}}(\mathbb{E}(|Z_i|^2))^{1/2}=\frac{1}{\sqrt{2}}$, and under this region $16K^4\geq 4K^2$. Reclacing $\frac{c_1}{16}$ to c_1, we obtain

$$\exists c_1>0\quad\forall\epsilon>0\qquad\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i^2-1\right|\geq\frac{\epsilon}{2}\right)\leq 2\exp\left[-\frac{c_1}{16K^4}\left(\epsilon^2\bigwedge\epsilon\right)n\right]\tag{3.42}$$

$$\begin{aligned}\text{L.H.S. of (3.41)}&=\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i^2-1\right|\geq\frac{\left(\delta\bigvee\delta^2\right)}{2}\right)\\&\underset{(3.42)}{\leq}2\exp\left[-\frac{c_1}{K^4}\underbrace{\left(\left(\delta\bigvee\delta^2\right)^2\bigwedge\left(\delta\bigvee\delta^2\right)\right)}_{=\delta^2}n\right]\\&=2\exp\left[-\frac{c_1}{K^4}C^2\left(\sqrt{\frac{d}{n}}+\sqrt{\frac{t}{n}}\right)^2n\right]\\&\leq 2\exp\left[-\frac{c_1}{K^4}C^2\left(d+t\right)\right]\\&=\exp\left[-\frac{c_1}{K^4}C^2\left(d+t\right)+\log 2\right]\end{aligned}$$

(3.41) can be shown by taking $C\equiv K^2\sqrt{\frac{\log 9}{c_1}}$,

$$2\exp\left[-\frac{c_1}{K^4}C^2\left(d+t\right)\right]\leq 2\exp(-\log 9(d+t))=2\cdot 9^{-t}\cdot 9^{-d}<e^{-t}9^{-d}$$

□

This bound depends on demension. The following is a demension free version.

Theorem 3.2.6 *Under the same assumption of Theorem 3.2.5,*

$$\exists C>0\ \forall t>1\quad P\left(\|\hat{\Sigma}-\Sigma\|\leq\|\Sigma\|\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\bigvee\frac{\mathbf{r}(\Sigma)}{n}\bigvee\sqrt{\frac{t}{n}}\bigvee\frac{t}{n}\right)\right)\geq 1-e^{-t}$$

,where $\mathbf{r}(\Sigma)\equiv\frac{(\mathbb{E}\|X\|)^2}{\|\Sigma\|}$

Example
In case when $\boldsymbol{X} \sim \mathcal{N}_3\,(0, I_3)$, $\|\Sigma\|$ = the maximum absolute eigen value = 1.

$$\begin{aligned}
\mathbb{E}\,\|\boldsymbol{X}\| &= \int_{\mathbb{R}^3} \sqrt{x^2+y^2+z^2}\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2+y^2+z^2}{2})dxdydz \\
&= \int_{r=0}^{\infty}\int_{\theta=0}^{2\pi}\int_{\phi=-\frac{1}{2}\pi}^{\frac{1}{2}\pi} r^2\frac{1}{\sqrt{2\pi}}\exp(-\frac{r^2}{2})\cos\phi d\phi d\theta dr \\
&= \sqrt{8\pi}\underbrace{\int_{r=0}^{\infty} r^2\exp(-\frac{r^2}{2})dr}_{=\sqrt{\frac{\pi}{2}}} \\
&= 2\pi
\end{aligned}$$

So, $\boldsymbol{r}(\Sigma) = 4\pi^2$

4 Robust Estimation on the framework of graphical models

Classical estimation methods rely heavily on assumptions about randomness, independenceness, distributional models, perhaps prior distributions for some unknown parameters, and so on. In fact, until the previous chapter, we have assumed $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ perfectly distribute from the distribution which interest us. Unfortunately, these assumption does not hold in practice, sometimes practical data contain outliers. When there are outliers in the data, classical estimators often have very poor performance.
In order to deal with such practical problems, alterative estimators which is less susceptible to the outliers have been developed (See [14]) . Distinguished with classical estimators, the estimators are called *robust* estimators.

4.1 Robust estimator for mean of random vector

At first, we consider estimators for the mean. In the following table, we describe methods to estimate mean of random vector, provide theoretical assumptions, computation time, and give the error bounds. As a basic setting, assume that

- we have i.i.d sample $(\boldsymbol{X}_i)_{i\in[n]}$ from the distribution P_{θ_0},
- $\boldsymbol{X}$ has finite covariance matrix.

The following shows several famous mean estimator.

$$\text{Mean estimator } \hat{\boldsymbol{\mu}} = \left(\hat{\mu}^{(1)}, \ldots, \hat{\mu}^{(d)}\right)^\top$$

	Method	Assumption	Computation	Error Bound
1	Simple mean	Sub Gaussian or bounded	$O\left(dn\right)$	*See Chapter 2*
2	Median of means	None	$O\left(dn \log k\right)$	$\lvert\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\rvert \leq \sqrt{32\mathrm{Tr}(\Sigma)\frac{\log(d/\delta)}{n}}$ (Proposition 1 in [5])
3	Trimmed mean	$n > (16/3)\log(8/\delta)$	$O\left(dn \log k\right)$	$\lvert\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\rvert \leq 9\sqrt{\mathrm{Tr}(\Sigma)\frac{\log(8/\delta)}{n}}$ (Theorem 6 in [5] and take union bound)

Remark 1 The finiteness of the variance of $\boldsymbol{X}$ means the finiteness of Median because

$$\begin{aligned} |\mathrm{Med}(\boldsymbol{X})| &\leq |\mathrm{Med}(\boldsymbol{X}) - \boldsymbol{\mu}| + |\boldsymbol{\mu}| \\ &\leq \mathbb{E}(|\mathrm{Med}(\boldsymbol{X}) - \boldsymbol{X}|) + |\boldsymbol{\mu}| \\ &\leq \mathbb{E}(|\boldsymbol{\mu} - \boldsymbol{X}|) + |\boldsymbol{\mu}| \\ &\leq \sqrt{\mathbb{E}((\boldsymbol{\mu} - \boldsymbol{X})^2)} + |\boldsymbol{\mu}| \quad = \quad \sqrt{\mathrm{Tr}(\Sigma\Sigma^\top)} + |\boldsymbol{\mu}| \,. \end{aligned}$$

The second and fourth inequalities holds from Jensen's inequality, and the third can be shown by the fact that median minimizes the function on $\mathbb{R}^d : a \mapsto \mathbb{E}(|\boldsymbol{X} - a|)$.

1. Sample mean In the previous chapter, we discussed sample mean.

$$\hat{\mu} \equiv \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i$$

2. Median of means Median of means[1] are defined depending on the number of blocks ($\equiv k$). The estimator is calculated by the following two steps. First, we partition $[n] = \{1, \dots, n\}$ into k blocks $B_1, \dots, B_k$, where we assume $\#B_i = m$ for all $i \in [k]$, so km equals to n. Next, we calculate the sample means in each blocks.

$$\boldsymbol{Z}_l \equiv \frac{1}{m} \sum_{i \in B_l} \boldsymbol{X}_i \quad \forall l \in [k]$$

Then, take the median of the k data.

$$\hat{\mu}^{(j)} \equiv \operatorname*{Med}_{l \in [k]} {Z_l}^{(j)} \quad \forall j \in [d]$$

3. Trimmed mean Trimmed mean[2] use $2n$ samples from $\boldsymbol{X}$. Let $X_1, \dots \boldsymbol{X}_n, \boldsymbol{X}_1, \dots \boldsymbol{X}_n$ be i.i.d. copies of $\boldsymbol{X}$. And it has a parameter $\epsilon \in [0, 0.5]$, , which we call *discarded rate*. The estimation procedure is the following. First, we use $(\boldsymbol{X}_i)_{i \in [n]}$ to set thresholds to set the criteria of outliers : For the discarded rate ϵ and for all $j \in [d]$, we calculate

$$\alpha^{(j)} \equiv \text{the } \epsilon\text{−quantile of } (\boldsymbol{X}_i^{(j)})_{i \in [b]} \qquad \beta^{(j)} \equiv \text{the } (1-\epsilon)\text{−quantile of } (\boldsymbol{X}_i^{(j)})_{i \in [b]}$$

Next, we calculate the mean of adjusted $(X_i^{(j)})_{i \in [n]}$:

$$\hat{\mu}^{(j)} \equiv \frac{1}{n} \sum_{i=1}^{n} \phi_{\alpha^{(j)}, \beta^{(j)}} \left(X_i^{(j)} \right),$$

where $\phi_{\alpha,\beta} : \mathbb{R} \to \mathbb{R}$ is defined by $\phi_{\alpha,\beta}(x) = \begin{cases} \beta & \text{if } x > \beta, \\ x & \text{if } x \in [\alpha, \beta], \\ \alpha & \text{if } x < \alpha, \end{cases}$

Remark Under the condition that $\boldsymbol{X}$ has finite covariance matrix, the expectation of Trimmed mean is also finite. We need a long discussion to prove in general case. But, under the assumption that $\alpha^{(j)}$'s and $\beta^{(j)}$'s are finite. It's trivial that the trimmed mean is finite.

4.2 Robust estimator for covariance matrix of random vector

Next, we consider estimating the covariance $\Sigma = (\sigma^{(i,j)})_{i,j}$. As is the case with mean, we suppose that

- we have i.i.d sample $(\boldsymbol{X}_i)_{i \in [n]}$ from the distribution P_{θ_0}
- $\boldsymbol{X}$ has finite variance.

$$\text{Covariance estimator } \hat{\Sigma} = \left(\hat{\sigma}^{(i,j)}\right)_{\{i,j\} \in [n] \times [n]}$$

[1] As the reference, e.g. see [5]
[2] As the reference, e.g. see [5]

The following shows several well-known covariance estimator.

	Method	Assumption	Computation	Error Bound
1	Sample covariance	None	$O\left(dn\right)$	$\|\hat{\Sigma}-\Sigma\| \leq C\|\Sigma\|\left(\sqrt{\frac{d}{n}} \vee \frac{d}{n} \vee \sqrt{\frac{-\log\delta}{n}} \vee \frac{-\log\delta}{n}\right)$
2	Covariance derived by Median absolute deviation	Gaussian	$O\left(dn\log n\right)$	The existency of the error boundis not trivial[3]
3	Covariance derived from order statistic of two samples defferences	Gaussian	$O\left(dn\log n\right)$	The existency of the error boundis not trivial[4]
4	Minsker's method	None	Infeasible	See the below

The correlation $\rho\left(\mathbf{X}^{(j_1)}, \mathbf{X}^{(j_2)}\right) = \rho_{j_1,j_2}$ can be written in terms of standard deviations of one dimensional variables $\sigma_X = \sqrt{\mathrm{Var}(X)}$ as the followings.

$$\rho_{i,j} = \rho_{\boldsymbol{X}^{(j_1)},\mathbf{X}^{(j_2)}} = \frac{\sigma^2_{a\mathbf{X}^{(j_1)}+b\mathbf{X}^{(j_2)}} - \sigma^2_{a\mathbf{X}^{(j_1)}-b\mathbf{X}^{(j_2)}}}{\sigma^2_{a\mathbf{X}^{(j_1)}+b\mathbf{X}^{(j_2)}} + \sigma^2_{a\mathbf{X}^{(j_1)}-b\mathbf{X}^{(j_2)}}}, \tag{4.1}$$

where $a \equiv \frac{1}{\sigma_{X^{(j_1)}}}$ and $b \equiv \frac{1}{\sigma_{X^{(j_2)}}}$.
And the relation between the covariance and the correlation is given by

$$\sigma^{(j_1,j_2)} = \rho_{j_1,j_2}\sigma_{X^{(j_1)}}\sigma_{X^{(j_2\bar{(}i))}}. \tag{4.2}$$

We use above relations to construct a (robust) estimator for the covariance :

First, we construct estimators of the variances (we will see several estimator for variances later) for all $j_1,\ j_2 \in [d]$,

$$\hat{\sigma}^2_{\mathbf{X}^{(j_1)}},\ \hat{\sigma}^2_{\mathbf{X}^{(j_2)}},\ \hat{\sigma}^2_{a\mathbf{X}^{(j_1)}+b\mathbf{X}^{(j_2)}},\ \hat{\sigma}^2_{a\mathbf{X}^{(j_1)}-b\mathbf{X}^{(j_2)}},$$

,where $a = \frac{1}{\hat{\sigma}_{X^{(j_1)}}}$ and $b = \frac{1}{\hat{\sigma}_{\mathbf{X}^{(j_2)}}}$.

Second, define $\hat{\sigma}^{(j_1,j_2)}$ by

$$\hat{\rho}_{j_1,j_2} \equiv \frac{\hat{\sigma}^2_{a\mathbf{X}^{(j_1)}+b\mathbf{X}^{(j_2)}} - \hat{\sigma}^2_{a\mathbf{X}^{(j_1)}-b\mathbf{X}^{(j_2)}}}{\hat{\sigma}^2_{a\mathbf{X}^{(j_1)}+b\mathbf{X}^{(j_2)}} + \hat{\sigma}^2_{a\mathbf{X}^{(j_1)}-b\mathbf{X}^{(j_2)}}}$$

$$\hat{\sigma}^{(j_1,j_2)} \equiv \hat{\rho}_{j_1,j_2}\hat{\sigma}_{X^{(j_1)}}\hat{\sigma}_{X^{(j_2)}} = \frac{\hat{\sigma}^2_{a\mathbf{X}^{(j_1)}+b\mathbf{X}^{(j_2)}} - \hat{\sigma}^2_{a\mathbf{X}^{(j_1)}-b\mathbf{X}^{(j_2)}}}{\hat{\sigma}^2_{a\mathbf{X}^{(j_1)}+b\mathbf{X}^{(j_2)}} + \hat{\sigma}^2_{a\mathbf{X}^{(j_1)}-b\mathbf{X}^{(j_2)}}}\hat{\sigma}_{X^{(j_1)}}\hat{\sigma}_{X^{(j_2)}}$$

Therefore, the transformation from variances to correlation by (4.1) makes our problem boil down to the problem to find the robust estimator of the variance of the each component. The following table shows several robust estimator for the variance.

1. Sample covariance As is well known, sample covariance $\hat{\Sigma}$ is defined by

$$\hat{\Sigma} \equiv \left(\hat{\sigma}^{(j_1,j_2)}\right)_{j_1,j_2\in[d]} \equiv \left(\frac{1}{n}\sum_{i=1}^{n}(X_i^{(j_1)} - \overline{X}^{(j_1)})(X_i^{(j_2)} - \overline{X}^{(j_2)})\right)_{j_1,j_2\in[d]}$$

This definition of sample covariance van be induced by the definition of sample variance $\hat{\sigma}_X \equiv \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$ and $(*)$. We can check this as the following.

$$\begin{aligned}
\hat{\rho}_{j_1,j_2} &= \frac{4ab\frac{1}{n}\sum_{i=1}^{n}(X_i^{(j_1)} - \overline{X}^{(j_1)})(X_i^{(j_2)} - \overline{X}^{(j_2)})}{2a^2\frac{1}{n}\sum_{i=1}^{n}(X_k^{(j_1)} - \overline{X}^{(j_1)})^2 + 2b^2\frac{1}{n}\sum_{i=1}^{n}(X_i^{(j_2)} - \overline{X}^{(j_2)})^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i^{(j_1)} - \overline{X}^{(j_1)})(X_i^{(j_2)} - \overline{X}^{(j_2)})}{\sqrt{\frac{1}{n}\sum_{k=1}^{n}(X_i^{(j_1)} - \overline{X}^{(j_1)})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i^{(j_2)} - \overline{X}^{(j_2)})^2}} \\
\hat{\sigma}^{(j_1,j_2)} &= \hat{\rho}_{j_1,j_2}\hat{\sigma}_{X^{(j_1)}}\hat{\sigma}_{X^{(j_2)}} \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i^{(j_1)} - \overline{X}^{(j_1)})(X_i^{(j_2)} - \overline{X}^{(j_2)})
\end{aligned}$$

2. Median absolute deviation When we know $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, one of robust covariance estimator is Median absolute deviation [5] (MAD), which is defined as

$$\hat{\sigma}^{(i)} \equiv c\mathrm{Med}_k\left\{\mathrm{Med}_l\left|\boldsymbol{X}_k^{(i)} - \boldsymbol{X}_l^{(i)}\right|\right\}.$$

in the case of Gaussian, $c \approx 1.1926$.

3. Alternative to MAD [6] Successively under the assumption of $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, another robust covariance estimator is defined as

$$\hat{\sigma}^{(i)} \equiv d\left\{\left|X_k^{(i)} - X_l^{(i)}\right|; k < l\right\}_{(z)}.$$

here $\{\boldsymbol{X}_k\}_{(z)}$ means the z-th largest value in $\{\boldsymbol{X}_k\}$ and $z \equiv \begin{pmatrix} \lfloor n/2 \rfloor + 1 \\ 2 \end{pmatrix}$. In case of Gaussian, $d \approx 2.2219$.

4. Minsker's method We consider the method described by Minsker in [16]. Different from the estimators in 2. Covariance derived by Median absolute deviation and 3.Covariance derived from order statistic of two samples differences, the Minsker's method directly estimates the covariance matrix. The definition is the following.

Definition 4.2.1 *Let $(\boldsymbol{X}_i)_{i\in[n]}$ are i.i.d. copies of a random vector $\boldsymbol{X} \in \mathbb{R}^d$. Minsker's estimator with parameter $\lambda = (\lambda_1, \lambda_2)$ is defined as*

$$\widehat{S}_\lambda \equiv \underset{S\in S_d}{\mathrm{argmin}} \min_{U_{i,j}\in S_d}\left[\frac{1}{n(n-1)}\sum_{i\neq j}\left\|\widetilde{\boldsymbol{X}}_{i,j}\widetilde{\boldsymbol{X}}_{i,j}^{\top} - S - \sqrt{n(n-1)}U_{i,j}\right\|_{\mathrm{F}}^2 + \lambda_1\|S\|_1 + \lambda_2\sum_{i\neq j}\left\|U_{i,j}\right\|_1\right], \tag{4.3}$$

here $\widetilde{\boldsymbol{X}}_{i,j} \equiv \frac{\boldsymbol{X}_i - \boldsymbol{X}_j}{\sqrt{2}}$, S_d is the whole set of real-valued symmetric $d \times d$ matrices.

The following theorem provides us the minimum value with respect to $U_{i,j}$.

[5] As the reference, see[6]

[6] As the reference, see [6]

Theorem 4.2.2 (See Remark 1 in [16]) *Let $(X_i)_{i\in[n]}$ are i.i.d. copies of a random vector $X \in \mathbb{R}^d$. We can express Minsker's estimator with parameter $\lambda = (\lambda_1, \lambda_2)$ as*

$$\widehat{S}_\lambda = \underset{S}{\operatorname{argmin}} \left\{ \frac{2}{n(n-1)} \operatorname{tr} \left[\sum_{i\neq j} \rho_{\frac{\sqrt{n(n-1)}\lambda_2}{2}} \left(\widetilde{X}_{i,j} \tilde{X}_{i,j}^\top - S \right) \right] + \lambda_1 \|S\|_1 \right\}, \tag{4.4}$$

where

$$\rho_\lambda(u) := \begin{cases} \frac{u^2}{2}, & |u| \leq \lambda \\ \lambda|u| - \frac{\lambda^2}{2}, & |u| > \lambda \end{cases} \qquad \forall u \in \mathbb{R}, \lambda \in \mathbb{R}^+$$

and for $f : \mathbb{R} \to \mathbb{R}$ and $A \in S_d$ with the spectral decomposition $A = \boldsymbol{U}\mathbf{diag}(\tau_1, \ldots, \tau_d)\boldsymbol{U}^\top$, define $f(A)$ as

$$f(A) \equiv \boldsymbol{U} \begin{pmatrix} f(\tau_1) & & \\ & \ddots & \\ & & f(\tau_d) \end{pmatrix} \boldsymbol{U}^\top. \tag{4.5}$$

Outline of proof
Consider the spectral decomposition of $\widetilde{X}_{i,j}\widetilde{X}_{i,j}^\top - S$ for all $i, j \in [n]$:

$$\widetilde{X}_{i,j}\widetilde{X}_{i,j}^\top - S = \sum_{k=1}^{d} \tau_k^{(i,j)} v_k^{(i,j)} v_k^{(i,j)\top}$$

where, $\tau_k^{(i,j)}$ is the k-th eigenvalue of $\widetilde{X}_{i,j}\widetilde{X}_{i,j}^\top - S$ and $v_k^{(i,j)}$ is the corresponding eigenvector. We can see that

$$\widetilde{U}_{i,j} \equiv \frac{1}{\sqrt{n(n-1)}} \sum_{k=1}^{d} \operatorname{sign}\left(\tau_k^{(i,j)}\right) \left(\left|\tau_k^{(i,j)}\right| - \frac{\sqrt{n(n-1)}\lambda_2}{2} \right)_+ v_k^{(i,j)} v_k^{(i,j)T} = \rho_{\frac{\sqrt{n(n-1)}\lambda_2}{2}} \left(\widetilde{Y}_{i,j} \widetilde{Y}_{i,j}^T - S \right)$$

minimizes the object function of (4.3). Plugging $\widetilde{U}_{i,j}$ in the object function, we get the statement. For details, see the D.1 in [16].

Practical calculation by Minsker's method In general, it is difficult to analytically find the minimizer S, so we solve this optimal problem by a numerical approach : Proximal Gradient Descent (PGD) method. Let' consider a general optimizing problem :

$$\operatorname{argmin}_x f(x) = \operatorname{argmin}_x g(x) + h(x), \tag{4.6}$$

where g is convex and differentiable, h is convex (not necessarily differentiable). The PDG method is a numerical way for solving the problem, which is proceeded by the following inductions.

- starts from an initial point $x^{(0)}$,
- updates as : $x^{(k)} = \operatorname{prox}_{\alpha_k h}\left(x^{(k-1)} - \alpha_k \nabla g\left(x^{(k-1)}\right)\right)$,

where

$$\operatorname{prox}_h(x) \equiv \underset{u}{\operatorname{argmin}} \left(h(u) + \frac{1}{2}\|u - x\|_2^2 \right),$$

and $\alpha_k > 0$ is the step size. The following lemma guarantees the convergence of $x^{(k)}$ to the minimizer x^*.

Lemma 4.2.3 *Assume that g and h satisfies the above conditions, and suppose that ∇g is Lipschitz continuous with constant $L > 0$, that is*

$$\|\nabla g(x) - \nabla g(y)\| \leq L\|x - y\|$$

and the optimal value f^* of (4.6) is finite and achieved at the point x^*. Take $\alpha_k = \alpha \leq L$. Then the PGD algorithm yield an $O(\frac{1}{k})$ convergence rate, i.e.

$$\exists C \in \mathbb{R} \quad \forall k \in \mathbb{N} \quad \left| f\left(x^{(k)}\right) - f^* \right| \leq \frac{C}{k}.$$

We can apply PGD method to our problem (4.4) by taking

$$\begin{aligned} g(S) &\equiv \operatorname{tr}\left[\sum_{i \neq j} \rho_{\frac{\sqrt{n(n-1)}\tau_2}{2}}\left(\widetilde{\boldsymbol{X}}_{i,j}\widetilde{\boldsymbol{X}}_{i,j}^\top - S\right)\right] \\ h(S) &\equiv \tau_1 \|S\|_1 \end{aligned}$$

Note that, g is convex and differentiable with respect to S. Recall that differentiability of multi dimensional function $f : \mathbb{R}^m \to \mathbb{R}^m$ is defined as the following.

Definition 4.2.4 (Gateaux differentiability) *$f : \mathbb{R}^m \to \mathbb{R}^m$ is called differentiable at $A \in \mathbb{R}^m$ when there exists a linear transformation $\mathcal{D}f(A) : \mathbb{R}^m \to \mathbb{R}^m$ such that*

$$\forall H \in \mathbb{R}^m \qquad \mathcal{D}f(A) = \left.\frac{d}{dt}\right|_{t=0} f(A + tH).$$

$\mathcal{D}f(A)$ is called Gateaux derivative.

Now, In order to check this, we use the following lemma.

Lemma 4.2.5 (See Theorem V.3.3 in [17]) *Let $f \in C^1(\mathbb{R})$ and let A be a symmetric matrix with the spectral decomposition $A = \boldsymbol{U}\mathbf{diag}(\lambda_1, \ldots, \lambda_d)\boldsymbol{U}^\top$. Then the map $f : S_d \longrightarrow S_d$ induced by (4.5) is differentiable at A, and the Gateux derivative of f at A satisfies*

$$\forall H \quad \mathcal{D}f(A)(H) = U\left[f'(\mathbf{diag}(\lambda_1, \ldots, \lambda_d)) \circ U^\top H U\right] U^\top,$$

where $\circ$ is entry-wise products.

Now, from the definition of ρ and g, g is also differentiable and takes over the convexity of ρ. And

$$\begin{aligned} \|\nabla g(S_1) - \nabla g(S_2)\|_F &= \|\mathcal{D}g(S_1) - \mathcal{D}g(S_2)\|_F \\ &\leq C\|\mathcal{D}g(S_1) - \mathcal{D}g(S_2)\| \quad \text{for an absolute constant } C \in \mathbb{R}_+ \\ &\leq C\|S_1 - S_2\|. \end{aligned}$$

So, applying 4.2.3, the convergence of the following algorithm is guaranteed. Thus, we get the following algorithm.

Algorithm : Proximal gradient descent (PGD) method

Input: number of iterations T, tuning parameters λ_1 and λ_2, initial estimation S^0, samples $(\boldsymbol{X}_i)_{i\in[n]} \in \mathbb{R}^d$.
For $t = 1, 2, \ldots, T$ **do**

(1) Compute $G_t = -\dfrac{2}{n(n-1)} \displaystyle\sum_{0 \leq i_t < j_t \leq n} \nabla g_{i,j}\left(S^t\right) = -\dfrac{2}{n(n-1)} \sum_{0 \leq i_t < j_t \leq n} \rho'_{\frac{\sqrt{n(n-1)}\lambda_2}{2}}\left(\widetilde{\boldsymbol{X}}_{i,j}\widetilde{\boldsymbol{X}}_{i,j}^\top - S^t\right)$.

(2) (gradient update) $T^{t+1} = S^t - G_t$.

(3) (proximal update)

$$S^{t+1} = \underset{S}{\operatorname{argmin}}\left\{\frac{1}{2}\left\|S - T^{t+1}\right\|_F^2 + \frac{\lambda_1}{2}\|S\|_1\right\} = \gamma_{\frac{\lambda_1}{2}}\left(T^{t+1}\right),$$

where $\gamma_\lambda(u) = \operatorname{sign}(u)(|u| - \lambda)_+$.

End for

Output: S^{T+1}

We are interested in the deviation of $\left\|\widehat{S}_\lambda - \Sigma\right\|$. The following statement provides as an answer.

Theorem 4.2.6 (See Theorem 4 in [16]) *Let $(\boldsymbol{X}_i)_{i\in[n]}$ are i.i.d. copies of a random vector $\boldsymbol{X} \in \mathbb{R}^d$, For any t, σ which satisfies*

$$\left.\begin{array}{rcl} 1 \le t & \le & \frac{c_3 n}{r_H} \text{ for some enough small constant} c_3, \text{where } r_H \equiv \frac{\mathrm{tr}\left(\mathbb{E}\left[\left(H_{1,2}-\Sigma\right)^2\right]\right)}{\left\|\mathbb{E}\left[\left(H_{1,2}-\Sigma\right)^2\right]\right\|} \\ \sigma & \ge & \left\|\mathbb{E}\left[\left(H_{1,2}-\Sigma\right)^2\right]\right\|^{\frac{1}{2}}, \\ n & \ge & \left\{64a^2 r_H t \bigvee \frac{4b^2t^2\|\Sigma\|^2}{\sigma^2}\right\} \text{ for some enough large constants } a, b, \\ \lambda_1 & \le & (\frac{\sigma}{4})\sqrt{\frac{n}{t}} \\ \lambda_2 & \ge & \frac{\sigma}{\sqrt{(n-1)t}} \end{array}\right\} (**)$$

Then

$$\mathbb{P}\left(\left\|\widehat{S}_\lambda - \Sigma\right\| \le \frac{20}{39}\lambda_1 + \frac{80}{39}\sigma\sqrt{\frac{t}{n}} + \frac{40}{39}\lambda_2 t\right) \ge 1 - \left(\frac{8r_H}{3}+1\right)e^{-t},$$

where $H_{i,j} \equiv \frac{(\boldsymbol{X}_i-\boldsymbol{X}_j)(\boldsymbol{X}_i-\boldsymbol{X}_j)^\top}{2}$ for $i, j \in [n]$.

Outline of proof
For a given $t \in \mathbb{N}$ in the PGD algorithm,

$$\begin{array}{rcl} \left\|S^{t+1}-\Sigma\right\| & \le & \left\|S^{t+1}-T^{t+1}\right\| + \left\|T^{t+1}-\Sigma\right\| \\ & \le & \left\|S^{t+1}-T^{t+1}\right\| + \left\|\frac{1}{n(n-1)\theta_2}\sum_{i\neq j}\left[\rho'\left(\theta_2\left(H_{i,j}-S^t\right)\right)-\rho'\left(\theta_2\left(H_{i,j}-\Sigma\right)\right)\right] + S^t - \Sigma\right\| \\ & & \qquad + \left\|\frac{1}{n(n-1)\theta_2}\sum_{i\neq j}\rho'\left(\theta_2\left(H_{i,j}-\Sigma\right)\right)\right\|, \end{array}$$

where $\theta_2 \equiv \frac{2}{\lambda_2\sqrt{n(n-1)}}$. Calculating the bound of the each three terms in the right hand side. We obtains the statement. For details, see the D.1 in [16].
Replacing $\left(\frac{8r_H}{3}+1\right)e^{-t}$ to δ, we immediately obtain the following Corollary.

Corollary 4.2.7 *Let $(\boldsymbol{X}_i)_{i\in[n]}$ are i.i.d. copies of a random vector $\boldsymbol{X} \in \mathbb{R}^d$. For a given $\delta \in [0,1]$, define $t = \log\left(\frac{8r_H+3}{3\delta}\right)$. Assume that t, σ satisfies (**) in Theorem 4.2.6. Then,*

$$\mathbb{P}\left(\left\|\widehat{S}_\lambda - \Sigma\right\| \le \frac{20}{39}\lambda_1 + \frac{80}{39}\sigma\sqrt{\frac{\log\left(\frac{8r_H+3}{3\delta}\right)}{n}} + \frac{40}{39}\lambda_2\log\left(\frac{8r_H+3}{3\delta}\right)\right) \ge 1-\delta.$$

Implementation of Minsker's method by PGD algorithm To see the behavior of estimates by Minsker method's, we set the following three cases as the population distribution.

- 1 dimensional Gaussian distribution : $(\boldsymbol{X}_i)_{i\in[n]} \sim \mathcal{N}(1,3)$

  $n = 100, S_0 = 10, T = 200, \lambda_1 = 0, \lambda_2 = 0.02, 0.1, 1,$ and 20. We simulate 100 times.

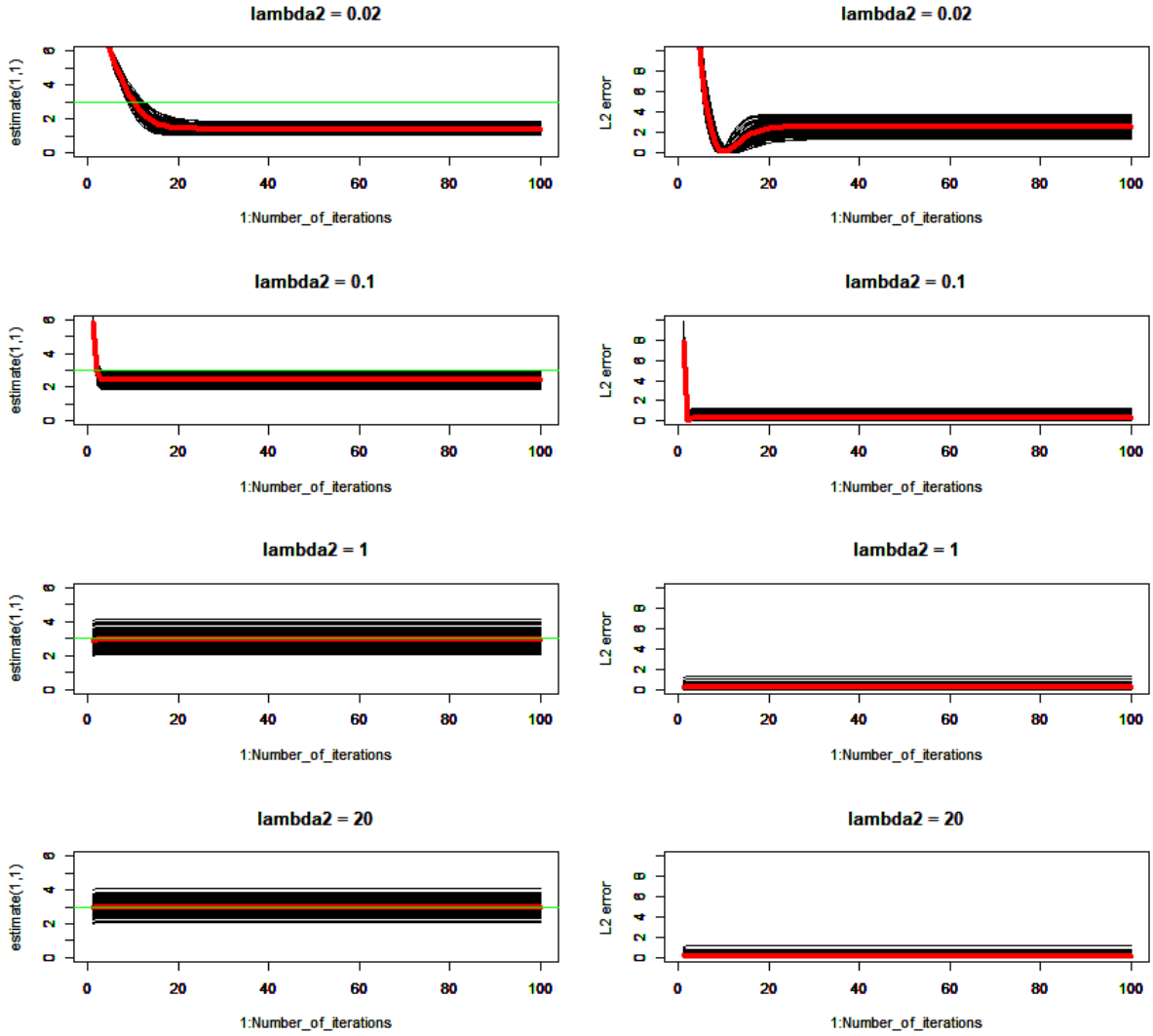
- 1 dimensional Pareto distribution : $(\boldsymbol{X}_i)_{i\in[n]} \sim Pareto(x_m = 2, \alpha = 3)$.

  $n = 100, S^0 = 10, T = 200, \lambda_1 = 0, \lambda_2 = 0.02, 0.1, 1,$ and 20

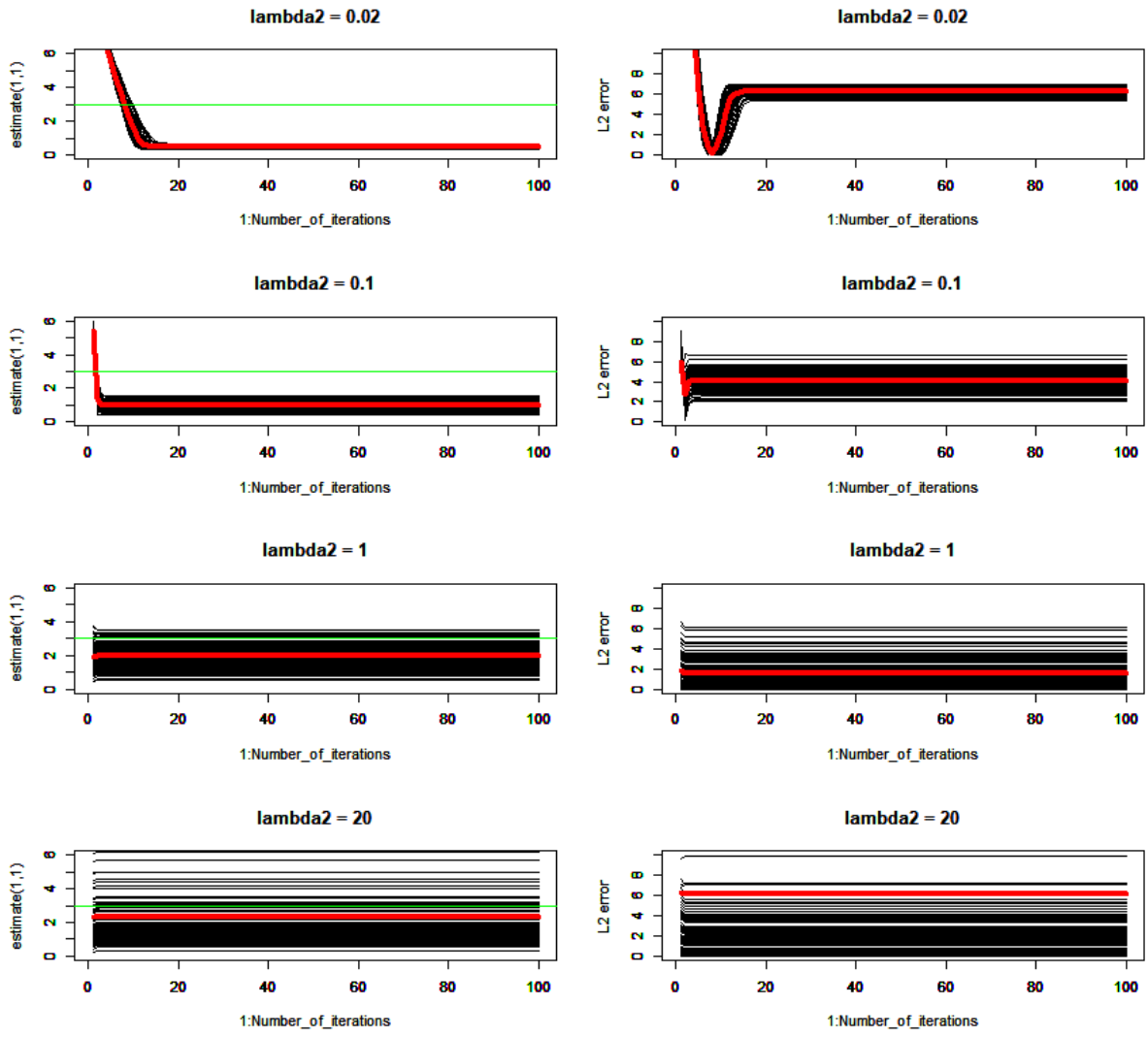
  Note that the mean $= \frac{\alpha \boldsymbol{x}_m}{\alpha-1} = \frac{3\cdot 2}{3-1} = 3$ and the variance $= \frac{x_m^2\cdot\alpha}{(\alpha-1)^2(\alpha-2)} = 3$. We simulate 100 times.

- 3 dimensional Gaussian distribution : $(X_i)_{i\in[n]} \sim \mathcal{N}(\mu, \Sigma)$ where $\mu = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1 & 0 & 0.5 \\ 0 & 2 & 1 \\ 0.5 & 1 & 1 \end{pmatrix}$. $n = 100, S^0 = 10, T = 200, \lambda_1 = 0, \lambda_2 = 0.02, 0.1, 1,$ and 20

In each case, we simulate 100 times. From the next page, the results (the pass of the operator norm and the pass of $(S^t - S_0)^2$) are shown. Red line represents the average.
1 dimensional normal distribution : $(X_i)_{i\in[n]} \sim \mathcal{N}(1, 3)$



1 dimensional pareto distribution : $(X_i)_{i\in[n]} \sim Pareto(x_m = 2, \alpha = 3)$



Implementation of Minsker'm method by SPGD algorithm Computational cost of PGD is so large. To avoid this problem, we replace the step (1) to a stochastic one.

Algorithm : Mini-batch Stochastic proximal gradient descent (SPGD) method

---

Input: Number of iterations T, step size $(\eta_t)_{t\in[T]}$, batch size $B \in \mathbb{N}$, tuning parameters λ_1 and λ_2, initial estimation S^0, samples $(\boldsymbol{X}_i)_{i\in[n]} \in \mathbb{R}^d$.

For $t = 1, 2, \ldots, T$ **do**

(1) **For** $b = 1, \ldots, B$ **do**

Randomly pick i_t and j_t s.t. $0 \le i_t < j_t \le n$.

Compute $G_t^b = -\nabla g_{i,j}\left(S^t\right) = -\rho'_{\frac{\sqrt{n(n-1)}\lambda_2}{2}}\left(\widetilde{\boldsymbol{X}}_{i,j}\widetilde{\boldsymbol{X}}_{i,j}^\top - S^t\right)$.

(2) $G_t = \frac{1}{B}\sum_{b=1}^B G_t^b$

(3) (gradient update) $T^{t+1} = S^t - G_t$.

(4) (proximal update)

$$S^{t+1} = \underset{S}{\operatorname{argmin}}\left\{\frac{1}{2}\left\|S - T^{t+1}\right\|_F^2 + \frac{\lambda_1}{2}\|S\|_1\right\} = \gamma_{\frac{\lambda_1}{2}}\left(T^{t+1}\right),$$

where $\gamma_\lambda(u) = \operatorname{sign}(u)(|u| - \lambda)_+$.

End for

Output: S^{T+1}

Same as the previous section, we set the following three cases as the population distribution.

- 1 dimensional normal distribution : $(\boldsymbol{X}_i)_{i\in[n]} \sim \mathcal{N}(1, 3)$

  $n = 100, S_0 = 10, T = 200, \lambda_1 = 0, \lambda_2 = 0.02, 0.1, 1,$ and 20. We simulate 100 times.

- 1 dimensional pareto distribution : $(\boldsymbol{X}_i)_{i\in[n]} \sim Pareto(x_m = 2, \alpha = 3)$.

  $n = 100, S^0 = 10, T = 200, \lambda_1 = 0, \lambda_2 = 0.02, 0.1, 1,$ and 20

  Note that the mean $= \frac{\alpha \boldsymbol{x_m}}{\alpha-1} = \frac{3\cdot 2}{3-1} = 3$ and the variance $= \frac{x_m^2 \cdot \alpha}{(\alpha-1)^2(\alpha-2)} = 3$. We simulate 100 times.

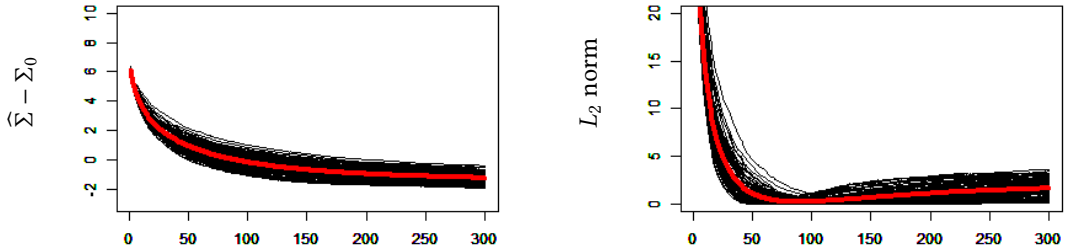
- 3 dimensional gaussian distribution : $(\boldsymbol{X}_i)_{i\in[n]} \sim \mathcal{N}(\mu, \Sigma)$ where $\mu = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1 & 0 & 0.5 \\ 0 & 2 & 1 \\ 0.5 & 1 & 1 \end{pmatrix}$.

  $n = 100, S^0 = 10, T = 200, \lambda_1 = 0, \lambda_2 = 0.02, 0.1, 1,$ and 20

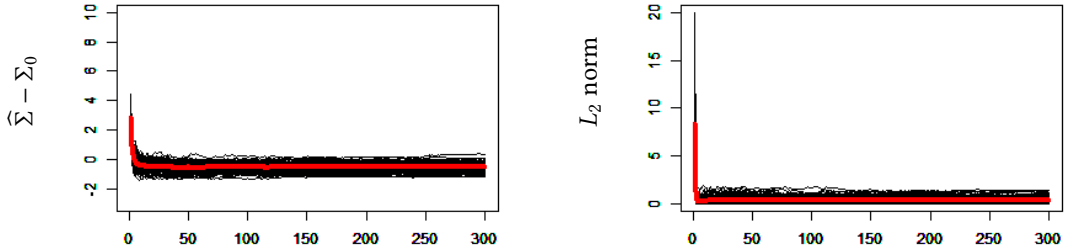
In each case, we simulate 100 times and we set the step size $B = 20$ and $\eta_t \equiv t^{-\frac{2}{3}}$. From the next page, the results (the pass of the operator norm and the pass of $(S^t - S_0)^2$) are shown. Red line represents the average.

1 dimensional normal distribution : $(X_i)_{i\in[n]} \sim \mathcal{N}(1,3)$

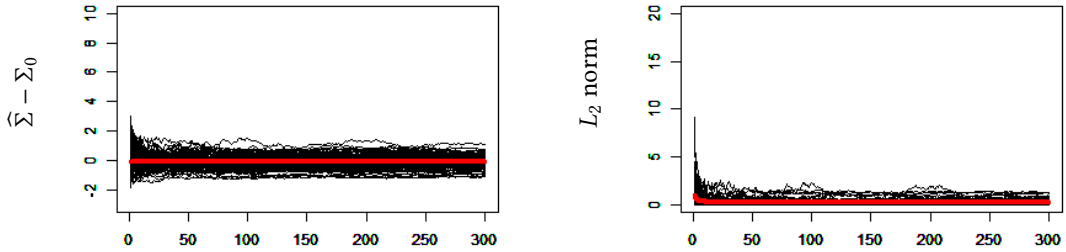
$\lambda_1 = 0 \quad \lambda_2 = 0.02$



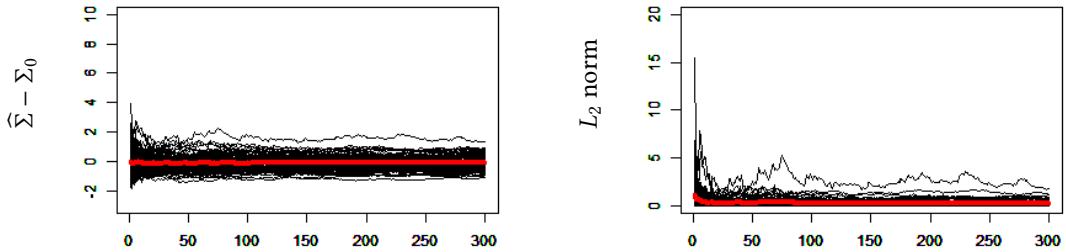
$\lambda_1 = 0 \quad \lambda_2 = 0.1$



$\lambda_1 = 0 \quad \lambda_2 = 1$

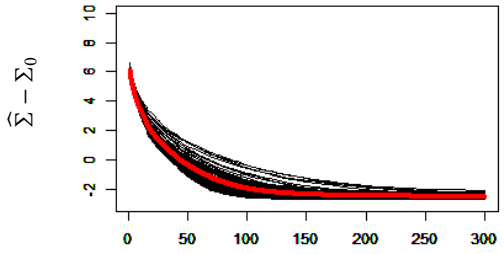


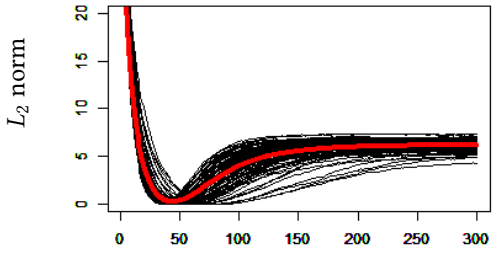
$\lambda_1 = 0 \quad \lambda_2 = 20$



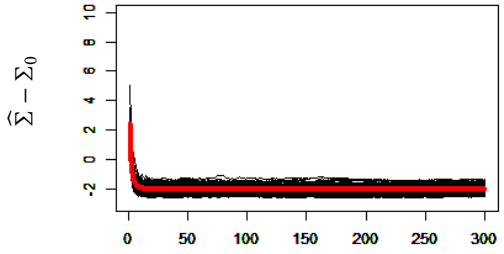
1 dimensional pareto distribution : $(X_i)_{i\in[n]} \sim Pareto(x_m = 2, \alpha = 3)$

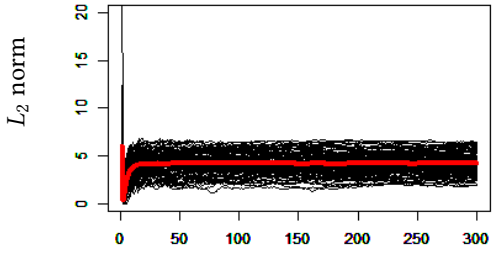
$\lambda_1 = 0 \quad \lambda_2 = 0.02$



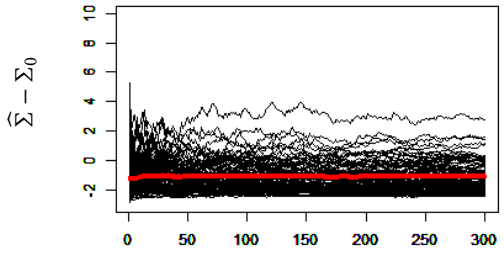


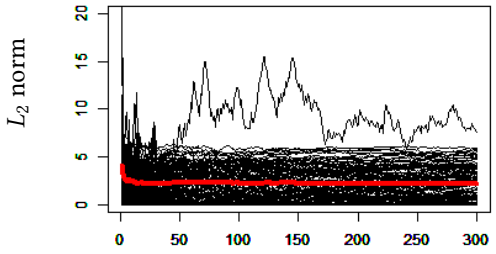
$\lambda_1 = 0 \quad \lambda_2 = 0.1$



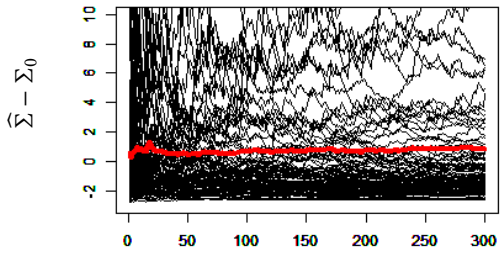


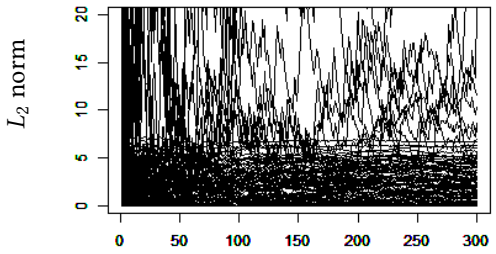
$\lambda_1 = 0 \quad \lambda_2 = 1$





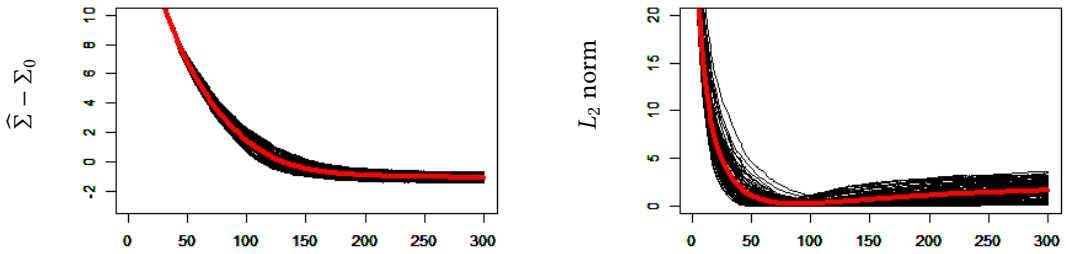
$\lambda_1 = 0 \quad \lambda_2 = 20$



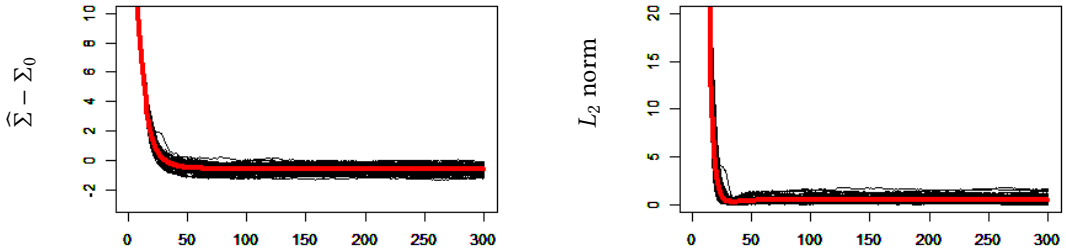


3 dimensional gaussian distribution : $(X_i)_{i\in[n]} \sim \mathcal{N}(\mu, \Sigma)$ where $\mu = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1 & 0 & 0.5 \\ 0 & 2 & 1 \\ 0.5 & 1 & 1 \end{pmatrix}$.

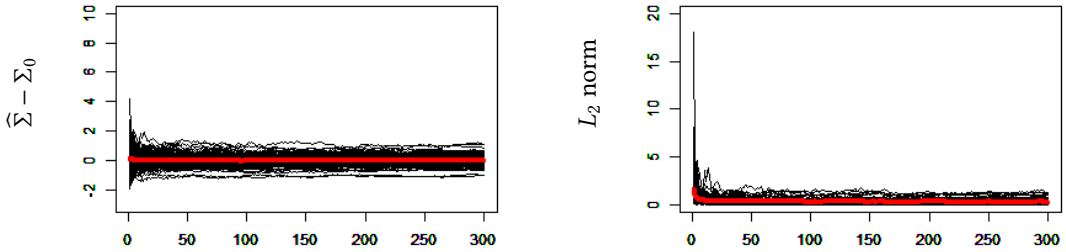
$\lambda_1 = 0 \quad \lambda_2 = 0.05$



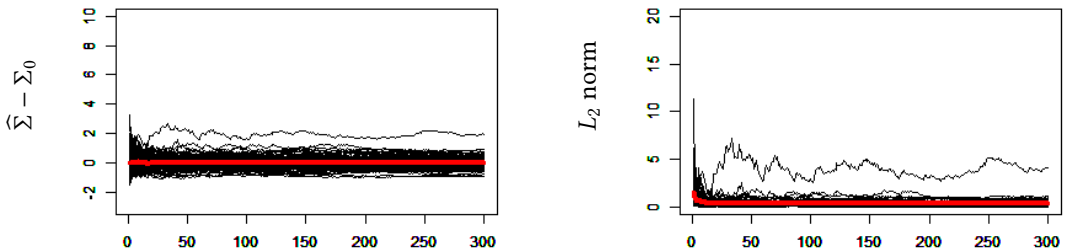
$\lambda_1 = 0 \quad \lambda_2 = 0.1$



$\lambda_1 = 0 \quad \lambda_2 = 1$



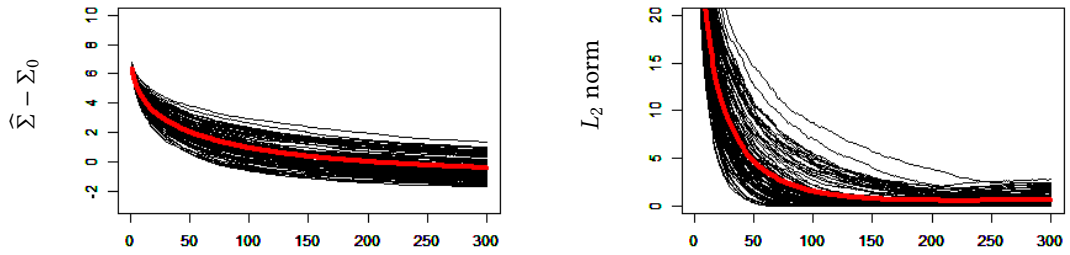
$\lambda_1 = 0 \quad \lambda_2 = 20$



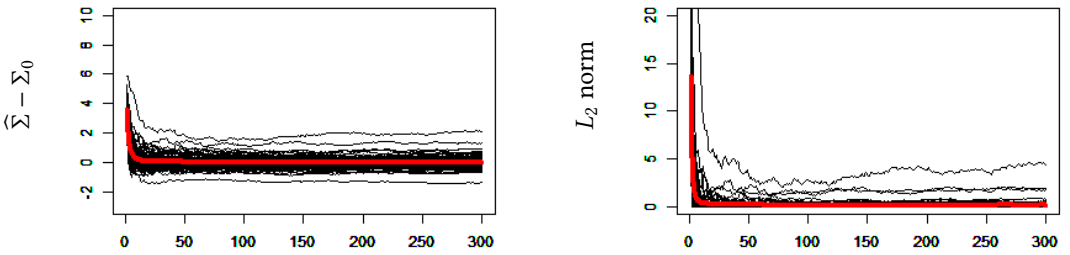
Estimation in contamination model Next, we consider the case in which the sample contains contamination. More precisely, we consider the situation in which we have samples $(Z_i)_{i\in[n]}$ with $Z_i \sim (1-\delta_i)\boldsymbol{X}_i + \delta_i C_i$ Here, $\boldsymbol{X}_i$ is an underling distribution and C_i is an error distribution, and $(\delta_i)_{i\in[n]}$ independently distribute, taking 1 with probability δ and 0 with probability $1-\delta$. The followings are the result of interpretation for contamination models with several underling distributions.

1 dimensional Gaussian distribution : $(\boldsymbol{X}_i)_{i\in[n]} \sim \mathcal{N}(1,3)$

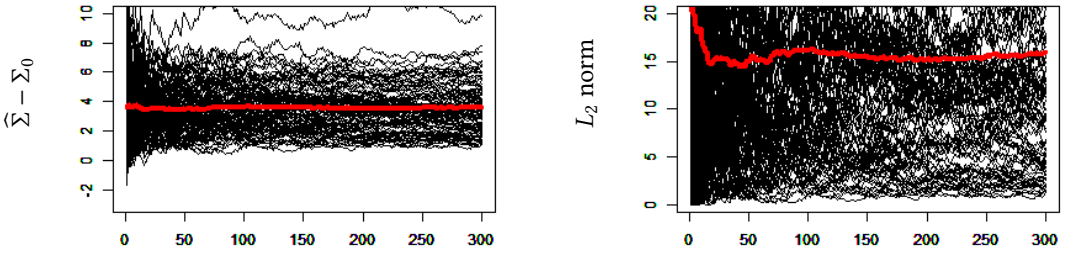
$\lambda_1 = 0 \quad \lambda_2 = 0.02$



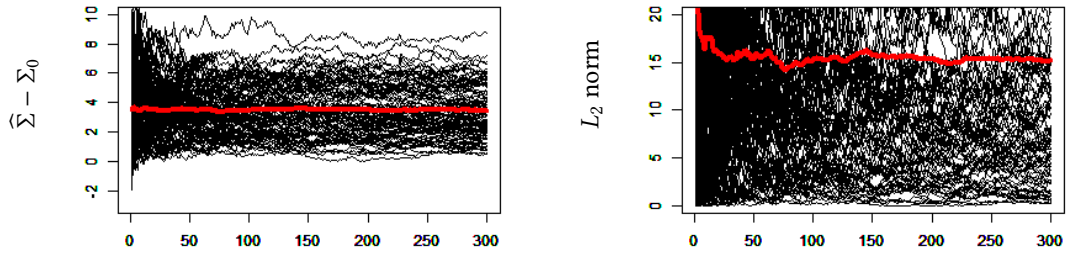
$\lambda_1 = 0 \quad \lambda_2 = 0.1$



$\lambda_1 = 0 \quad \lambda_2 = 1$

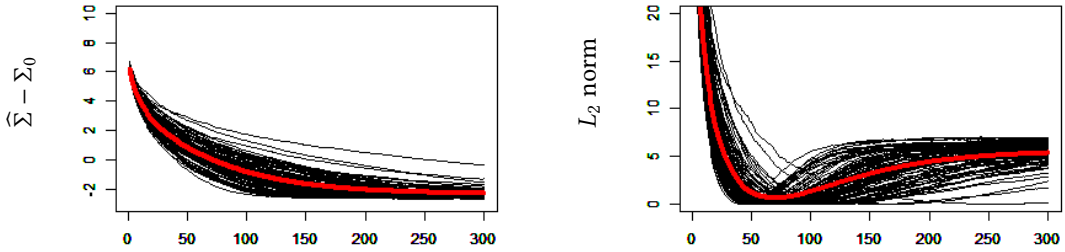


$\lambda_1 = 0 \quad \lambda_2 = 20$

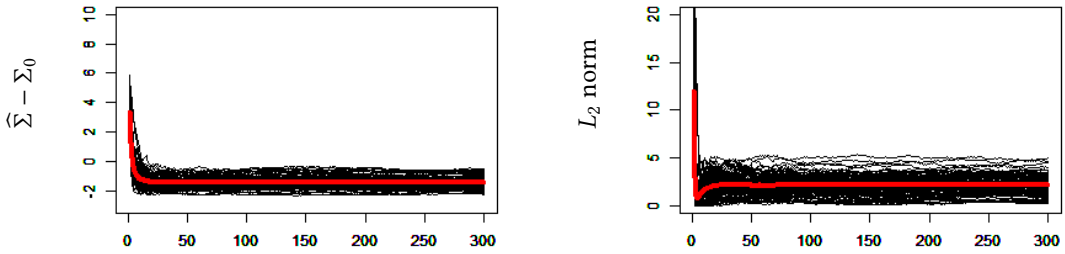


1 dimensional pareto distribution : $(X_i)_{i\in[n]} \sim Pareto(x_m = 2, \alpha = 3)$

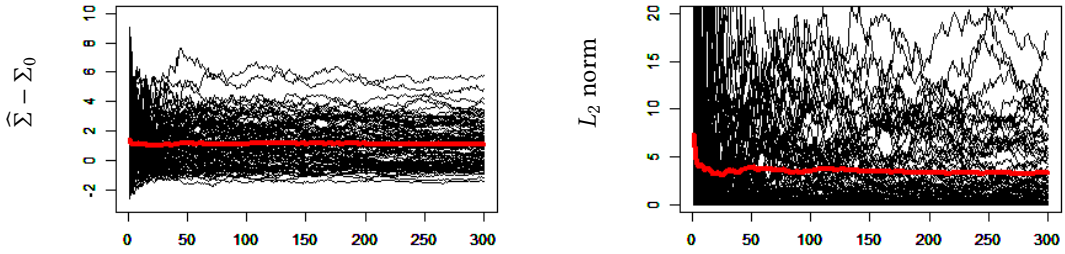
$\lambda_1 = 0 \quad \lambda_2 = 0.02$



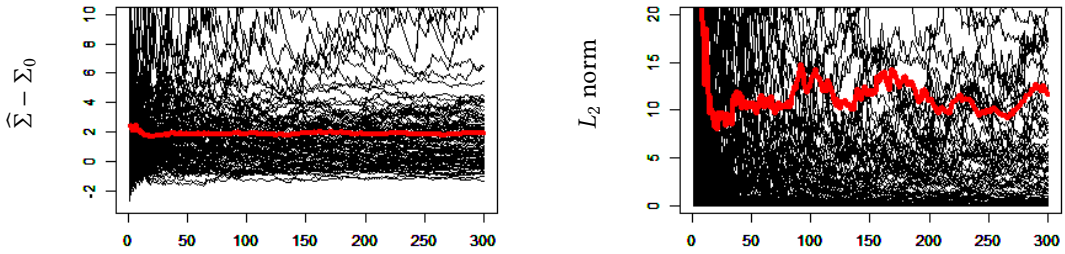
$\lambda_1 = 0 \quad \lambda_2 = 0.1$



$\lambda_1 = 0 \quad \lambda_2 = 1$

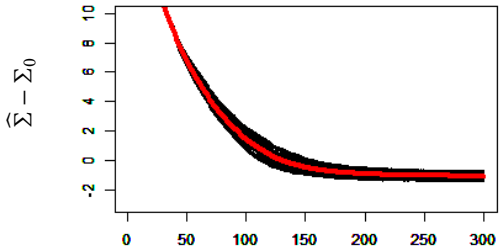


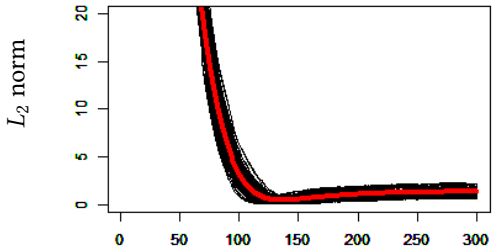
$\lambda_1 = 0 \quad \lambda_2 = 20$



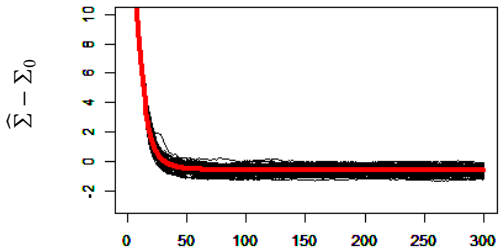
3 dimensional gaussian distribution : $(X_i)_{i\in[n]} \sim \mathcal{N}(\mu, \Sigma)$ where $\mu = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1 & 0 & 0.5 \\ 0 & 2 & 1 \\ 0.5 & 1 & 1 \end{pmatrix}$.

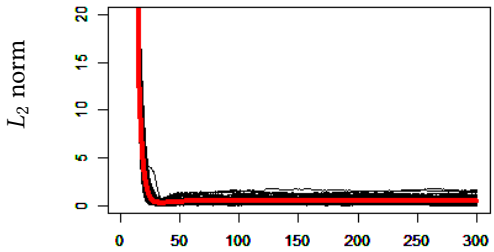
$\lambda_1 = 0 \quad \lambda_2 = 0.05$



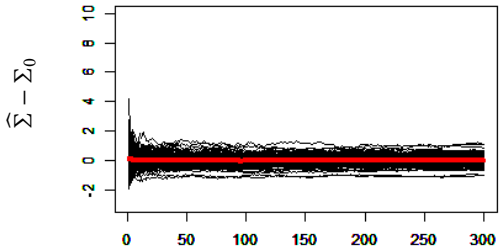


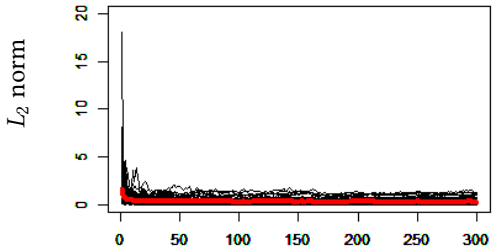
$\lambda_1 = 0 \quad \lambda_2 = 0.1$



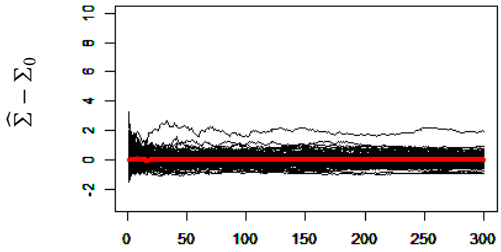


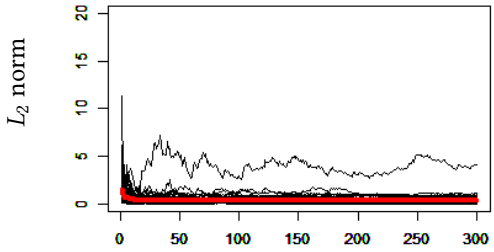
$\lambda_1 = 0 \quad \lambda_2 = 1$





$\lambda_1 = 0 \quad \lambda_2 = 20$





5 Score Matching

In this section, at first we introduce *score matching method*, in which *score matching loss* is introduced, and *score matching estimator* is defined as the minimizer of the empirical version of *score matching loss*. Our goal in this chapter is to robustify score matching estimator.

5.1 Construction of Score matching estimator

Definition 5.1.1 (Score matching loss : See [7] and [8][1]) *Suppose that distribution $P_{\boldsymbol{\theta}}$ has a twice continuously differentiable density $p_{\boldsymbol{\theta}}$ over $\mathbb{R}^d$ for all $\boldsymbol{\theta} \in \Theta$. Let $h_1, \ldots, h_d : \mathbb{R} \to \mathbb{R}$ be a.s. positive functions with respect to P_{θ_0} and absolutely continuous, and set $\boldsymbol{h}(\boldsymbol{x}) = \left(h_1\left(x^{(1)}\right), \ldots, h_d\left(x^{(d)}\right)\right)^\top$. Score matching loss with respect to $\boldsymbol{h}$ and the distribution $P_{\boldsymbol{\theta}}$ with density $p_{\boldsymbol{\theta}}$, $J_h(p_{\boldsymbol{\theta}})$ is defined as*

$$J_h(p_{\boldsymbol{\theta}}) \equiv \int_{\mathbb{R}^d} p_{\theta_0}(\boldsymbol{x}) \left| \nabla \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \circ \boldsymbol{h}(\boldsymbol{x})^{1/2} - \nabla \log p_{\theta_0}(\boldsymbol{x}) \circ \boldsymbol{h}(\boldsymbol{x})^{1/2} \right|^2 d\boldsymbol{x}, \tag{5.1}$$

here $\nabla\cdot$ means the divergence, and $\circ$ means pairwise products.

Note $J_h(p_{\boldsymbol{\theta}}) = 0$ if and only if $p_{\boldsymbol{\theta}}(x) = p_{\theta_0}(x)$ for almost every $\boldsymbol{x} \in \mathbb{R}^d$.

For the later discussion, we add the following assumption on $\boldsymbol{h}$.

$$(A1) \quad \forall j \quad \lim_{x^{(j)} \nearrow +\infty} p_{\theta_0}(x) h_j(x^{(j)}) \partial_j \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \lim_{x^{(j)} \searrow -\infty} p_{\theta_0}(x) h_j(x^{(j)}) \partial_j \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = 0,$$

$$(A2) \quad \int_{\mathbb{R}^d} \left\| \nabla \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \circ \boldsymbol{h}(\boldsymbol{x})^{1/2} \right\|_2^2 d\boldsymbol{x} < \infty, \int_{\mathbb{R}^d} \left\| (\nabla \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \circ \boldsymbol{h}(\boldsymbol{x}))' \right\|_1 d\boldsymbol{x} < \infty.$$

We consider another representation of (5.1).

$$\begin{aligned} J_h(p) = & \underbrace{\int_{\mathbb{R}^d} p_{\theta_0}(\boldsymbol{x}) \sum_{j=1}^d \left[\frac{h_j(x^{(j)})(\partial_j \log p_{\theta_0}(\boldsymbol{x}))^2}{2} \right] \mathbf{d}\boldsymbol{x}}_{\equiv \mathrm{T}_1} + \underbrace{\int_{\mathbb{R}^d} p_{\theta_0}(\boldsymbol{x}) \sum_{j=1}^d \left[\frac{h_j(x^{(j)})(\partial_j \log p_{\theta_0}(\boldsymbol{x}))^2}{2} \right] \mathbf{d}\boldsymbol{x}}_{\equiv \mathrm{T}_2} \\ & \underbrace{- \int_{\mathbb{R}^d} p_{\theta_0}(\boldsymbol{x}) \sum_{j=1}^d \left[h_j(x^{(j)}) \partial_j \log p(\boldsymbol{x}) \partial_j \log p_{\theta_0}(\boldsymbol{x}) \right] \mathbf{d}\boldsymbol{x}}_{\equiv \mathrm{T}_3}. \end{aligned}$$

[1] A specific case that $\boldsymbol{h} \equiv \mathbb{1}_d$ was introduced in [?], the generalized version was in [8].

[1] For proof, see the proposition 2 in [8].

Here $\partial_j \cdot$ means the partial derivative with respect to x_j. The second term is constant with respect to the choice of p, the third term is reformulated as the following under this assumptions (A1) and (A2).

$$
\begin{aligned}
\mathrm{T}_3 &= -\sum_{j=1}^{d} \int_{\mathbb{R}^d} p_{\theta_0}(\boldsymbol{x}) \Big[h_j(x^{(j)}) \partial_j \log p(\boldsymbol{x}) \partial_j \log p_{\theta_0}(\boldsymbol{x}) \Big] \mathbf{d}\boldsymbol{x} \\
&\underset{Fubini\ and\ (A2)}{=} -\sum_{j=1}^{d} \int_{\mathbb{R}^{d-1}} \Big[\int_{\mathbb{R}} p_{\theta_0}(\boldsymbol{x}) h_j(x^{(j)}) \partial_j \log p(\boldsymbol{x}) \partial_j \log p_{\theta_0}(\boldsymbol{x}) \mathbf{d}x^{(j)} \Big] \mathbf{d}\boldsymbol{x}_{-\mathrm{j}} \\
&\underset{\partial_j \log p_{\theta_0} = \partial_j p_{\theta_0} / p_{\theta_0}}{=} -\sum_{j=1}^{d} \int_{\mathbb{R}^{d-1}} \Big[\int_{\mathbb{R}} \partial_j p_{\theta_0}(\boldsymbol{x}) h_j(x^{(j)}) \partial_j \log p(\boldsymbol{x}) \mathbf{d}x^{(j)} \Big] \mathbf{d}\boldsymbol{x}_{-j} \\
&\underset{Integration\ by\ part}{=} -\sum_{j=1}^{d} \int_{\mathbb{R}^{d-1}} \Big[\lim_{x^{(j)} \nearrow +\infty} p_{\theta_0}(\boldsymbol{x}_{-j}; \mathrm{x}^{(j)}) h_j(x^{(j)}) \partial_j \log p(\boldsymbol{x}_{-j}; x^{(j)}) \\
&\qquad - \lim_{x^{(j)} \searrow -\infty} p_{\theta_0}(\boldsymbol{x}_{-j}; \mathrm{x}^{(j)}) h_j(x^{(j)}) \partial_j \log p(\boldsymbol{x}_{-j}; x^{(j)}) \\
&\qquad - \int_{\mathbb{R}} p_{\theta_0}(\boldsymbol{x}) \partial_j (h_j(x^{(j)}) \partial_j \log p(\boldsymbol{x})) \mathbf{d}x^{(j)} \Big] \mathbf{d}\boldsymbol{x}_{-j} \\
&\underset{(\mathrm{A1})}{=} \sum_{j=1}^{d} \int_{\mathbb{R}^{d-1}} \Big[\int_{\mathbb{R}} p_{\theta_0}(\boldsymbol{x}) \partial_j (h_j(x^{(j)}) \partial_j \log p(\boldsymbol{x})) \mathbf{d}\mathrm{x}^{(j)} \Big] \mathbf{d}\boldsymbol{x}_{-j} \\
&= \sum_{j=1}^{d} \int_{\mathbb{R}^d} p_{\theta_0}(\boldsymbol{x}) \partial_j (h_j(x^{(j)}) \partial_j \log p(\boldsymbol{x})) \mathbf{d}\boldsymbol{x} \\
&\underset{Product\ rule}{=} \sum_{j=1}^{d} \int_{\mathbb{R}^d} p_{\theta_0}(\boldsymbol{x}) h_j{}'(x^{(j)}) \partial_j \log p(\boldsymbol{x}) + \sum_{j=1}^{d} \int_{\mathbb{R}^d} p_{\theta_0}(\boldsymbol{x}) h_j(x^{(j)}) \partial_{jj} \log p(\boldsymbol{x}),
\end{aligned}
$$

where $\mathbf{d}\boldsymbol{x}_{-j} = \mathbf{d}x_1 \cdots \mathbf{d}x_{j-1} \mathbf{d}x_{j+1} \cdots \mathbf{d}x_n$. Then, *Score matching loss* is discribed as

$$
\begin{aligned}
J_h(p) &= \int_{\mathbb{R}^d} p_{\theta_0}(\boldsymbol{x}) \sum_{j=1}^{d} \Big[h_j{}'(x^{(j)}) \partial_j \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) + h_j(x^{(j)}) \partial_{jj} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) + \frac{h_j(x^{(j)}) (\partial_j \log p(\boldsymbol{x}))^2}{2} \Big] \mathbf{d}\boldsymbol{x} + \text{const.} \\
&= \mathbb{E}_{p_{\theta_0}} \sum_{j=1}^{d} \Big[h_j{}'(X^{(j)}) \partial_j \log p_{\boldsymbol{\theta}}(\boldsymbol{X}) + h_j(X^{(j)}) \partial_{jj} \log p_{\boldsymbol{\theta}}(\boldsymbol{X}) + \frac{h_j(X^{(j)}) (\partial_j \log p_{\theta_0}(\boldsymbol{X}))^2}{2} \Big] + \text{const.}
\end{aligned} \tag{5.2}
$$

5.2 Score matching estimator for exponential family

Next, we consider tha case in which $\mathcal{P}$ is an *exponential family*. Exponential family is defined as the following.

Definition 5.2.1 *We call $\mathcal{P} \equiv \{p_\theta : \boldsymbol{\theta} \in \Theta\}$ an exponential family when the density p_θ is represented by*

$$
\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{t}(\boldsymbol{x}) - a(\boldsymbol{\theta}) + b(\boldsymbol{x}) \tag{5.3}
$$

Here, $\boldsymbol{\theta}$ and $\boldsymbol{t}(\boldsymbol{x})$ are called *canonical parameter* and *sufficient statistics*. An example of exponential family is multivariate gaussian distribution, which we check later. In our setting we assume the canonical parameter

of the population distribution θ_0 belongs to Θ. In case of an exponential family, (5.2) can be rewritten as follows (omitting constant term).

$$\begin{aligned}
&\mathbb{E}_{p_{\theta_0}} \sum_{j=1}^{d} \left[h_j'(X^{(j)})\{\boldsymbol{\theta}^\top \boldsymbol{t}_j'(\boldsymbol{X}) + b_j'(\boldsymbol{X})\} + h_j(X^{(j)})\{\boldsymbol{\theta}^\top \boldsymbol{t}_j''(\boldsymbol{X}) + b_j''(\boldsymbol{X})\} + \frac{1}{2}\{h_j(X^{(j)})(\boldsymbol{\theta}^\top \boldsymbol{t}_j'(\boldsymbol{X}) + b_j'(\boldsymbol{X}))^2 \right] \\
= \quad &\mathbb{E}_{p_{\theta_0}} \sum_{j=1}^{d} \Big[h_j'(X^{(j)})\{\boldsymbol{\theta}^\top \boldsymbol{t}_j'(\boldsymbol{X}) + \underbrace{b_j'(\boldsymbol{X})}_{\text{const. w.r.t } \boldsymbol{\theta}} \} + h_j(X^{(j)})\{\boldsymbol{\theta}^\top \boldsymbol{t}_j''(\boldsymbol{X}) + \underbrace{b_j''(\boldsymbol{X})}_{\text{const. w.r.t } \boldsymbol{\theta}} \} \\
&\qquad + \frac{1}{2} h_j(X^{(j)})(\boldsymbol{\theta}^\top \boldsymbol{t}_j'(\boldsymbol{X}))^2 + (\boldsymbol{\theta}^\top \boldsymbol{t}_j'(\boldsymbol{X})) b_j'(\boldsymbol{X}) + \underbrace{\frac{1}{2} b_j'(\boldsymbol{X})^2}_{\text{const. w.r.t } \boldsymbol{\theta}} \Big] \\
= \mathbb{E}_{p_{\theta_0}} &\sum_{j=1}^{d} \left[h_j'(X^{(j)})\{\boldsymbol{\theta}^\top \boldsymbol{t}_j'(\boldsymbol{X})\} + h_j(X^{(j)})\{\boldsymbol{\theta}^\top \boldsymbol{t}_j''(\boldsymbol{X})\} + \frac{1}{2} h_j(X^{(j)})(\boldsymbol{\theta}^\top \boldsymbol{t}_j'(\boldsymbol{X}))^2 + h_j(X^{(j)})(\boldsymbol{\theta}^\top \boldsymbol{t}_j'(\boldsymbol{X})) b_j'(\boldsymbol{X}) \right]
\end{aligned}$$

Here $\boldsymbol{t}_j' \equiv \partial_j \boldsymbol{t}$, $b_j' \equiv \partial_j b$, $\boldsymbol{t}_j'' \equiv \partial_{jj} \boldsymbol{t}$, and $b_j'' \equiv \partial_{jj} b$. Summarizing the above discussion, we get the following theorem.

Theorem 5.2.2 *In the exponential family of the distributions, we have that*

$$J_h(p_\theta) = \frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\Gamma} \boldsymbol{\theta} - \boldsymbol{g}^\top \boldsymbol{\theta} + \text{const.} \tag{5.4}$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^{d \times d}$ *and* $\boldsymbol{g} \in \mathbb{R}^d$ *are*

$$\boldsymbol{\Gamma} \equiv \mathbb{E}_{p_{\theta_0}} \sum_{j=1}^{d} h_j(X^{(j)}) \boldsymbol{t}_j'(\boldsymbol{X}) \boldsymbol{t}_j'(\boldsymbol{X})^\top, \tag{5.5}$$

$$\boldsymbol{g} \equiv -\mathbb{E}_{p_{\theta_0}} \sum_{j=1}^{d} [h_j(X^{(j)}) b_j'(\boldsymbol{X}) \boldsymbol{t}_j'(\boldsymbol{X}) + h_j(X^{(j)}) \boldsymbol{t}_j''(\boldsymbol{X}) + h_j'(X^{(j)}) \boldsymbol{t}_j'(\boldsymbol{X})]. \tag{5.6}$$

Corollary 5.2.3 *Let $\boldsymbol{X}$ distribute from an exponential family, define $\boldsymbol{\Gamma}$ by* (5.5) *and $\boldsymbol{g}$ by* (5.6)*. Under the condition that $\boldsymbol{\Gamma}$ is invertible. Then*

$$\operatorname*{argmin}_{\boldsymbol{\theta}} J_h(p_\theta) = \boldsymbol{\Gamma}^{-1} \boldsymbol{g}$$

From the definition of the score matching loss function, the minimizer coincides with the true parameter $\boldsymbol{\theta}_0$. The above discussion provides us an idea to construct an estimator of the true parameter $\boldsymbol{\theta}_0$: The population quantities $\boldsymbol{\Gamma}, \boldsymbol{g}$ define the true parameter $\boldsymbol{\theta}_0$. So, in order to find the estimator for $\boldsymbol{\theta}_0$, we apply plugin principle and replace correspondent population analogs by their empirical counterparts.

Definition 5.2.4 *Let X be a random vector with the distribution P_{θ_0} which belongs to an exponential family $\mathcal{P}$. For independent copies $(\boldsymbol{X}_i)_{i \in [n]}$ of $\boldsymbol{X}$, we define an estimator $\hat{\boldsymbol{\theta}}_{SM}$ by*

$$\hat{\boldsymbol{\theta}}_{\mathrm{SM}} \equiv \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \left(\frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\Gamma}(\boldsymbol{X}) \boldsymbol{\theta} - \boldsymbol{g}(\boldsymbol{X})^\top \boldsymbol{\theta} \right)$$

here

$$\begin{aligned}
\boldsymbol{\Gamma}(\boldsymbol{X}) &\equiv \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} h_j(X_i^{(j)}) \boldsymbol{t}_j'(\boldsymbol{X}_i) \boldsymbol{t}_j'(\boldsymbol{X}_i)^\top, \\
\boldsymbol{g}(\boldsymbol{X}) &\equiv -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} [h_j(X_i^{(j)}) b_j'(\boldsymbol{X}_i) \boldsymbol{t}_j'(\boldsymbol{X}_i) + h_j(X_i^{(j)}) \boldsymbol{t}_j''(\boldsymbol{X}_i) + h_j(X_i^{(j)}) \boldsymbol{t}_j'(\boldsymbol{X}_i)].
\end{aligned}$$

For later discussion, we define the following notations.

Definition 5.2.5 *For random vectors $(\boldsymbol{X}_i)_{i\in[n]}$ and function $\boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}^d$ satisfying the assumptions of Definition 5.2.4 and for $i \in [n]$, we define*

$$\boldsymbol{\Gamma}(\boldsymbol{X}_i) \equiv \sum_{j=1}^{d} h_j(X_i^{(j)})\boldsymbol{t}_j'(\boldsymbol{X}_i)\boldsymbol{t}_j'(\boldsymbol{X}_i)^\top, \tag{5.7}$$

$$-\boldsymbol{g}(\boldsymbol{X}_i) \equiv \sum_{j=1}^{d} [h_j(X_i^{(j)})b_j'(\boldsymbol{X}_i)\boldsymbol{t}_j'(\boldsymbol{X}_i) + h_j(X_i^{(j)})\boldsymbol{t}_j''(\boldsymbol{X}_i) + h_j'(X_i^{(j)})\boldsymbol{t}_j'(\boldsymbol{X}_i)]. \tag{5.8}$$

Note (5.7) and (5.8), $\boldsymbol{\Gamma}(\boldsymbol{X})$ and $\boldsymbol{g}(\boldsymbol{X})$ can be written by

$$\boldsymbol{\Gamma}(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Gamma}(\boldsymbol{X}_i) \qquad \boldsymbol{g}(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{X}_i),$$

and $\boldsymbol{\Gamma}(\boldsymbol{X})$ and $\boldsymbol{\Gamma}(\boldsymbol{X}_i)$ are Hermitian matrices from the definition.

5.3 Error bound of Score matching estimator

Our interest is the behavior of the difference $\hat{\boldsymbol{\theta}}_{\mathrm{SM}} - \boldsymbol{\theta}_0$. The following lemma provides an error decomposition of $SME\boldsymbol{\theta} - \boldsymbol{\theta}_0$.

Lemma 5.3.1 *For a given exponential family $\mathcal{P} \equiv \{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ with the true parameter $\boldsymbol{\theta}_0$, define $\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}(\boldsymbol{X})$, $\boldsymbol{g}$, $\boldsymbol{g}(\boldsymbol{X})$ as the above, and assume $\boldsymbol{\Gamma}(\boldsymbol{X})$ is invertible a.s. Then*

$$\hat{\boldsymbol{\theta}}_{\mathrm{SM}} - \boldsymbol{\theta}_0 = \boldsymbol{\Gamma}(\boldsymbol{X})^{-1}(\boldsymbol{g}(\boldsymbol{X}) - \boldsymbol{g}) + \boldsymbol{\Gamma}(X)^{-1}(\boldsymbol{\Gamma}(\boldsymbol{X}) - \boldsymbol{\Gamma})\boldsymbol{\theta}_0. \tag{5.9}$$

Proof.

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathrm{SM}} - \boldsymbol{\theta}_0 &= \Gamma^{-1}(X)\,(g(X) - \Gamma(X)\boldsymbol{\theta}_0)\\ &= \Gamma^{-1}(X)\,(g(X) - g + g - \Gamma(X)\boldsymbol{\theta}_0)\\ &= \Gamma^{-1}(X)((g(X) - g)) + \Gamma^{-1}(X)(\Gamma - \Gamma(X))\boldsymbol{\theta}_0.\end{aligned}$$

□

From the triangle inequality and the properties of operator norm and Euclidean norm : $|Ax| \le \|A\|\,|x|$ for $A \in \mathbb{R}^{d\times d}$, $x \in \mathbb{R}^d$, we can derive the following inequality for the difference between $\boldsymbol{\theta}_{SM}$ and $\boldsymbol{\theta}_0$

Lemma 5.3.2

$$\left|\hat{\boldsymbol{\theta}}_{\mathrm{SM}} - \boldsymbol{\theta}_0\right| \le \left\|\Gamma^{-1}(X)\right\|\left(|g(X) - g| + \|\Gamma(X) - \Gamma\|\,|\boldsymbol{\theta}_0|\right).$$

Our goal is finding the concentration inequality of $\hat{\boldsymbol{\theta}}_{\mathrm{SM}}$, that is finding the (probabilistic) upper bound of $|\hat{\boldsymbol{\theta}}_{\mathrm{SM}} - \boldsymbol{\theta}_0|$. From the lemma 5.3.1, this problem boils down to the problem to find the upper bound of $|g(X) - g|$ and $\|\Gamma(X) - \Gamma\|$.

As an example, at first we consider the case in which the population distribution satisfies that $|\boldsymbol{g}(\boldsymbol{X}_1)|$ and $\|\boldsymbol{\Gamma}(\boldsymbol{X}_1)\|$ are bounded. In this case, there exist L and R such that $|\boldsymbol{g}(\boldsymbol{X}_1) - \boldsymbol{g}| \le L$ amd $\|\boldsymbol{\Gamma}(\boldsymbol{X}_1) - \boldsymbol{\Gamma}\| \le R$. Thus, Corollary 3.1.23 provides us with a bound of $|g(\boldsymbol{X}) - g|$. Define $C \equiv \frac{\mathrm{Tr}g(X_1)g(X_1)^\top}{d}$, then with probability $1-\delta$ it hold at least that

$$|\boldsymbol{g}(X) - \boldsymbol{g}| \le \frac{1}{n}\sqrt{dC}\sqrt{\log\frac{2}{\delta}} + \frac{2L}{n}\log\frac{2}{\delta}. \tag{5.10}$$

And Theorem 3.2.1 provide us with the bound of $\|\boldsymbol{\Gamma}(\boldsymbol{X}) - \boldsymbol{\Gamma}\|$. With probability $1-\delta$it holds that

$$\|\boldsymbol{\Gamma}(\boldsymbol{X}) - \boldsymbol{\Gamma}\| \le \sqrt{\frac{2}{n}R^2\,\log\left(\frac{d}{\delta}\right)} + \frac{2R}{3n}\log\left(\frac{d}{\delta}\right) \tag{5.11}$$

Replacing δ to $\frac{\delta}{2}$ of (5.10) and (5.11), and taking the union bound, we obtain the following.

Corollary 5.3.3 *Let $(X_i)_{i\in[n]} \in \mathbb{R}^d$ and $(h_j)_{j\in[d]} : \mathbb{R} \to \mathbb{R}$ be random vectors and positive functions which satisfy the assumption of Definition 5.1.1 . Define $\boldsymbol{g}(X)$, $\Gamma(X)$, and$\boldsymbol{g}(X_i)$, $\boldsymbol{\Gamma}(X_i)$, as in Definition 5.2.4 and Definition 5.2.5. Assume that*

$$|\boldsymbol{g}(X_i) - \boldsymbol{g}| \leq L \quad \textit{and} \quad \|\boldsymbol{\Gamma}(X_i) - \boldsymbol{\Gamma}\| \leq R$$

Then,

$$\left|\hat{\theta}_{\mathrm{SM}} - \theta_0\right| \;\leq\; \left\|\Gamma^{-1}(X)\right\| \left(\frac{1}{n}\sqrt{dC}\sqrt{\log\frac{4}{\delta}} + \frac{2L}{n}\log\frac{4}{\delta} + \sqrt{\frac{2}{n}R^2\,\log\left(\frac{2d}{\delta}\right)} + \frac{2R}{3n}\log\left(\frac{2d}{\delta}\right)|\theta_0| \right).$$

5.4 Score matching estimator in Gaussian case

5.4.1 Original score matching estimator in Gaussian case

In this section, we consider the case in which $\boldsymbol{X}$ is distributed as a Gaussian distribution $\mathcal{N}_d(\boldsymbol{\mu}, \Omega^{-1})$. At first we discuss the case of $\boldsymbol{h} \equiv \mathbb{1}_d$ The probability density function is given by

$$f(\boldsymbol{X}; \boldsymbol{\mu}, \Omega) \equiv \frac{1}{\sqrt{2\pi|\Omega|^{-1}}} \exp\left(-\frac{1}{2}(x-\boldsymbol{\mu})^\top \Omega (x-\boldsymbol{\mu})\right).$$

Then,

$$\begin{aligned} \partial^j \log f(\boldsymbol{X}; \boldsymbol{\mu}, \Omega) &= -(\Omega(\boldsymbol{X}-\boldsymbol{\mu}))_j \\ \partial^{jj} f(\boldsymbol{X}; \boldsymbol{\mu}, \Omega) &= -\omega_{jj} \\ h_j{}' &\equiv 0 \end{aligned}$$

, where $\left(\omega_{ij}\right)_{i,j\in[n]} \equiv \Omega$. Thus, we obtain

$$J_h(\boldsymbol{\mu}, \Omega) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\left\{-\omega_{jj} + \frac{1}{2}(X_i-\boldsymbol{\mu})^\top \Omega^2 (X_i-\boldsymbol{\mu})\right\}.$$

From the definition, the score matching estimator $\hat{\boldsymbol{\mu}}_{SM}$, $\hat{\Omega}_{SM}$ is defined as the minimizer of $J_h(\boldsymbol{\mu}, \Omega)$. The gradient is

$$\nabla_{\boldsymbol{\mu}} J_h(\boldsymbol{\mu}, \Omega) = \Omega^2\mu - \Omega^2 \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{5.12}$$

$$\nabla_{\Omega} J_h(\boldsymbol{\mu}, \Omega) = -I_d + \Omega\frac{1}{2n}\sum_{i=1}^{n}(\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_i-\boldsymbol{\mu})^\top + \frac{1}{2n}\left\{\sum_{i=1}^{n}(\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_i-\boldsymbol{\mu})^\top\right\}\Omega, \tag{5.13}$$

Ω^2 is positive definite, so

$$(5.12) = 0 \Leftrightarrow \boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

. Under this constraint with respect to $\boldsymbol{\mu}$,

$$(5.13) = 0 \quad \Leftrightarrow \quad \Omega = \left(\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_{SM})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_{SM})^\top\right)^{-1}.$$

Thus, the score maching estimator $\boldsymbol{\theta}_{SM} = (\hat{\boldsymbol{\mu}}_{SM}, \hat{\Sigma}_{\mathrm{SM}})$ is

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{\mathrm{SM}} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \\ \hat{\Sigma}_{\mathrm{SM}} &= \left(\hat{\Omega}_{\mathrm{SM}}\right)^{-1} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_{SM})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_{SM})^\top \end{aligned} \tag{5.14}$$

On the other hand, one of other estimation approaches is Maximum likelihood estimation. Maximum likelihood estimation is a method to estimate the parameter of the population distribution by searching for the maximizer of the likelihood function. In Gaussian case, the log-likelihood function is

$$\begin{aligned} \ell(\boldsymbol{\mu}, \Omega) &= const. + \frac{n}{2}\log|\Omega| - \sum_{i=1}^{n}\frac{1}{2}(\boldsymbol{X}_i - \boldsymbol{\mu})^\top \Omega(\boldsymbol{X}_i - \boldsymbol{\mu}) \\ &= const. + \frac{n}{2}\log|\Omega| - \sum_{i=1}^{n}\frac{1}{2}(\boldsymbol{X}_i - \overline{X} + \overline{X} - \boldsymbol{\mu})^\top \Omega(\boldsymbol{X}_i - \overline{X} + \overline{X} - \boldsymbol{\mu}) \\ &= const. + \frac{n}{2}\log|\Omega| - \frac{1}{2}\Big\{ \underbrace{\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{X})^\top \Omega(\boldsymbol{X}_i - \overline{X})}_{=n\mathrm{Tr}\left(\Omega\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{X})(\boldsymbol{X}_i - \overline{X})^\top\right)} + 2\underbrace{\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{X})^\top \Omega(\overline{X} - \boldsymbol{\mu})}_{=0} + \sum_{i=1}^{n}(\overline{X} - \boldsymbol{\mu})^\top \Omega(\overline{X} - \boldsymbol{\mu})\Big\} \\ &= const. + \frac{n}{2}\log|\Omega| - \frac{n}{2}\mathrm{Tr}(\Omega\overline{S}) - \frac{n}{2}(\overline{X} - \boldsymbol{\mu})^\top \Omega(\overline{X} - \boldsymbol{\mu}). \end{aligned}$$

Here, $\overline{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$, $\overline{S} \equiv \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{X})(\boldsymbol{X}_i - \overline{X})^\top$. The last term $= 0$ if and only if $\boldsymbol{\mu} = \overline{X}$ since Ω is positive definite. Thus, the Maximum likelihood estimator $\boldsymbol{\theta}_{ML} \equiv (\hat{\boldsymbol{\mu}}_{\mathrm{ML}}, \hat{\Sigma}_{\mathrm{ML}})$ is given by

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{\mathrm{ML}} &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i, \\ \hat{\Sigma}_{\mathrm{ML}} &= \left(\hat{\Omega}_{\mathrm{ML}}\right)^{-1} = \left(\underset{\Omega\in PD_d}{\arg\max}\ \log|\Omega| - \mathrm{Tr}(\Omega\overline{S})\right)^{-1}. \end{aligned}$$

Lemma 5.4.1 *Regarding $\hat{\Omega}_{ML}$, if $\overline{S}$ is positive definite, then $\hat{\Omega}_{ML} = \overline{S}^{-1}$. And if $\overline{S}$ is singular, then $\hat{\Omega}_{ML}$ doesn't exist.*

Proof.
(First case) Since $\overline{S}$ is positive definite, there exists a orthogonal matrix Q s.t. $\overline{S} = Q\Lambda Q^\top$,where Λ is a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ with the eigen values $(\lambda_j)_{j\in[d]}$ of $\hat{\Omega}$ in the diagonal entries. When we define a map $F : \mathbb{R}^{d\times d} \to \mathbb{R}$ by $F_{\overline{S}}(\Omega) \equiv \log|\Omega| - \mathrm{Tr}(\Omega\overline{S})$,

$$F_{\overline{S}}(\Omega) = F_{Q^\top\overline{S}Q}(Q^\top\Omega Q)$$

Thus,

$$\underset{\Omega\in PD_d}{\arg\max} F_{\overline{S}}(\Omega) = Q \cdot \underset{\Omega\in PD_d}{\arg\max} F_{\Lambda}(\Omega) \cdot Q^\top$$

Let $\Omega = AA^\top$ be the Cholesky decomposition, where A is lower-triangular with the (i, j) entry $= \begin{cases} a_{i,j} \text{ for } i \geq j \\ 0 \ other \end{cases}$.

Then

$$F_{\Lambda}(\Omega) = \sum_{i=1}^{d}\left(\log a_{ii}^2 - a_{ii}^2\lambda_i - \sum_{j<i} a_{ij}^2\lambda_i\right). \tag{5.15}$$

From the assumption on $\overline{S}$, $\lambda_i > 0$ for all $i \in [d]$. For $\lambda > 0$, the function $\log a - a\lambda$ is strictly concave for $a \in (0, \infty)$ and takes its unique maximum for $a = 1/\lambda$. The term $-\sum_{i<j}\left(a_{ij}^2\lambda_i\right)$ attains its maximum when all $a_{ij} = 0$. Thus, F_Λ is maximized at $\tilde{\Omega} = \mathrm{diag}\,(1/\lambda_1, \ldots, 1/\lambda_d)$. We conclude that $F_{\overline{S}}$ takes the maximum at

$$\Omega = Q\tilde{\Omega}Q^\top = \left(Q^\top \Lambda Q\right)^{-1} = \overline{S}^{-1}$$

(Second case) If $\overline{S}$ is singular, then there exists an eigenvalue $\lambda_i = 0$, and the i-th term in (5.15) is $\log a_{ii}^2 \to \infty$ as $a_{ii} \to \infty$. Hence, $F_{\overline{S}}(\Omega)$ is unbounded.

□

Thus, under the condition that

1. the distribution comes from a Gaussian distribution,
2. $\boldsymbol{h} \equiv \mathbb{1}_d$,
3. no constraint on the parameters except for the symmetry and positive definiteness of Ω,

the score-matching estimator coincides with the maximum likelihood estimator.

5.4.2 Generalized score matching estimator in Gaussian case

Next we consider the case in which the distribution is Gaussian but the function $\boldsymbol{h}$ is general. Note that $\mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$ has d mean parameters and $\frac{d(d+1)}{2}$ covariance parameters. We write the $\frac{d(d+1)}{2} + d$ parameters by

$$\boldsymbol{\theta} \equiv (\omega_{11}, \ldots, \omega_{1d}, \eta_1\ ,\ \omega_{22}, \ldots, \omega_{2d}, \eta_2\ ,\ \ldots\ ,\ \omega_{dd}, \eta_d)^\top \quad \in \quad \mathbb{R}^{\frac{d(d+1)}{2}+d}$$

In order to make the next discussion clear, we introduce the following map R.

Definition 5.4.2 *We define* $\mathcal{R} : R^{\frac{d(d+1)}{2}+d} \to \mathbb{R}^{(d+1)d}$ *by*

$$R(\boldsymbol{\theta}) \equiv (\omega_{11}, \ldots, \omega_{1d}, \eta_1\ ,\ \omega_{12}, \omega_{22}, \ldots, \omega_{2d}, \eta_2\ ,\ \ldots,\ \omega_{1d}, \ldots, \omega_{dd}, \eta_d)^\top$$

Note that R *can be written as* $R \equiv \mathrm{vec} \circ R'$, *where* $R' : \mathbb{R}^{\frac{d(d+1)}{2}+d} \to \mathbb{R}^{(d+1)\times d}$ *is defined by*

$$R'(\boldsymbol{\theta}) \equiv \begin{pmatrix} \omega_{11} & \cdots & \omega_{1d_2} \\ \vdots & \ddots & \vdots \\ \omega_{d_1 1} & \cdots & \omega_{d_1 d_2} \\ \eta_{d_1 1} & \cdots & \eta_{d_1 d_2} \end{pmatrix}$$

The logarithm density function of $\mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$ is given by

$$\begin{aligned}
\log f(\boldsymbol{x}) &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \text{const.} \\
&= -\frac{1}{2}\boldsymbol{x}^\top \Sigma^{-1}\boldsymbol{x} + \boldsymbol{\mu}^\top \Sigma^{-1}\boldsymbol{x} + \text{const.} \\
&= -\frac{1}{2}\boldsymbol{x}^\top \Omega \boldsymbol{x} + \eta^\top \boldsymbol{x} + \text{const.} \qquad \textit{here, } \Omega \equiv \Sigma^{-1} \textit{and } \eta \equiv \Sigma^{-1}\boldsymbol{\mu} \\
&= -\frac{1}{2}\mathrm{Tr}(\Omega \boldsymbol{x}\boldsymbol{x}^\top) + \eta^\top \boldsymbol{x} + \text{const.} \\
&= (\boldsymbol{R}\boldsymbol{\theta})^\top \left(\frac{{x_1}^2}{2}, \ldots, \frac{x_1 x_d}{2}, x_1, \frac{x_2 x_1}{2}, \ldots, \frac{x_2 x_d}{2}, x_2, \ldots, \frac{x_d x_1}{2}, \ldots, \frac{x_d x_d}{2}, x_d\right)^\top + \text{const.} \\
&= \underbrace{\boldsymbol{\theta}^\top \boldsymbol{R}^\top \left(\frac{{x_1}^2}{2}, \ldots, \frac{x_1 x_d}{2}, x_1, \frac{x_2 x_1}{2}, \ldots, \frac{x_2 x_d}{2}, x_2, \ldots, \frac{x_d x_1}{2}, \ldots, \frac{x_d x_d}{2}, x_d\right)^\top}_{=\boldsymbol{t}(\boldsymbol{x})} + \text{const.}
\end{aligned}$$

Note Since the map $PD_d \times \mathbb{R}^d \to PD_d \times \mathbb{R}^d \ : \ (\Sigma, \boldsymbol{\mu}) \mapsto (\Omega, \eta)$ is bijection, the problem to find the parameter $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma)$ which minimizes score matching loss $J_h(p_{\boldsymbol{\theta}})$ can be replaced by the problem to find the parameter $(\Omega, \boldsymbol{\eta})$ which minimizes $J_h(p_{\boldsymbol{\theta}})$.
The above transformation illustrates that Gaussian distribution is an exponential family with the canonical parameter $\boldsymbol{\theta}$ and $b(\boldsymbol{x}) = 0$. Applying theorem 5.2.2, we get the following.

Corollary 5.4.3 (Score matching estimator for Gaussian distribution) *Let X come from a Gaussian distribution $\mathcal{N}(\eta\Omega^{-1}, \Omega^{-1})$ and $(X_i)_{1\leq i\leq n}$ be independent copeis of $\boldsymbol{X}$. Then, $\boldsymbol{\Gamma}(\boldsymbol{X})$ and $\boldsymbol{g}(X)$ for the score matching estimator (Ω, η) is given by*

$$\boldsymbol{\Gamma}(\boldsymbol{X}) = \boldsymbol{R}^\top \begin{pmatrix} \boldsymbol{\Gamma}_1(\boldsymbol{X}) & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{\Gamma}_m(\boldsymbol{X}) \end{pmatrix} \boldsymbol{R}, \qquad \boldsymbol{g}(\boldsymbol{X}) = \boldsymbol{R}^\top \begin{pmatrix} \boldsymbol{g}_{1,1}(\boldsymbol{X}) \\ g_{2,1}(\boldsymbol{X}) \\ \vdots \\ \boldsymbol{g}_{1,d}(\boldsymbol{X}) \\ g_{2,d}(\boldsymbol{X}) \end{pmatrix}. \tag{5.16}$$

Here

$$\begin{aligned} \boldsymbol{\Gamma}_j(\boldsymbol{X}) &\equiv \begin{bmatrix} \boldsymbol{\Gamma}_{11,j} & \boldsymbol{\Gamma}_{12,j} \\ \boldsymbol{\Gamma}_{21,j} & \boldsymbol{\Gamma}_{22,j} \end{bmatrix} \equiv \frac{1}{n}\sum_{i=1}^n \begin{bmatrix} h_j(X_j^{(i)})\boldsymbol{X}^{(i)}\boldsymbol{X}^{(i)T} & -h_j(X_j^{(i)})\boldsymbol{X}^{(i)} \\ -h_j(X_j^{(i)})\boldsymbol{X}^{(i)T} & h_j(X_j^{(i)}) \end{bmatrix} \qquad \in \mathbb{R}^{(d+1)\times(d+1)}, \\ \boldsymbol{g}_{1,j} &\equiv \frac{1}{n}\sum_{i=1}^n h_j'(X_j^{(i)})X_j^{(i)} + h_j(X_j^{(i)})\mathbf{e}_{j,d} \qquad \in \mathbb{R}^d, \\ g_{2,j} &\equiv \frac{1}{n}\sum_{i=1}^n h_j'(X_j^{(i)}) \qquad \in \mathbb{R}. \end{aligned}$$

R is defined in Definition 5.4.2, and $\mathbf{e}_{j,d} \in \mathbb{R}^d$ is 1 for the j-th entry and 0 otherwise, that is, $\mathbf{e}_{j,d} \equiv (0, \ldots, 0, \underset{\text{j th}}{1}, 0, \ldots, 0)^\top$
.

I case of $\boldsymbol{h} \equiv 1$, this corollary can be written as the following.

Corollary 5.4.4 (Score matching estimator for Gaussian distribution in case of $h \equiv \mathbb{1}_d$) *Let X come from a Gaussian distribution $\mathcal{N}(\eta\Omega^{-1}, \Omega^{-1})$ and $(X_i)_{1\leq i\leq n}$ be independent copeis of $\boldsymbol{X}$. Take $h \equiv \mathbb{1}_d$. Then, $\boldsymbol{\Gamma}(\boldsymbol{X})$ and $\boldsymbol{g}(X)$ for the score matching estimator (Ω, η) is given by*

$$\boldsymbol{\Gamma}(\boldsymbol{X}) = \boldsymbol{R}^\top \begin{pmatrix} \boldsymbol{\Gamma}_0(\boldsymbol{X}) & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{\Gamma}_0(\boldsymbol{X}) \end{pmatrix} \boldsymbol{R}, \qquad \boldsymbol{g}(\boldsymbol{X}) = \boldsymbol{R}^\top \begin{pmatrix} \boldsymbol{g}_{1,1}(\boldsymbol{X}) \\ g_{2,1}(\boldsymbol{X}) \\ \vdots \\ \boldsymbol{g}_{1,d}(\boldsymbol{X}) \\ g_{2,d}(\boldsymbol{X}) \end{pmatrix}. \tag{5.17}$$

Here

$$\begin{aligned} \boldsymbol{\Gamma}_0(\boldsymbol{X}) &\equiv \begin{bmatrix} \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22} \end{bmatrix} \equiv \frac{1}{n}\sum_{i=1}^n \begin{bmatrix} \boldsymbol{X}^{(i)}\boldsymbol{X}^{(i)\top} & -\boldsymbol{X}^{(i)} \\ -\boldsymbol{X}^{(i)\top} & 1 \end{bmatrix} \qquad \in \mathbb{R}^{(d+1)\times(d+1)}, \\ \boldsymbol{g}_{1,j} &\equiv \mathbf{e}_{j,d} \qquad \in \mathbb{R}^d, \\ g_{2,j} &\equiv 0 \qquad \in \mathbb{R}. \end{aligned}$$

$\mathbf{e}_{j,d} \in \mathbb{R}^d$ and $\boldsymbol{R}$ are similarly defined as in corollary 5.4.3.

Remark : [7] shows more generalized case of Cororally 5.4.3, in which $\boldsymbol{X}$ is distributed as pairwise interaction power model. And, when we set $h \equiv \mathbb{1}_d$, the corporally 5.4.3 coincides with (5.14).

Error decomposition in Gaussian case Now we go consider the decomposition of $\left|\hat{\theta}_{\mathrm{SM}} - \theta_0\right|$ in case of $h \equiv \mathbb{1}_d$. Remember that from the the error $\left|\hat{\theta}_{\mathrm{SM}} - \theta_0\right|$ can be upper bounded by the terms which include $\|\Gamma(X) - \Gamma\|$ and $|g(X) - g|$.

$$
\begin{aligned}
\|\Gamma(X) - \Gamma\| &= \left\| R^\top \begin{pmatrix} \Gamma_0(X) - \Gamma_0 & & 0 \\ & \ddots & \\ 0 & & \Gamma_0(X) - \Gamma_0 \end{pmatrix} R \right\| \quad \text{here } \Gamma_0(X) \equiv \frac{1}{n}\sum_{i=1}^n \begin{bmatrix} X^{(i)}X^{(i)T} & -X^{(i)} \\ -X^{(i)T} & 1 \end{bmatrix},\ \Gamma_0 \equiv \mathbb{E}\begin{bmatrix} XX^T & -X \\ -X^T & 1 \end{bmatrix} \\
&\leq \|R\|^2 \left\| \begin{pmatrix} \Gamma_0(X) - \Gamma_0 & & 0 \\ & \ddots & \\ 0 & & \Gamma_0(X) - \Gamma_0 \end{pmatrix} \right\| \\
&= \|R\|^2 \|\Gamma_0(X) - \Gamma_0\| \\
&\leq \|R\|^2 \left\{ \underbrace{\left\| \frac{1}{n}\sum_{i=1}^n X^{(i)}X^{(i)\top} - \mathbb{E}(XX^\top) \right\|}_{\equiv a} + 2 \underbrace{\left| \frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}(X) \right|}_{\equiv b} + 1 \right\}
\end{aligned}
$$

$$
\begin{aligned}
a &= \left\| \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})(X_i - \overline{X})^\top - \mathbb{E}(X-\mu)(X-\mu)^\top + \overline{X}\overline{X}^\top - \mu\mu^\top \right\| \\
&\leq \left\| \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})(X_i - \overline{X})^\top - \mathbb{E}(X-\mu)(X-\mu)^\top \right\| + \left\| \overline{X}\overline{X}^\top - \mu\mu^\top \right\| \\
&= \left\| \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})(X_i - \overline{X})^\top - \mathbb{E}(X-\mu)(X-\mu)^\top \right\| + \left\| (\overline{X}-\mu)(\overline{X}-\mu)^\top + (\overline{X}-\mu)\mu^\top + \mu(\overline{X}-\mu)^\top \right\| \\
&= \underbrace{\left\| \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})(X_i - \overline{X})^\top - \mathbb{E}(X-\mu)(X-\mu)^\top \right\|}_{\equiv c} + \underbrace{\left|\overline{X}-\mu\right|^2}_{=b^2} + 2\underbrace{\left|\overline{X}-\mu\right|}_{=b}|\mu|
\end{aligned}
$$

From Theorem 3.1.25, with probability at least $1-\delta$, $b \leq \frac{2}{n}\mathrm{Tr}(\Sigma)\log\frac{2}{\delta} + \sqrt{\frac{2}{n}\mathrm{Tr}(\Sigma)\log\frac{2}{\delta}}$. From Theorem 3.2.5, with probability at least $1-\delta$, $c \leq C\|\Sigma\|\left(\sqrt{\frac{d}{n}} \vee \frac{d}{n} \vee \sqrt{\frac{-\log\delta}{n}} \vee \frac{-\log\delta}{n}\right)$. Replacing δ to $\frac{\delta}{2}$ and taking the union bound, we obtain

$$
\begin{aligned}
\|\Gamma(X) - \Gamma\| &\leq \|R\|^2 \left\{ c + b^2 + 2b(|\mu| + 1) + 1 \right\} \\
&\leq \|R\|^2 \left\{ C\|\Sigma\| \left(\sqrt{\frac{d}{n}} \vee \frac{d}{n} \vee \sqrt{\frac{-\log\delta/2}{n}} \vee \frac{-\log\delta/2}{n} \right) \right. \\
&\qquad \left. + \left(\frac{2}{n}\mathrm{Tr}(\Sigma)\log\frac{4}{\delta} + \sqrt{\frac{2}{n}\mathrm{Tr}(\Sigma)\log\frac{4}{\delta}} \right)^2 + 2\left(\frac{2}{n}\mathrm{Tr}(\Sigma)\log\frac{4}{\delta} + \sqrt{\frac{2}{n}\mathrm{Tr}(\Sigma)\log\frac{4}{\delta}} \right)(|\mu| + 1) + 1 \right\}
\end{aligned}
$$

with probability at least $1-\delta$. Since $|\boldsymbol{g}(X)-\boldsymbol{g}|=0$ in case that $\boldsymbol{X}$ is distributed as Gaussian and h is identity function,

$$\left|\hat{\boldsymbol{\theta}}_{\mathrm{SM}}-\boldsymbol{\theta}_0\right| \le \left\|\Gamma^{-1}(X)\right\| |\boldsymbol{\theta}_0| \, \|R\|^2 \left\{ C\|\Sigma\| \left(\sqrt{\frac{d}{n}} \bigvee \frac{d}{n} \bigvee \sqrt{\frac{-\log\delta/2}{n}} \bigvee \frac{-\log\delta/2}{n} \right) \right.$$
$$\left. + \left(\frac{2}{n}\mathrm{Tr}(\Sigma)\log\frac{4}{\delta} + \sqrt{\frac{2}{n}\mathrm{Tr}(\Sigma)\log\frac{4}{\delta}} \right)^2 + 2\left(\frac{2}{n}\mathrm{Tr}(\Sigma)\log\frac{4}{\delta} + \sqrt{\frac{2}{n}\mathrm{Tr}(\Sigma)\log\frac{4}{\delta}} \right)(|\boldsymbol{\mu}|+1)+1 \right\}.$$

5.4.3 Robust score matching estimator in Gaussian case

Our aim is to obtain robust version of $\hat{\boldsymbol{\theta}}_{\mathrm{SM}}$. For this aim, we consider to robustify $\boldsymbol{\Gamma}_0(X)$ in **??**. consists of empirical mean and empirical covariance, we replace these to robust ones. More precisely, we replace

$$\boldsymbol{\Gamma}_0(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} \boldsymbol{X}^{(i)}\boldsymbol{X}^{(i)\top} & -\boldsymbol{X}^{(i)} \\ -\boldsymbol{X}^{(i)\top} & 1 \end{bmatrix} \qquad \in \mathbb{R}^{(d+1)\times(d+1)}$$

to

$$\hat{\boldsymbol{\Gamma}}_0(\boldsymbol{X}) \equiv \begin{bmatrix} \hat{\Sigma}(\boldsymbol{X}^{(i)}) + \hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})\hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})^\top & -\hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)}) \\ -\hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})^\top & 1 \end{bmatrix} \qquad \in \mathbb{R}^{(d+1)\times(d+1)}. \tag{5.18}$$

Here $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ are robust estimator for mean and covariance such as those we listed up in the previous chapter. Note that this robustification procedure can be also applied to general exponential families in case that we have robust estimators for $\mathbb{E}_{p_{\boldsymbol{\theta}_0}}\boldsymbol{t}_j'(X)\boldsymbol{t}_j'(X)^\top = \mathrm{Cov}_{p_{\boldsymbol{\theta}_0}}(\boldsymbol{t}_j'(X)) + \mathbb{E}_{p_{\boldsymbol{\theta}_0}}(\boldsymbol{t}_j'(X))\mathbb{E}_{p_{\boldsymbol{\theta}_0}}(\boldsymbol{t}_j'(X))^\top$, $\mathbb{E}_{p_{\boldsymbol{\theta}_0}} b_j'(X)\boldsymbol{t}_j'(X)$, and $\mathbb{E}_{p_{\boldsymbol{\theta}_0}}\boldsymbol{t}_j''(X)$ for all $j \in [d]$ by replacing the corresponding empirical means and covariance matrix to them.

On the other hand, We can also obtain a robust estimator for $\boldsymbol{\mu}$ and Σ by simply replacing $\hat{\boldsymbol{\mu}}_{\mathrm{ML}}$ and $\hat{\Sigma}_{\mathrm{ML}}$ to a robust mean and robust covariance. Robustifying process are different, but we obtain the same estimator, i.e. the following lemma holds.

Lemma 5.4.5 *Let X is distributed as $\mathcal{N}(\boldsymbol{\mu},\Sigma)$ a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_0,\Sigma_0)$ and $(\boldsymbol{X}_i)_{1\le i\le n}$ be independent copies of $\boldsymbol{X}$. Define*

$$\hat{\boldsymbol{\Gamma}}(\boldsymbol{X}) \equiv \boldsymbol{R}^\top \begin{pmatrix} \hat{\boldsymbol{\Gamma}}_0(\boldsymbol{X}) & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \hat{\boldsymbol{\Gamma}}_0(\boldsymbol{X}) \end{pmatrix} \boldsymbol{R}, \qquad \boldsymbol{g}(\boldsymbol{X}) \equiv \boldsymbol{R}^\top \begin{pmatrix} \boldsymbol{g}_{1,1}(\boldsymbol{X}) \\ g_{2,1}(\boldsymbol{X}) \\ \vdots \\ \boldsymbol{g}_{1,d}(\boldsymbol{X}) \\ g_{2,d}(\boldsymbol{X}) \end{pmatrix}, \tag{5.19}$$

$$\begin{aligned} \textit{here} \qquad \hat{\boldsymbol{\Gamma}}_0(\boldsymbol{X}) &\equiv \begin{bmatrix} \hat{\Sigma}(\boldsymbol{X}^{(i)}) + \hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})\hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})^\top & -\hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)}) \\ -\hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})^\top & 1 \end{bmatrix} \qquad \in \mathbb{R}^{(d+1)\times(d+1)}, \\ \boldsymbol{g}_{1,j} &\equiv \mathbf{e}_{j,d} \qquad \in \mathbb{R}^d, \\ g_{2,j} &\equiv 0 \qquad \in \mathbb{R}. \end{aligned}$$

$\mathbf{e}_{j,d} \in \mathbb{R}^d$ and $\boldsymbol{R}$ are similarly defined as in Corollary 5.4.3. $\hat{\boldsymbol{\mu}}(\boldsymbol{X})$ and $\hat{\Sigma}(\boldsymbol{X})$ are estimators for the mean and the covariance. Suppose $\hat{\boldsymbol{\Gamma}}(\boldsymbol{X})$ is positive definite. Then,

$$\boldsymbol{R}\left(\operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \left(\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Gamma}(\boldsymbol{X})\boldsymbol{\theta} - \boldsymbol{g}(\boldsymbol{X})^\top\boldsymbol{\theta} \right) \right) = \left(\hat{\Sigma}(\boldsymbol{X})^{-1}\hat{\boldsymbol{\mu}}(\boldsymbol{X}) \, , \, \hat{\Sigma}(\boldsymbol{X})^{-1} \right),$$

here $\boldsymbol{\theta} \equiv (\omega_{11},\dots,\omega_{1d},\eta_1 \, , \, \omega_{22},\dots,\omega_{2d},\eta_2 \, , \, \dots \, , \, \omega_{dd},\eta_d)^\top \quad \in \quad \mathbb{R}^{\frac{d(d+1)}{2}+d}$.

Proof.
Since $\hat{\boldsymbol{\Gamma}}(\boldsymbol{X})$ is positive definite, $J_h(\boldsymbol{\theta}) \equiv \frac{1}{2}\boldsymbol{\theta}^\top \hat{\boldsymbol{\Gamma}}(\boldsymbol{X})\boldsymbol{\theta} - \boldsymbol{g}(\boldsymbol{X})^\top \boldsymbol{\theta}$ is strictly convex function with respect to $\boldsymbol{\theta}$. So, it is enough to show $\frac{\partial J_h}{\partial \boldsymbol{\theta}} = 0$ at $\boldsymbol{\theta} = \hat{\boldsymbol{\Sigma}}(\boldsymbol{X})^{-1}\hat{\boldsymbol{\mu}}(\boldsymbol{X})$.

$$
\begin{aligned}
\frac{\partial J_h}{\partial \boldsymbol{\theta}} = 0 \quad &\Leftrightarrow \quad \frac{\partial J_h}{\partial (\boldsymbol{R}\boldsymbol{\theta})}\frac{\partial (\boldsymbol{R}\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial J_h}{\partial (\boldsymbol{R}\boldsymbol{\theta})}\boldsymbol{R} = 0 \\
&\Leftrightarrow \quad \frac{\partial J_h}{\partial (\boldsymbol{R}\boldsymbol{\theta})} = 0 \qquad\qquad\qquad \textit{since } \mathrm{kernel}(\boldsymbol{R}) = 0 \\
&\Leftrightarrow \quad (\boldsymbol{R}\boldsymbol{\theta})^\top \hat{\boldsymbol{\Gamma}}(\boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{X})^\top = 0 \\
&\Leftrightarrow \quad \hat{\boldsymbol{\Gamma}}(\boldsymbol{X})(\boldsymbol{R}\boldsymbol{\theta}) = \boldsymbol{g}(\boldsymbol{X}) \\
&\Leftrightarrow \quad \begin{pmatrix}
\left(\hat{\Sigma}(\boldsymbol{X}^{(i)}) + \hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})\hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})^\top\right)\begin{pmatrix}\omega_{11}\\ \vdots \\ \omega_{1d}\end{pmatrix} - \hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})\eta_1 \\
\hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})^\top\begin{pmatrix}\omega_{11}\\ \vdots \\ \omega_{1d}\end{pmatrix} + \eta_1 \\
\vdots \\
\left(\hat{\Sigma}(\boldsymbol{X}^{(i)}) + \hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})\hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})^\top\right)\begin{pmatrix}\omega_{d1}\\ \vdots \\ \omega_{dd}\end{pmatrix} - \hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})\eta_d \\
\hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})^\top\begin{pmatrix}\omega_{d1}\\ \vdots \\ \omega_{dd}\end{pmatrix} + \eta_d
\end{pmatrix} = \begin{pmatrix}\mathbf{e}_{1,d}\\ 0 \\ \\ \vdots \\ \\ \mathbf{e}_{d,d}\\ 0\end{pmatrix} \qquad \in \mathbb{R}^{(d+1)\times d} \\
&\Leftrightarrow \quad \begin{pmatrix}
\left(\hat{\Sigma}(\boldsymbol{X}^{(i)}) + \hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})\hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})^\top\right)\begin{pmatrix}\omega_{11}\\ \vdots \\ \omega_{1d}\end{pmatrix} - \hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})\eta_1 \\
\vdots \\
\left(\hat{\Sigma}(\boldsymbol{X}^{(i)}) + \hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})\hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})^\top\right)\begin{pmatrix}\omega_{d1}\\ \vdots \\ \omega_{dd}\end{pmatrix} - \hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})\eta_d \\
-\hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})^\top\begin{pmatrix}\omega_{11}\\ \vdots \\ \omega_{1d}\end{pmatrix} + \eta_1 \\
\vdots \\
-\hat{\boldsymbol{\mu}}(\boldsymbol{X}^{(i)})^\top\begin{pmatrix}\omega_{d1}\\ \vdots \\ \omega_{dd}\end{pmatrix} + \eta_d
\end{pmatrix} = \begin{pmatrix}\mathbf{e}_{1,d}\\ \vdots \\ \mathbf{e}_{d,d}\\ 0 \\ \vdots \\ 0\end{pmatrix} \qquad \in \mathbb{R}^{(d+1)\times d}
\end{aligned}
$$

$$\Leftrightarrow \quad \left(\hat{\Sigma}(X^{(i)}) + \hat{\mu}(X^{(i)})\hat{\mu}(X^{(i)})^\top\right) \underbrace{\begin{pmatrix} \omega_{11} & \dots & \omega_{d1} \\ \vdots & \ddots & \vdots \\ \omega_{1d} & \dots & \omega_{dd} \end{pmatrix}}_{\Omega} - \hat{\mu}(X^{(i)}) \underbrace{\begin{pmatrix} \eta_1 \\ \vdots \\ \eta_d \end{pmatrix}^\top}_{\eta^\top} = I_d$$

$$\text{and} \quad -\hat{\mu}(X^{(i)})^\top \underbrace{\begin{pmatrix} \omega_{11} & \dots & \omega_{d1} \\ \vdots & \ddots & \vdots \\ \omega_{1d} & \dots & \omega_{dd} \end{pmatrix}}_{\Omega} + \underbrace{\begin{pmatrix} \eta_1 \\ \vdots \\ \eta_d \end{pmatrix}^\top}_{\eta^\top} = \mathbf{0}^\top$$

$(\eta, \Omega) = \left(\hat{\Sigma}(X)^{-1}\hat{\mu}(X)\ ,\ \hat{\Sigma}(X)^{-1}\right)$ satisfies the both equations. So, this pair minimizes J_h

□

Note that this lemma holds under the condition that the distribution comes from Gaussian and $h \equiv 1$.

Error bound of robust score matching estimator Similar to non-robust case, the error bound is bounded by the error of robust mean and robust covariance. For example, when we replace sample mean to Median of means and sample covariance to Minsker estimator, under that each parameters satisfies the assumption in the table in section 4.1 and (∗∗) in Theorem 4.2.6,

$$\left|\hat{\theta}_{\mathrm{SM}} - \theta_0\right| \le \left\|\Gamma^{-1}(X)\right\| |\theta_0| \, \|R\|^2 \left\{ C\|\Sigma\| \left(\frac{20}{39}\lambda_1 + \frac{80}{39}\sigma\sqrt{\frac{\log\left(\frac{8r_H+3}{3\delta}\right)}{n}} + \frac{40}{39}\lambda_2 \log\left(\frac{8r_H+3}{3\delta}\right) \right) \right.$$
$$\left. + \left(\frac{2}{n}\mathrm{Tr}(\Sigma)\log\frac{4}{\delta} + \sqrt{\frac{2}{n}\mathrm{Tr}(\Sigma)\log\frac{4}{\delta}} \right)^2 + 2\left(\sqrt{32\mathrm{Tr}(\Sigma)\frac{\log(d/\delta)}{n}} \right)(|\mu| + 1) + 1 \right\}.$$

5.5 Simulation in Gaussian case

5.5.1 Non-robust estimator

In the simulation, we consider the case in which the the population distribution X is distributed as $\mathcal{N}_3(\mu_0, \Sigma_0)$, where

$$\mu = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \ \Sigma = \begin{pmatrix} 1 & 0 & 0.5 \\ 0 & 2 & 1 \\ 0.5 & 1 & 1 \end{pmatrix},$$

and we set $h_j \equiv 1$ for $j \in [3]$. Suppose that we don't know both μ_0 amd Σ_0, which we estimate by using sample data. We compare the score matching estimators in case that the sample size $n = 10, n = 100$, and $n = 1000$. For each case, we estimate 100 times.

From Corollary 5.4.4, the score matching estimator is given by

$$\Gamma(X) = R^\top \begin{pmatrix} \Gamma_0(X) & & 0 \\ & \ddots & \\ 0 & & \Gamma_0(X) \end{pmatrix} R, \qquad g(X) = R^\top \begin{pmatrix} g_{1,1}(X) \\ g_{2,1}(X) \\ \vdots \\ g_{1,d}(X) \\ g_{2,d}(X) \end{pmatrix}. \tag{5.20}$$

Here

$$\begin{aligned}
\Gamma_0(X) &\equiv \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} \equiv \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} X^{(i)}X^{(i)\top} & -X^{(i)} \\ -X^{(i)\top} & 1 \end{bmatrix} \qquad \in \mathbb{R}^{(d+1)\times(d+1)}, \\
g_{1,j} &\equiv \mathbf{e}_{j,d} \qquad \in \mathbb{R}^d, \\
g_{2,j} &\equiv 0 \qquad \in \mathbb{R}.
\end{aligned}$$

Thus, we can estimate $\boldsymbol{\mu}$ and Σ by the following procedure.

Procedure 1 Score maching method in Gaussian distribution (100 times trials)

1: INPUT: 100 $\times$ n samples $(\boldsymbol{x}_{\ell:i})_{\ell\in[100]}, i \in [n]$ from $\mathcal{N}_3(\boldsymbol{\mu}_0, \Sigma_0)$
2: OUTPUT: Estimated means $(\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell})$ and covariances $(\hat{\Sigma}_{\mathrm{SM}_\ell})_{\ell\in[100]}$ by score mathing method
3: FOR $\ell = 1$ *to* 100
Calculate $\Gamma(\boldsymbol{x}_\ell)$ and $\boldsymbol{g}(\boldsymbol{x}_\ell)$ in Corollary 5.19
$\hat{\boldsymbol{\theta}}_{\mathrm{SM}_\ell} = \Gamma(\boldsymbol{x}_\ell)^{-1}\boldsymbol{g}(\boldsymbol{x}_\ell)$
Find $\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell}$ and $\hat{\Sigma}_{\mathrm{SM}_\ell}$ corresponding to $\hat{\boldsymbol{\theta}}_{\mathrm{SM}_\ell}$
Calculate $\hat{\Omega}_{\mathrm{SM}_\ell} = \hat{\Sigma}^{-1}_{\mathrm{SM}_\ell}$

The following is the result. As we proved in Section 3.2, when $h_j(x) = i\ \forall j$, $\theta_{SM} = \theta_{ML}$. So, we obrtain exactly the same result from the following procedure.

Procedure 2 Maximum likelihood estimation method for Gaussian Graphical models (100 times trials)

1: INPUT: 100 $\times$ n samples $(\boldsymbol{x}_{\ell:i})_{\ell\in[100],i\in[n]}$ from $\mathcal{N}_3(\boldsymbol{\mu}_0, \Sigma_0)$
2: OUTPUT: Estimated means $(\hat{\boldsymbol{\mu}}_{\mathrm{ML}_\ell})_{\ell\in[100]}$ and covariances $(\hat{\Sigma}_{\mathrm{ML}_\ell})_{\ell\in[100]}$ by maximum likelihood method
3: FOR $\ell = 1$ *to* 100
$\hat{\boldsymbol{\mu}}_{\mathrm{ML}_\ell} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_{\ell:i}$
$\hat{\Sigma}_{\mathrm{ML}_\ell} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_{\ell:i} - \hat{\boldsymbol{\mu}}_{\mathrm{ML}_\ell})(\boldsymbol{x}_{\ell:i} - \hat{\boldsymbol{\mu}}_{\mathrm{ML}_\ell})^\top$
Calculate $\hat{\Omega}_{\mathrm{ML}_\ell} = \hat{\Sigma}^{-1}_{\mathrm{ML}_\ell}$

L^2 error For the evaluation of the errors of $\hat{\boldsymbol{\mu}}_{\mathrm{SM}} - \boldsymbol{\mu}_0$ and $\hat{\Sigma}_{\mathrm{SM}} - \Sigma_0$, we use L^2 error. The following is the algorithm to calculate it.

Procedure 3 Measuring the L^2 errors

1: INPUT: $(\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell})_{\ell\in[100]}$ and $(\hat{\Sigma}_{\mathrm{SM}_\ell})_{\ell\in[100]}$ obtained by Procedure 1
2: OUTPUT: $L^2(\hat{\boldsymbol{\mu}}_{\mathrm{SM}} - \boldsymbol{\mu}_0)$, $L^2(\hat{\Sigma}_{\mathrm{SM}} - \Sigma_0)$, and $L^2(\hat{\Omega}_{\mathrm{SM}} - \Omega_0)$
3: FOR $\ell = 1$ to 100, calculate

$$\left|\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell} - \boldsymbol{\mu}_0\right|_2 \equiv \left(\sum_{i=1}^{3}\left|\hat{\mu}^{(i)}_{\mathrm{SM}_\ell} - \mu^{(i)}\right|^2\right)^{\frac{1}{2}} \quad \left\|\hat{\Sigma}_{\mathrm{SM}_\ell} - \Sigma_0\right\|_2 \equiv \left(\sum_{i,j=1}^{3}\left|\hat{\sigma}^{(ij)}_{\mathrm{SM}_\ell} - \sigma_0^{(ij)}\right|^2\right)^{\frac{1}{2}} \quad \left\|\hat{\Omega}_{\mathrm{SM}_\ell} - \Omega_0\right\|_2 \equiv \left(\sum_{i,j=1}^{3}\left|\hat{\omega}^{(ij)}_{\mathrm{SM}_\ell} - \omega_0^{(ij)}\right|^2\right)^{\frac{1}{2}}$$

4: For $k = 1$ to n, calculate

$$L_2(\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell} - \boldsymbol{\mu}_0) \equiv \frac{1}{100}\sum_{\ell=1}^{100}\left|\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell} - \boldsymbol{\mu}_0\right|_2 \quad L_2(\hat{\Sigma}_{\mathrm{SM}_\ell} - \Sigma_0) \equiv \frac{1}{100}\sum_{\ell=1}^{100}\left\|\hat{\Sigma}_{\mathrm{SM}_\ell} - \Sigma_0\right\|_2 \quad L_2(\hat{\Omega}_{\mathrm{SM}_\ell} - \Omega_0) \equiv \frac{1}{100}\sum_{\ell=1}^{100}\left\|\hat{\Omega}_{\mathrm{SM}_\ell} - \Omega_0\right\|_2$$

Result :

n	$L_2(\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell} - \boldsymbol{\mu}_0)$	$L_2(\hat{\Sigma}_{\mathrm{SM}_\ell} - \Sigma_0)$
100	0.17042293	0.4270580
300	0.09780322	0.2619216
1000	0.05625493	0.1384555

(5.21)

TPR and FPR In a covariance estimation, we are especially interested in a covariance is 0 or not. In our example, $(1,2)$ entry and $(2,1)$ entry of Σ_0 are 0. We decide that $\hat{\Sigma}_{\mathrm{SM}}^{(ij)} = 0$ when $\left|\hat{\Sigma}_{\mathrm{SM}}^{(ij)}\right| \leq \epsilon$. To compare the result from the perspective how correctly the estimator distinguish each parameters is zero or not, we use the following criteria.

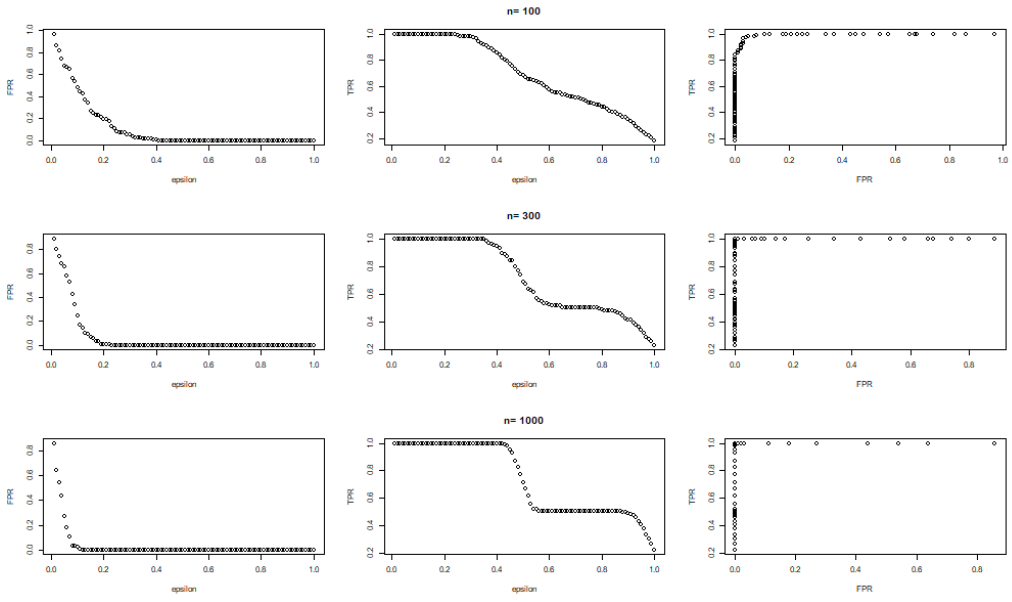
$$\begin{aligned}
\mathrm{FPR}(\hat{\Sigma}_{\mathrm{SM}}) &\equiv \frac{\#\text{ of non-diagonal entries not equal 0 in } \hat{\Sigma}_{\mathrm{SM}} \text{ but the decision is wrong}}{\#\text{ of non-diagonal entries equal 0 in } \Sigma_0} = \frac{\left|\hat{S}_{\mathrm{off}} \setminus S_{0,\mathrm{off}}\right|}{d(d-1) - \left|S_{0,\mathrm{off}}\right|} \\
\mathrm{TPR}(\hat{\Sigma}_{\mathrm{SM}}) &\equiv \frac{\#\text{ of non-diagonal entries not equal 0 in } \hat{\Sigma}_{\mathrm{SM}} \text{ and the decision is correct}}{\#\text{ of non-diagonal entries not equal 0 in } \Sigma_0} = \frac{\left|\hat{S}_{\mathrm{off}} \cap S_{0,\mathrm{off}}\right|}{\left|S_{0,\mathrm{off}}\right|}
\end{aligned}$$

where $S_{0,\mathrm{off}} \equiv \left\{(i,j) : i \neq j \wedge \sigma_{0,ij} \neq 0\right\}$, and $\hat{S}_{\mathrm{off}} \equiv \left\{(i,j) : i \neq j \text{ and } \left|\hat{\sigma}_{\mathrm{SM}}^{(ij)}\right| > \epsilon\right\}$. Since $\hat{\Sigma}_{\mathrm{SM}}$ depends on the sample, FPR and TPR fluctuate. So, we take the mean of them.
Analogically $\mathrm{TPR}(\hat{\Omega}_{\mathrm{SM}})$ and $\mathrm{FPR}(\hat{\Omega}_{\mathrm{SM}})$ are also defined.

Procedure 4 Calculate the FPR and TPR

1: INPUT: $(\hat{\Sigma}_{\mathrm{SM}_\ell})_{\ell \in [100]}$ obtained by Procedure 1, $\epsilon \in [0,1]$
2: OUTPUT: Mean of FPR and Mean of TPR
3: FOR $l = 1$ to 100
Calculate $\mathrm{FPR}(\hat{\Sigma}_{\mathrm{SM}_\ell})$ and $\mathrm{TPR}(\hat{\Sigma}_{\mathrm{SM}_\ell})$
4: Calculate Mean of FPD $\equiv \frac{1}{100}\sum_{\ell=1}^{100} \mathrm{FPR}(\hat{\Sigma}_{\mathrm{SM}_\ell})$ and Mean of TPR $\equiv \frac{1}{100}\sum_{\ell=1}^{100} \mathrm{TPR}(\hat{\Sigma}_{\mathrm{SM}_\ell})$

Result :



5.5.2 Robust estimator

The performance of the robust score matching estimator provided by (5.18) is the followings.

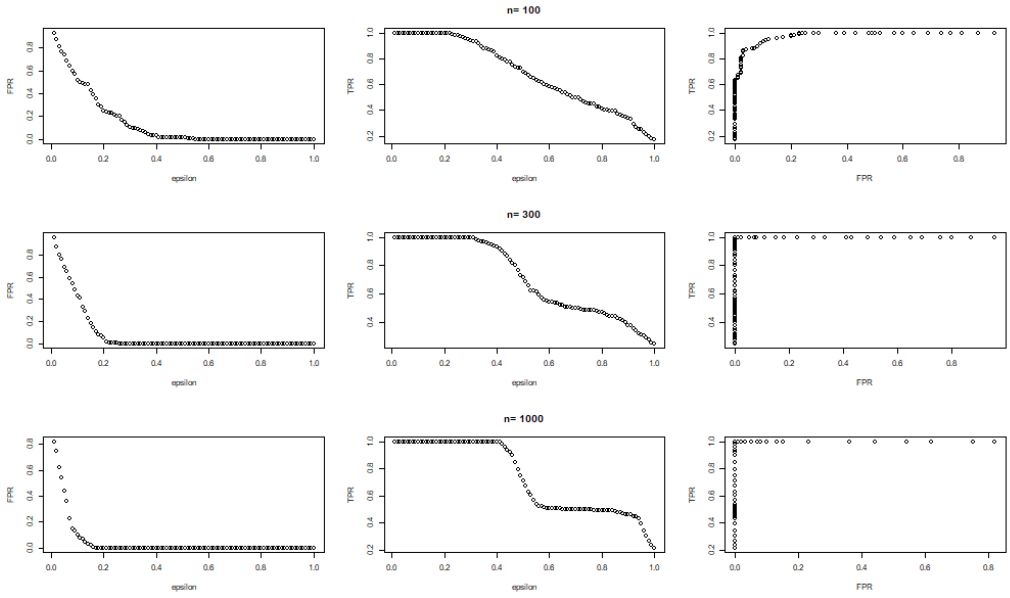
Median of Mean, Median absolute deviation

L^2 error

# of samples	$L_2(\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell} - \boldsymbol{\mu}_0)$	$L_2(\hat{\Sigma}_{\mathrm{SM}_\ell} - \Sigma_0)$
100	0.2008869	0.5691329
300	0.1156710	0.3521662
1000	0.06558571	0.1891582

(5.22)

TPR and FPR



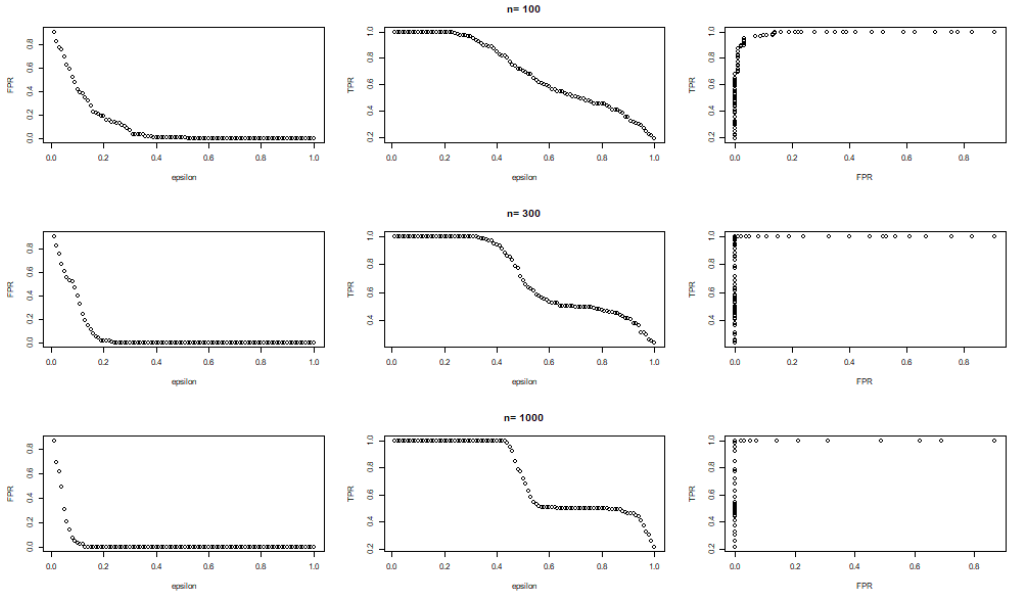
Median of Mean, Altenative MAD

L^2 error

# of samples	$L_2(\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell} - \boldsymbol{\mu}_0)$	$L_2(\hat{\Sigma}_{\mathrm{SM}_\ell} - \Sigma_0)$
100	0.2008869	0.468477
300	0.1156710	0.3012807
1000	0.06558571	0.1600284

(5.23)

TPR and FPR



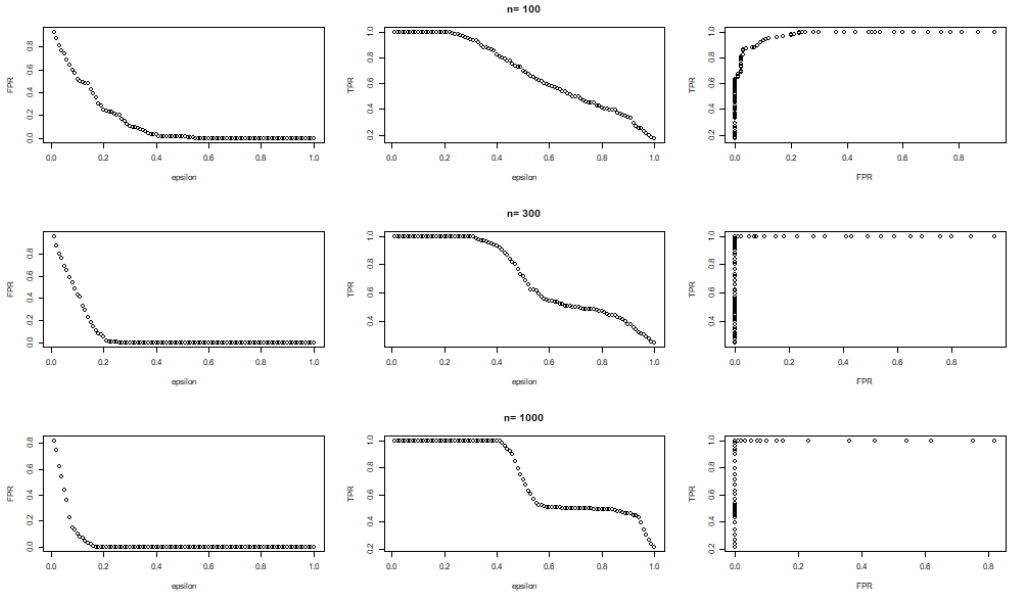
Trimmed Means, Median absolute deviation

L^2 error

# of samples	$L_2(\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell} - \boldsymbol{\mu}_0)$	$L_2(\hat{\Sigma}_{\mathrm{SM}_\ell} - \Sigma_0)$
100	0.1917967	0.5691329
300	0.1051795	0.3521662
1000	0.06180664	0.1891582

(5.24)

TPR and FPR



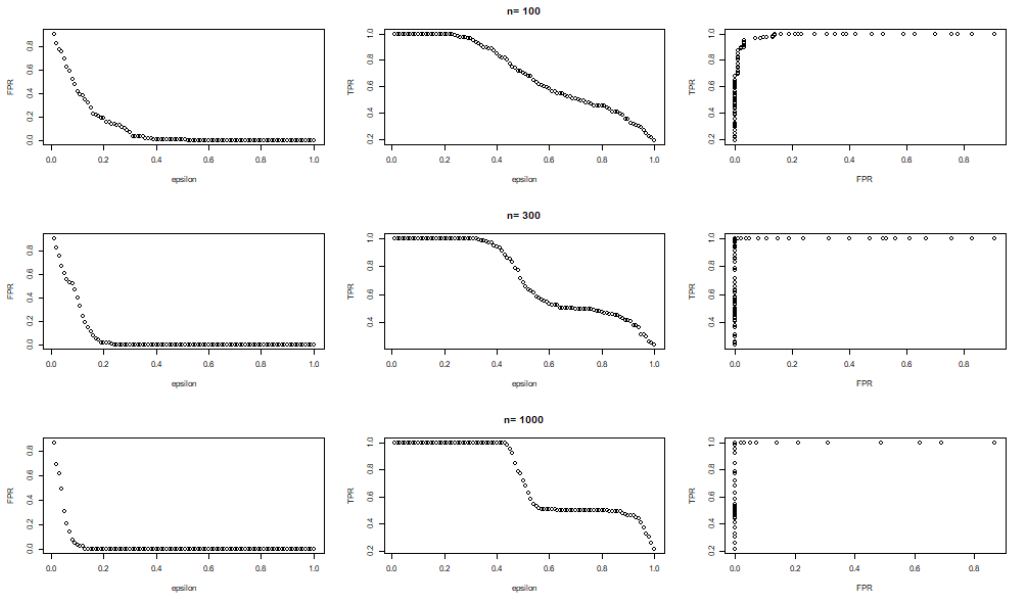
Trimmed Means, Altenative MAD

L^2 error

# of samples	$L_2(\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell} - \boldsymbol{\mu}_0)$	$L_2(\hat{\Sigma}_{\mathrm{SM}_\ell} - \Sigma_0)$
100	0.1917967	0.4684770
300	0.1051795	0.3012807
1000	0.06180664	0.1600284

(5.25)

TPR and FPR



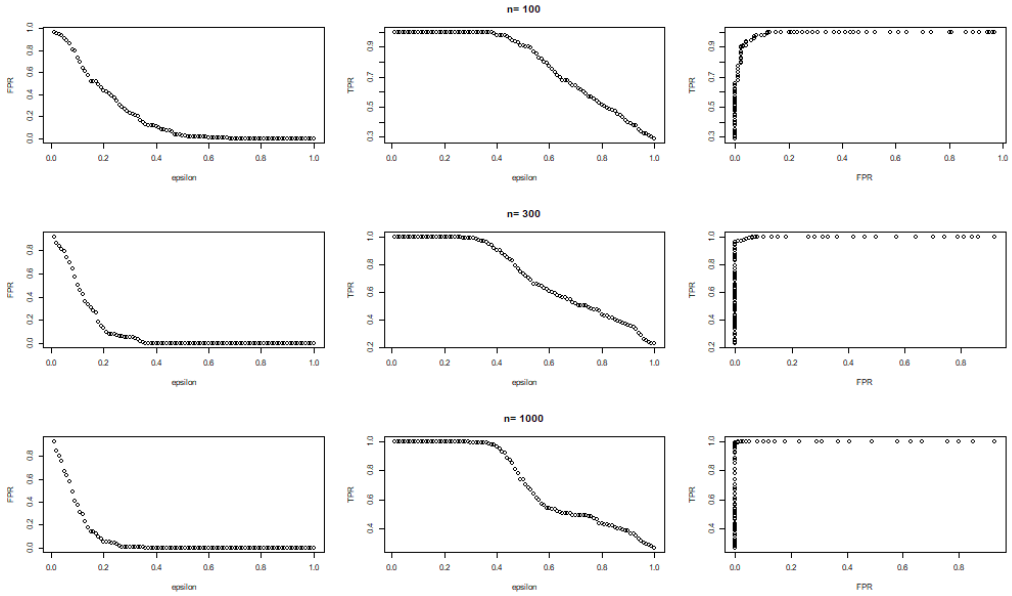
Medians of Means, Minsker's Method In Minsker's method, we set the number of iterations is 30, $\lambda_1 \equiv 0, \lambda_2 \equiv 0.1, B \equiv 10, \eta_t \equiv t^{-\frac{2}{3}}$

L^2 error

# of samples	$L_2(\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell} - \boldsymbol{\mu}_0)$	$L_2(\hat{\Sigma}_{\mathrm{SM}_\ell} - \Sigma_0)$
100	0.2008869	0.5669076
300	0.1156710	0.4381501
1000	0.06558571	0.3836059

(5.26)

TPR and FPR



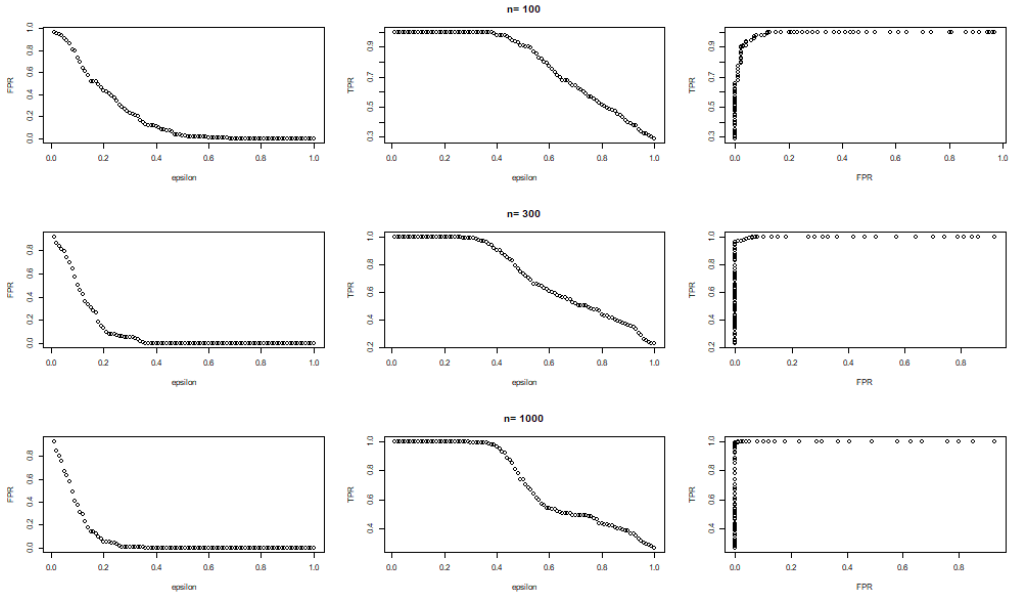
Trimmed Means, Minsker's Method In Minsker's method, we set the number of iterations is 30, $\lambda_1 \equiv 0$, $\lambda_2 \equiv 0.1$, $B \equiv 10$, $\eta_t \equiv t^{-\frac{2}{3}}$

L^2 error

# of samples	$L_2(\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell} - \boldsymbol{\mu}_0)$	$L_2(\hat{\Sigma}_{\mathrm{SM}_\ell} - \Sigma_0)$
100	0.19179671	0.5669076
300	0.10517949	0.4381501
1000	0.06180664	0.3836059

(5.27)

TPR and FPR



Next we consider the model which have contamination.

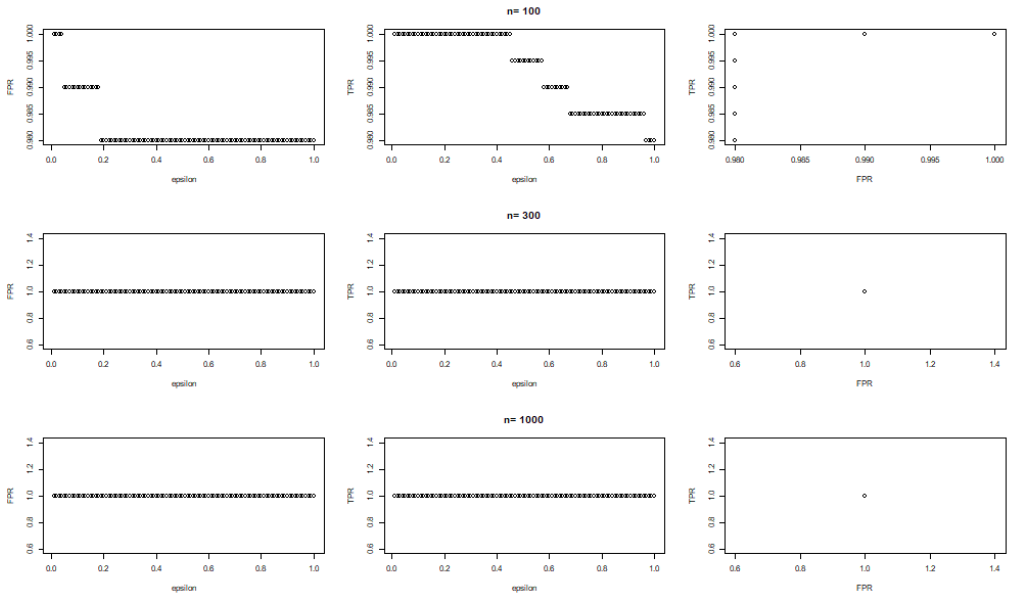
Non-robust score matching estimator

L^2 error

# of samples	$L_2(\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell} - \boldsymbol{\mu}_0)$	$L_2(\hat{\Sigma}_{\mathrm{SM}_\ell} - \Sigma_0)$
100	4.249416	345.7046
300	4.375587	358.147
1000	4.299185	353.4891

(5.28)

TPR and FPR



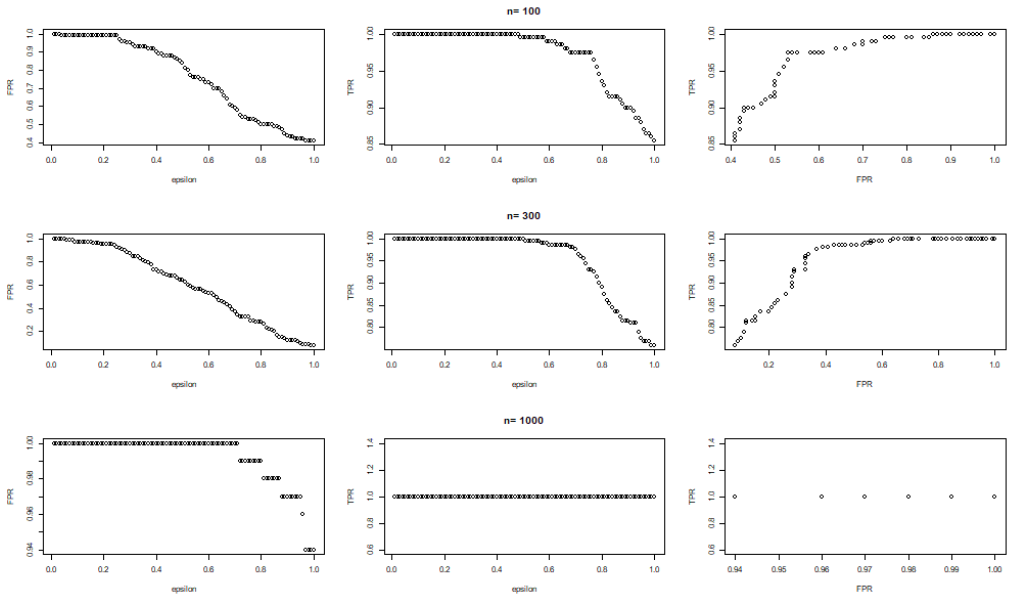
Median of Means, Minsker's Method in contamination model In Minsker's method, we set the number of iterations is 30, $\lambda_1 \equiv 0$, $\lambda_2 \equiv 0.1$, $B \equiv 10, \eta_t \equiv t^{-\frac{2}{3}}$. The distribution is contaminated with 5 percent of outlier $(50, 50, 50)$.

L^2 error

# of samples	$L_2(\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell} - \boldsymbol{\mu}_0)$	$L_2(\hat{\Sigma}_{\mathrm{SM}_\ell} - \Sigma_0)$
100	2.676537	2.837128
300	4.096872	1.615904
1000	4.214861	4.938987

(5.29)

TPR and FPR



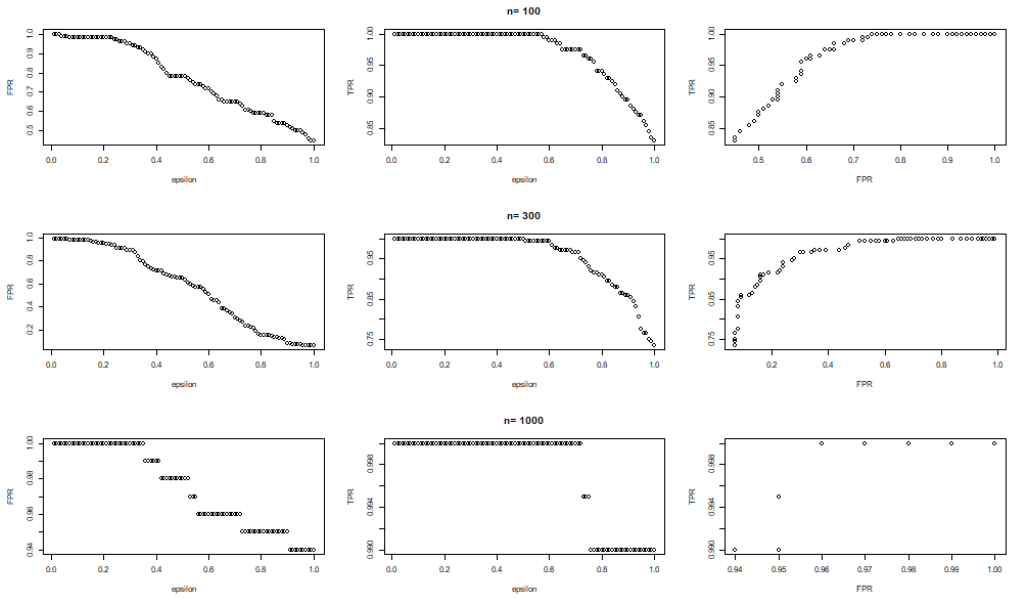
Trimmed Means, Minsker's Method in contamination model In Minsker's method, we set the number of iterations is 30, $\lambda_1 \equiv 0$, $\lambda_2 \equiv 0.1$, $B \equiv 10, \eta_t \equiv t^{-\frac{2}{3}}$. The distribution is contaminated with 5 percent of outlier $(50, 50, 50)$.

L^2 error

# of samples	$L_2(\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell} - \boldsymbol{\mu}_0)$	$L_2(\hat{\Sigma}_{\mathrm{SM}_\ell} - \Sigma_0)$
100	0.30125	2.837128
300	0.2033159	1.615904
1000	0.1751119	4.938987

(5.30)

TPR and FPR



Comparing with Non-robust estimator, we can see that the robust estimator is more efficient.

5.5.3 Estimation of Precision matrix

Although we have estimated the mean $\boldsymbol{\mu}$ and the covariance matrix Σ of Gaussian distribution until here, at the same time the estimating the precision matrix $\Omega \equiv \Sigma^{-1}$ is also our interest because in Gaussian case, each entries in the precision matrix is related to the statistical graph. We can estimate the Ω of the population distribution by taking the inverse of $\hat{\Sigma}$/
we consider the case in which the the population distribution $\boldsymbol{X}$ is distributed as $\mathcal{N}_3(\mu_0, \Sigma_0)$, where

$$\mu = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \ \Omega = \begin{pmatrix} 1 & 0 & 0.5 \\ 0 & 2 & 1 \\ 0.5 & 1 & 1 \end{pmatrix}, \quad \left(\text{then } \Sigma = \Omega^{-1} = \begin{pmatrix} 2 & 1 & -2 \\ 1 & 1.5 & -2 \\ -2 & -2 & 4 \end{pmatrix} \right)$$

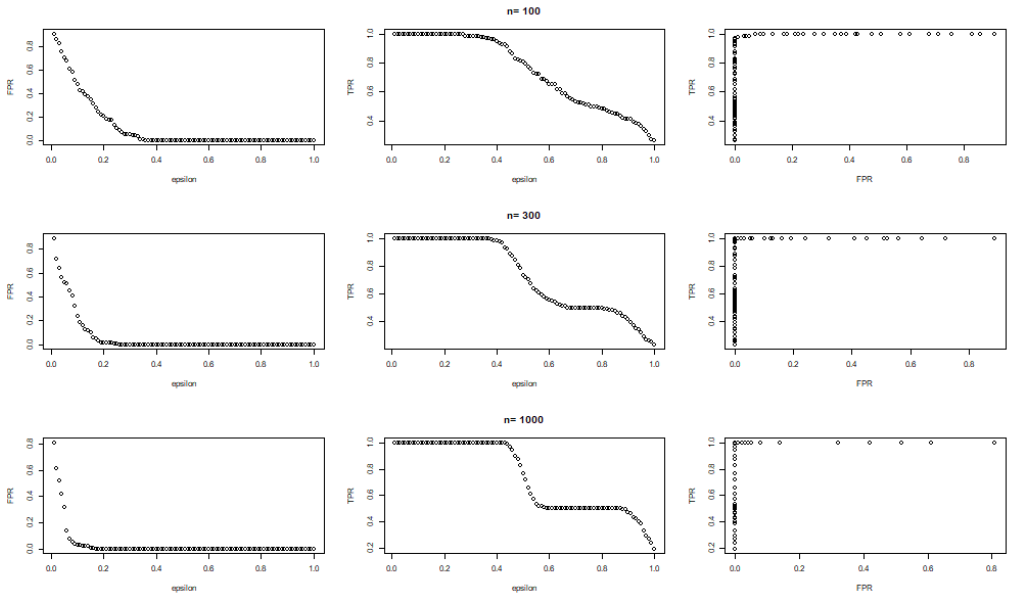
and we set $h_j \equiv 1$ for $j \in [3]$. The followings are the result of this estimation

Non-robust case In Minsker's method, we set the number of iterations is 30, $\lambda_1 \equiv 0$, $\lambda_2 \equiv 0.1$, $B \equiv 10, \eta_t \equiv t^{-\frac{2}{3}}$

L^2 error

# of samples	$L_2(\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell} - \boldsymbol{\mu}_0)$	$L_2(\hat{\Omega}_{\mathrm{SM}_\ell} - \Omega_0)$
100	0.2291308	0.5343700
300	0.1286565	0.2836951
1000	0.0769031	0.1544092

TPR and FPR

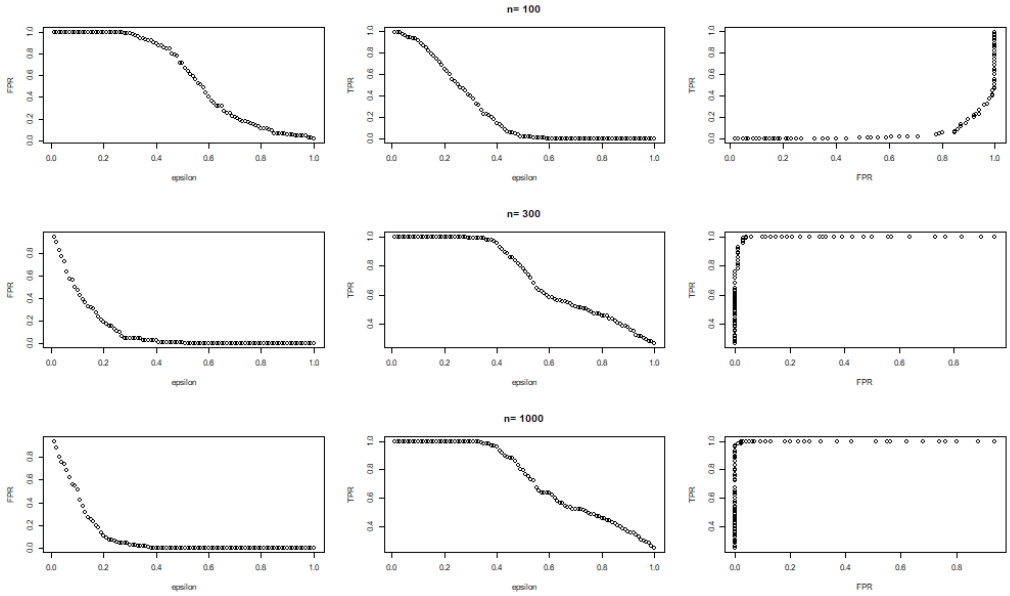


Robust case : Medians of Means, Minsker's Method In Minsker's method, we set the number of iterations is 30, $\lambda_1 \equiv 0$, $\lambda_2 \equiv 0.1$, $B \equiv 10$, $\eta_t \equiv t^{-\frac{2}{3}}$

L^2 error

# of samples	$L_2(\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell} - \boldsymbol{\mu}_0)$	$L_2(\hat{\Omega}_{\mathrm{SM}_\ell} - \Omega_0)$
100	0.27581564	1.5124890
300	0.15455185	0.4772902
1000	0.08734777	0.4015194

TPR and FPR

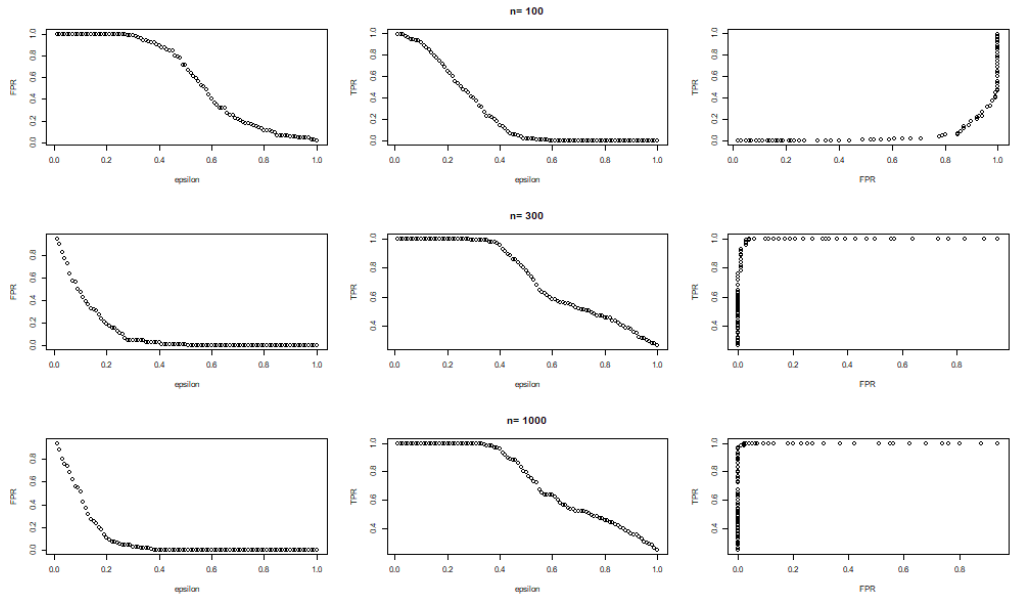


Robust case : Trimmed Mean, Minsker's Method In Minsker's method, we set the number of iterations is 30, $\lambda_1 \equiv 0$, $\lambda_2 \equiv 0.1$, $B \equiv 10$, $\eta_t \equiv t^{-\frac{2}{3}}$

L^2 error

# of samples	$L_2(\hat{\boldsymbol{\mu}}_{\mathrm{SM}_\ell} - \boldsymbol{\mu}_0)$	$L_2(\hat{\boldsymbol{\Omega}}_{\mathrm{SM}_\ell} - \boldsymbol{\Omega}_0)$
100	0.26785685	1.5124890
300	0.14696800	0.4772902
1000	0.08505598	0.4015194

TPR and FPR



5.6 Score matching estimator in Pareto case

Next, we consider non-Gaussian case. First, let's consider the maximum likelihood estimator The the density function of the Pareto distribution with the parameter $\boldsymbol{\theta} = (\alpha, \beta)$ is given by

$$p_{\alpha,\beta}(x; \alpha, \beta) = \alpha \frac{\beta^\alpha}{x^{\alpha+1}}, \quad x \geq \beta, \quad \alpha, \beta > 0. \tag{5.31}$$

α is called *scale parameter*, and β is *shape parameter*.

$$\log p_{\alpha,\beta}(x; \alpha, \beta) = \log(\alpha) + \alpha \log(\beta) - (\alpha + 1) \log(x).$$

Pareto distribution is an exponential family with the canonical parameter $-(\alpha+1)$. The log-likelihood function $\ell(\alpha,\beta;x)$ of the Pareto distribution given a sample $x=(x_1,\ldots,x_n)$ is given by

$$\begin{aligned}\log\prod_{i=1}^{n}p_{\alpha,\beta}(x_i;\alpha,\beta) &= \sum_{i=1}^{n}\log\left(\alpha\frac{\beta^\alpha}{x_i^{\alpha+1}}\right)\\ &= n\log(\alpha)+n\alpha\log(\beta)-(\alpha+1)\sum_{i=1}^{n}\log(x_i)\,. \end{aligned}\tag{5.32}$$

Let's consider the maximum likelihood estimator. At first, we consider the case case that β is known. For α, we can find the maximizer by solving set the partial derivative equal to 0 :

$$\frac{\partial\ell(\alpha,\beta)}{\partial\alpha}=\frac{n}{\alpha}+n\log(\beta)-\sum_{i=1}^{n}\log(x_i)=0.$$

Thus, we get the maximum likelihood estimator for Pareto distribution

$$\begin{aligned}\hat{\alpha}_{\mathrm{ML}} &= \arg\min_{\alpha} n\log(\alpha)+n\alpha\log(\beta)-(\alpha+1)\sum_{i=1}^{n}\log(x_i)\\ &= \frac{n}{\sum_{i=1}^{n}\log(x_i)-n\log(\beta)}\end{aligned}$$

In case that both α and β are unknown, the higher β, the higher (5.32), and $\beta\le x_i$ for all i. Under this restriction, the β maximizing $\ell(x;\alpha,\beta)$ is $\hat{\beta}=\min_i x_i$.

$$\begin{aligned}\hat{\beta}_{\mathrm{ML}} &= \min_i x_i.\\ \hat{\alpha}_{\mathrm{ML}} &= \frac{n}{\sum_{i=1}^{n}\log(x_i)-n\log(\hat{\beta})}\end{aligned}$$

Next, we consider the score matching estimator for Pareto distribution. Note that in order to apply Score matching method, we need to chose h which satisfies (A1) and (A2). In fact, tn Pareto case, $h(x)=1$ does not satisfy (A1) since $\lim_{x\to\beta}\log p_{\alpha,\beta}(x)\partial\log p_{\alpha,\beta}(x)\neq 0$ So, we need to consider a function h which satisfies $\lim_{x\to\beta}h(x)=0$.

$$\begin{aligned}\text{(A1)} &\Leftrightarrow \lim_{x\to\beta,+\infty}p_{\theta_0}(x)h(x)\partial\log p_{\theta}(\boldsymbol{x})=0\\ &\Leftrightarrow \lim_{x\to\beta,+\infty}\alpha\frac{\beta^\alpha}{x^{\alpha+1}}h(x)\frac{-(\alpha+1)}{x}=0\\ &\Leftrightarrow \lim_{x\to\beta,+\infty}\frac{h(x)}{x^{\alpha+2}}=0\end{aligned}$$

For example, $h(x) \equiv (x-\beta)^k$ for $0 < k \leq \alpha + 2$ and $h(x) \equiv (x-\beta)x^k$ for $k \leq \alpha + 1$ satisfies this condition. For such h, we can calculate $\boldsymbol{\Gamma}(X)$ and $\boldsymbol{g}(X)$ in Definition 5.2.4 as follows. Then

$$\begin{aligned}
\boldsymbol{\Gamma}(X) &= \frac{1}{n}\sum_{i=1}^{n} h(X_i)\boldsymbol{t}'(\boldsymbol{X}_i)\boldsymbol{t}'(\boldsymbol{X}_i)^\top \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{h(X_i)}{X_i^2}, \\
\boldsymbol{g}(X) &= -\frac{1}{n}\sum_{i=1}^{n} \left[h(X_i)b'(\boldsymbol{X}_i)\boldsymbol{t}'(\boldsymbol{X}_i) + h(X_i)(X_i)\boldsymbol{t}''(\boldsymbol{X}_i) + h'(X_i)\boldsymbol{t}'(\boldsymbol{X}_i)\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n} \left[-\frac{h(X_i)}{X_i^2} + \frac{h'(X_i)}{X_i}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \left[\frac{h(X_i)}{X_i^2} - \frac{h'(X_i)}{X_i}\right].
\end{aligned}$$

Then, the score matching estimator $\hat{\boldsymbol{\theta}}_{\mathrm{SM}}$ for $\theta = -(\alpha+1)$ is given by

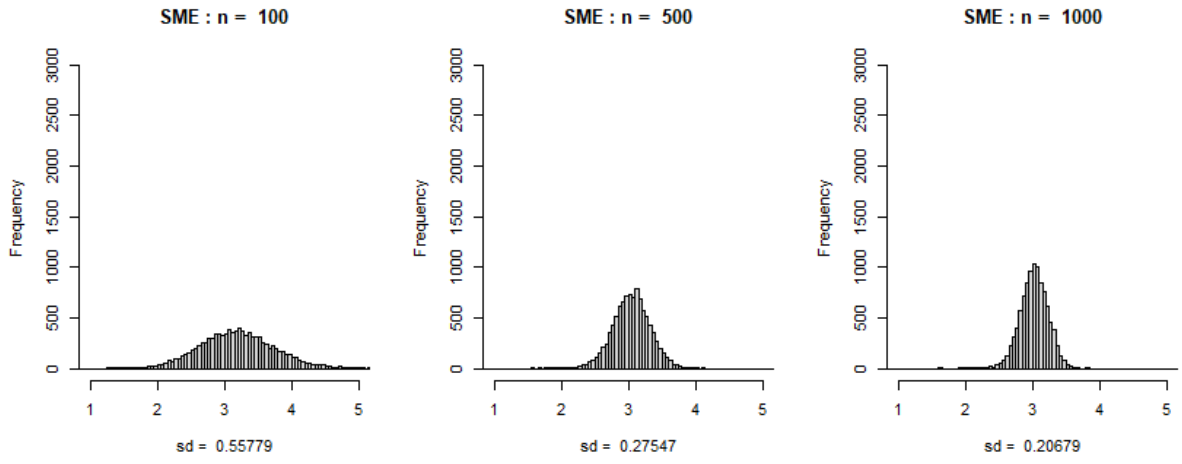
$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{\mathrm{SM}} &= \left(\frac{1}{n}\sum_{i=1}^{n} \frac{h(X_i)}{X_i^2}\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \left[\frac{h(X_i)}{X_i^2} - \frac{h'(X_i)}{X_i}\right] \\
&= 1 - \left(\sum_{i=1}^{n} \frac{h(X_i)}{X_i^2}\right)^{-1} \sum_{i=1}^{n} \frac{h'(X_i)}{X_i}.
\end{aligned}$$

Thus,

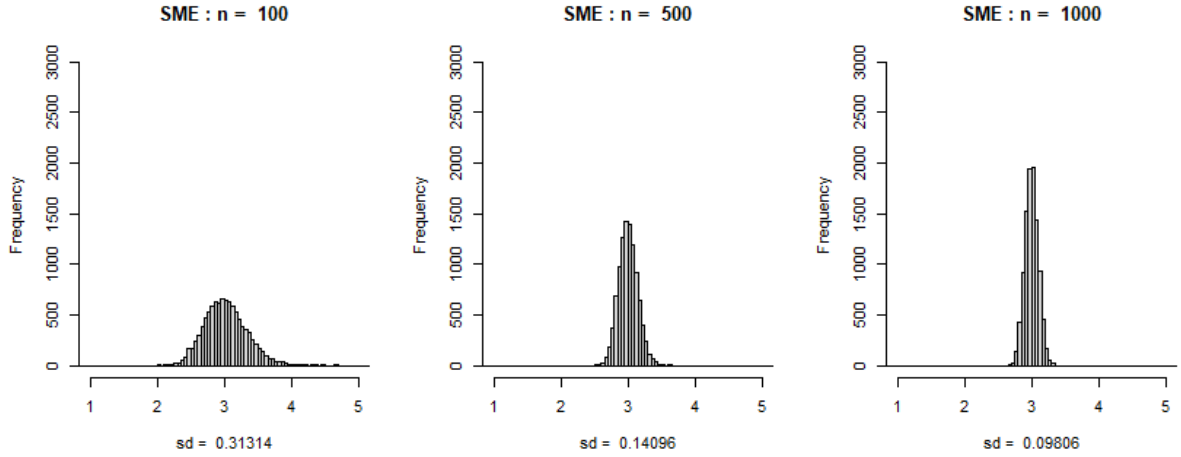
$$\hat{\alpha}_{\mathrm{SM}} = \left(\sum_{i=1}^{n} \frac{h(X_i)}{X_i^2}\right)^{-1} \sum_{i=1}^{n} \frac{h'(X_i)}{X_i} - 2. \tag{5.33}$$

The following are the result of each estimator. Here, the population distribution is Pareto distribution with $\alpha = 3, \beta = 1$. We know the β, but α is unknown. We estimate α with $100, 500, 1000$ samples. We calculate 10,000 times for each simulation, and the followings are the results.

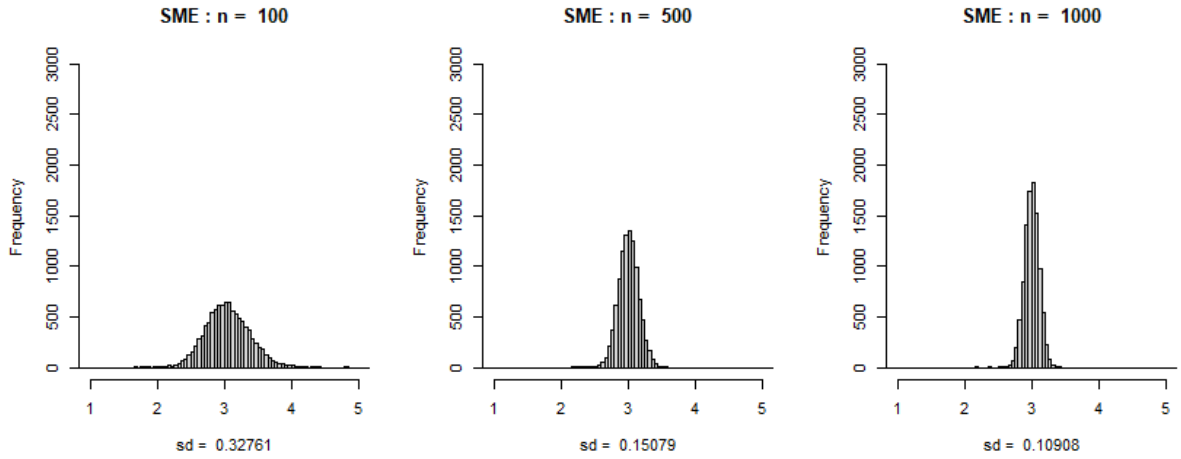
SME with $h(x) \equiv (x-\beta)^3$

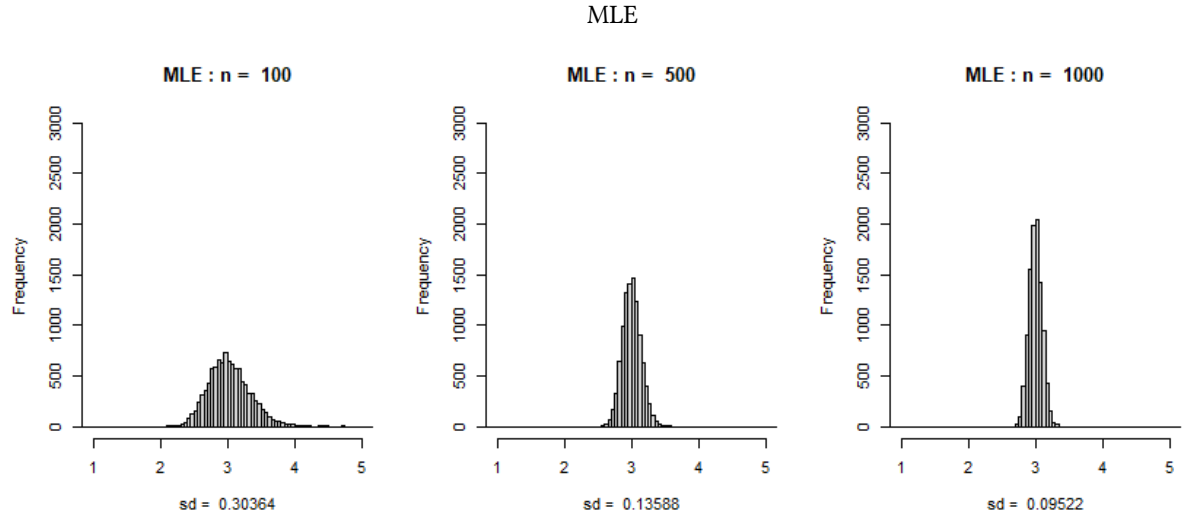


SME with $h(x) \equiv (x - \beta)x$



SME with $h(x) \equiv (x - \beta)x^2$





5.6.1 Robust Score matching estimator in Pareto case

Note that the final case can be considered under the situation in which we have a previous information that $\alpha \geq 0$. Now, we consider $h(x) \equiv (x-\beta)x^2$. In this case, (5.33) can be written as

$$\begin{aligned}\hat{\alpha}_{\text{SM}} &= \left(\sum_{i=1}^{n} \frac{X_i^2(X_i-\beta)}{X_i^2}\right)^{-1} \sum_{i=1}^{n} \frac{3X_i^2 - 2\beta X_i}{X_i} - 2 \\ &= \left(\sum_{i=1}^{n}(X_i-\beta)\right)^{-1} \sum_{i=1}^{n}(3X_i - 2\beta) - 2 \\ &= \left(\overline{X}-\beta\right)^{-1}\left(3\overline{X} - 2\beta\right) - 2\end{aligned}$$

By replacing the empirical mean $\overline{X}$ to a robust version, we can obtain a robust estimator for $\hat{\alpha}_{\text{SM}}$.

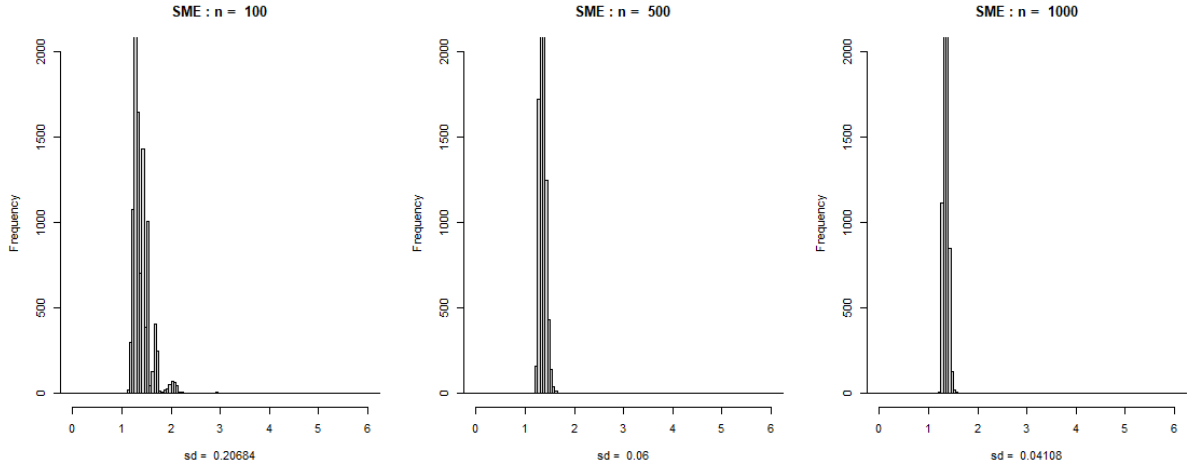
$$\hat{\alpha}_{\text{SM}_{\text{rob}}} = \left(\overline{X}_{\text{rob}} - \beta\right)^{-1}\left(3\overline{X}_{\text{rob}} - 2\beta\right) - 2.$$

The followings are the result with $\boldsymbol{X}_{\text{rob}}$ is Median of Mean and Trimmed mean. The population distribution, the number of samples, and the number of trials are the same as the above.
To see the performance of these estimators, we consider a contamination model, that is each sample $(X_i)_{i\in[n]}$ identically distributes

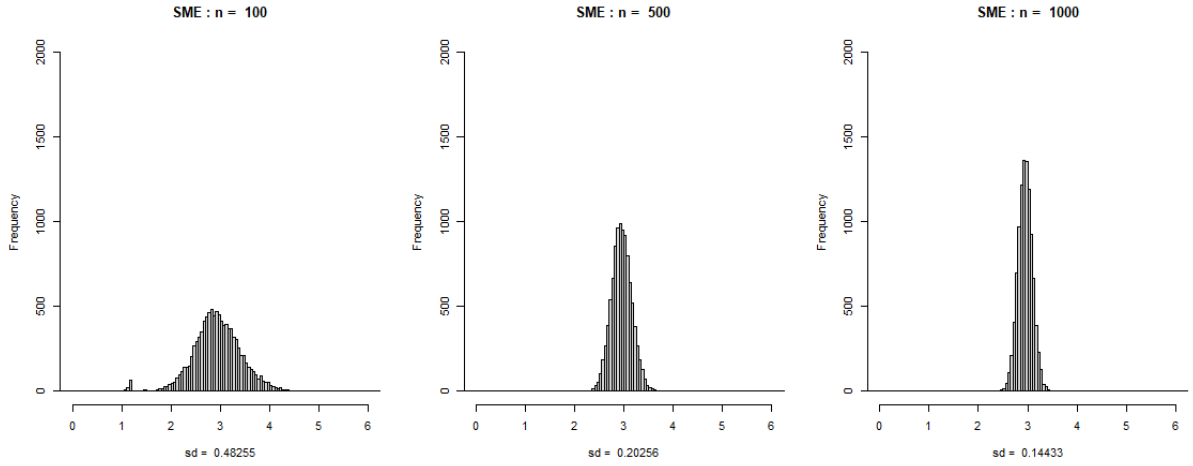
$$X_i = (1-\delta_i)Y_i + \delta_i 50 \qquad Y_i \overset{\text{i.i.d.}}{\sim} \text{Pareto}(\alpha = 3, \beta = 1) \quad \delta_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(0.05).$$

Then, the results are the followings, which show that while the estimate by non-robust score matching method is distorted by the outlier, the robust score matching estimator can reduce the effect of the outliers.

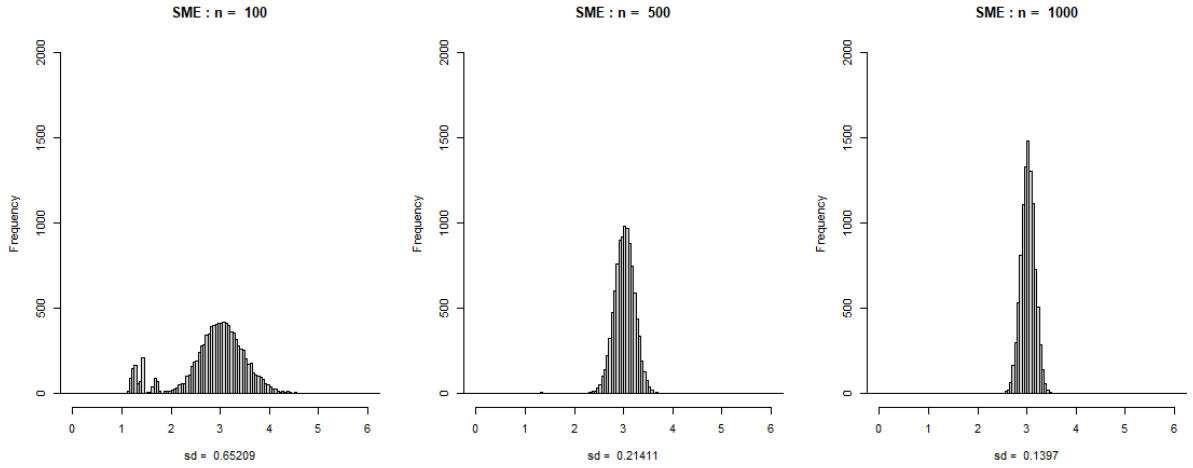
Non-robust SME



Robust SME with Median of Mean (The number of sample of each block is 5)



Robust SME with Trimmed mean (Removal rate is 0.1)



6 Summary

First of all in this thesis, we conducted a review of various concentration inequalities. Starting with discussions on the inequalities involving empirical means and empirical covariances, we proceeded to review several robust estimators that can serve as alternatives to them. In particular, as an alternative to the empirical covariance, we focused on the estimator introduced by Minsker [16]. The estimator was defined in (4.3) as a penalized least-squares estimator S with parameter λ_2 as the penalty weight for the sample deviation from the Gaussianity of the distribution.

Our primary focus in this thesis is on score matching estimator (SME). The original version of SME is introduced in [7], and generalized version was introduced in [8], which is defined as the minimizer of the empirical version of (5.1). When dealing with a exponential distribution family, SME can be explicitly represented by empirical means of multi-liniear forms of the sufficient statistics, as shown in Theorem 5.2.2. This representation provides us with the error bound of $\hat{\theta}_{\mathrm{SM}}$ from the true parameter $\boldsymbol{\theta}_0 \in \Theta$ given in Lemma 5.3.2, where Θ is a parameter space and a subset of $\mathbb{R}^m$. In case the distribution is $\mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$, $\Theta = \mathbb{R}^d \times PD_d$ where $PD_d \in \mathbb{R}^{d\times d}$ is the set of whole d dimensional symmetric and positive definite matrices. This decomposition allows to control the error $\left|\hat{\boldsymbol{\theta}}_{\mathrm{SM}} - \boldsymbol{\theta}_0\right|$ in terms of controlling the deviations of the mean vector $\boldsymbol{g}$ and covariance matrix $\boldsymbol{\Gamma}$ from their empirical analogues.

As an example we consider the case of Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with unknown parameters $\boldsymbol{\mu}, \Sigma$ We checked that the original SME for the parameters of Gaussian distribution coincides with the most likelihood estimator (MLE). By applying Theorem 5.2.2, we explicitly derived the expression for the generalized score matching estimator for the parameters of the Gaussian distribution (Corollary 5.4.3). For original score matching method, that is, we take $h(x) \equiv (1, \ldots, 1)^\top$ as the function $\boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}$ which was introduced in the (5.1), the SME can be expressed as a composition of empirical mean and empirical covariance as shown in Corollary 5.4.4.

One of the contributions of this thesis is a construction of a robust version of score matching estimator (RSM) for the unknown parameter $\boldsymbol{\mu}$ and Σ of $\mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$. For this goal, as we saw in Section 5.4.1, we replaced the empirical quantities corresponding to the empirical mean and covariance matrix in the generalized SME with robust alternatives. As shown in many books like [14] and papers such as [3], [5], and [16], there is a lot of candidates as a robust empirical mean and a robust covariance matrix estimator. In this thesis we chose the Median of Means and Trimmed Mean, to which we referred in Section 4.1, as alternatives to the empirical mean, and and employed Minsker's approach, which we introduce in Section 4.2, to construct a robust version of the empirical covariance estimator. This choice was made because the error bounds for these robust statistics can be explicitly calculated. And we showed that in case that $h(x) = (1, \ldots, 1)^\top$ and the underling distribution is Gaussian, RSE matches the alterative mean and covariance matrix estimators used in the robustification procedure. (Lemma 5.19)

To analyze the behavior of RSM, we adopted two approaches: error bounds and numerical simulations. By applying Lemma 5.3.2, we obtained an explicit error bound for RSM for the unknown parameters of Gaussan or Pareto distribution. In numerical approach, we considered the situation in which we have samples $(Z_i)_{i\in[n]}$ with $Z_i \sim (1-\delta_i)\boldsymbol{X}_i + \delta_i C_i$, where X_i distributes from Gaussian or Pareto distribution, C_i is some outlier, and $\delta \in [0, 1]$. In other words, the sample contains contaminations at the rate δ. The estimation error in this model differs from the uncontaminated model as we saw in (5.21) and (5.28), however (5.29) and (5.30) shows that RSM prove to be efficient. The idea to robustify the score matching

estimator by replacing the components of the estimator to robust alternatives can be extended to the case of general exponential distribution families because the score matching estimator for an exponential distribution family can be written as a liniear combination of empirical means and covariance matrix of derivatives of sufficient statistics. By substituting them with robust counterparts, we derive the RSM for the parameters of the exponential family.

In the last of the thesis, in Section 5.6 we considered the problem of estimating the shape parameter of Pareto distribution, the original SME can not be represented by empirical mean and covariance matrix. However generalized SME with suitable h enables us to write down the estimator as the composition of them, and we can replace them to robust alternatives and robust score matching estimators for the shape parameter of Pareto distribution. Even here, we consider a Pareto distribution with contaminations, and we confirmed the RSM's resilience to outliers .

There are several issues that we could not address in this thesis. As a RSM for the parameters of Gaussian distribution, we used Minsker method, which have parameter λ_2. It is interesting to determine the best parameter for effectively mitigating the influence of outliers. The choice of the most suitable method will depend on the sample size. To address this question, we need to simulate data using various methods for different sample sizes. And, here we only discussed about the specific h. When h is general, $\boldsymbol{\Gamma}(\boldsymbol{X})$ and $\boldsymbol{g}(\boldsymbol{X})$ include the term related to $h(\boldsymbol{X})$, SME for the parameters of Gaussian distribution cannot be expressed in terms of empirical mean and empirical covariances. For instance, when the function $h : \mathbb{R}^d \to \mathbb{R}^d$ s.t. $h(\boldsymbol{x}) = x$ as the function h in (5.1) , then we need a robust estimator for the 3rd moment at least. So we can not directly apply our idea dealing only with robust mean and covariance matrix estimator. I would like to leave these problems to the future researches.

As I conclude this thesis, I would like to express my gratitude to Prof. Mathias Drton and Dr. Oleksandr Zadorozhnyi. Professor Mathias offered me the opportunity to delve into the captivating subject of robustifying the score matching estimator. Dr. Zadorozhnyi consistently scheduled meetings and provided unwavering support throughout my thesis. He also generously shared valuable advice and made revisions to my drafts. I am profoundly thankful for all the insightful discussions with them.

7 Appendix : R code

7.1 Robust estimators

The following is functions to estimate the parameters from a given observations X by robust methods

```
library(plyr)
############################
### Robust Mean estimator###
############################
#Input : X:samples data nblocks:number of blocks
#Output: Median of mean
Median_of_mean_1D <- function(X,nblocks){#(data(n × 1),number of blocks)
  number_of_sample <- length(X) #input vector
  m <- floor(number_of_sample/nblocks)
  if(m*nblocks != number_of_sample){
    message("Number of samples can not be devided by number of blocks")
  }
  X_dev <- array(X,dim=c(m,nblocks))
  Z <- array(0,dim=c(1,nblocks))
  for(i in 1:nblocks){
    Z[i] <- mean(X_dev[,i])
  }
  return(median(Z))
}

#Input : X:samples data threshold:outlier rate
#Output: Trimmed mean
Trimmed_mean_1D <- function(X,e){
  number_of_sample <- length(X) #input vector
  under_threshold <- floor(number_of_sample/2 * e)
  #upper_threshold <- number_of_sample/2 - under_threshold
  if(under_threshold == 0){
    return(mean(X))
  }else{
  alfa <- sort(X[1:(number_of_sample/2)],partial =
                 under_threshold)[under_threshold]
  beta <- -sort(- X[1:(number_of_sample/2)], partial =
                  under_threshold)[under_threshold]
  total <- 0
  for(i in ((number_of_sample/2)+1):number_of_sample){
    total <- total + max(alfa,min(beta,X[i]))
  }
  return(total/length(((number_of_sample/2)+1):number_of_sample))
  }
}
#################################
```

```
### Robust Covariance estimator###
#################################
covariance_estimation <- function(X_multi_dim){
  #X is (number of data * dim )matrix
  result<- matrix(0, nrow = d, ncol = d)

  for(w in 1:d){
    result[w,w] = (hatt_sigma(X_multi_dim[,w]))^2
  }
  for(w1 in 2:d){
    for(w2 in 1:(w1-1)){
      sigma_plus  = hatt_sigma(X_multi_dim[,w1]/sqrt(result[w1,w1])+
                      X_multi_dim[,w2]/sqrt(result[w2,w2]))
      sigma_minus = hatt_sigma(X_multi_dim[,w1]/sqrt(result[w1,w1])-
                      X_multi_dim[,w2]/sqrt(result[w2,w2]))
      result[w1,w2] = (sigma_plus^2 -sigma_minus^2)/(sigma_plus^2 +
                        sigma_minus^2)*sqrt(result[w1,w1])*sqrt(result[w2,w2])
      result[w2,w1] =  result[w1,w2]
    }
  }
  return(result)
}

Median_absolute_deviation_1D <- function(X){#(data(n × 1),number of blocks)
  Const = 1.1926
  number_of_sample <- length(X) #input vector
  Y <- matrix(0,nrow = number_of_sample,ncol=number_of_sample)
  for(s in 1 : number_of_sample){
    for(t in 1: number_of_sample){
      Y[s,t] = abs(X[s]-X[t])
    }
  }
  partial_median <- rep(0,number_of_sample)
  for(s in 1 : number_of_sample){
      partial_median[s] = median(Y[s,])
  }
  return(Const*median(partial_median))
}

Altenative_MAD_1D <- function(X){#(data(n × 1),number of blocks)

  Const = 2.2219
  number_of_sample <- length(X) #input vector
  Y <- array(0,dim=c(number_of_sample,number_of_sample))
  for(s in 1:number_of_sample){
    for(t in 1:number_of_sample){
      Y[s,t] <- abs(X[s]-X[t])
    }
  }

  h <- (floor(number_of_sample/2) + 1)*floor(number_of_sample/2)
```

```
  return(Const*sort(Y,partial = h)[h])
}
```

7.2 Minsker method

The followings are the functions to conduct Minsker's method from a given observations X.

```
Huber_loss_prime <- function(lambda,u){
  if(abs(u) <= lambda){
    return(u)
  }else{
    if(u >= 0){return(lambda)}
    else{return(-lambda)}
  }
}

Huber_loss_prime_matrix <- function(lambda,Y){# M is d x d matrix
  output_Y <- matrix(0,nrow = dim(Y)[1], ncol = dim(Y)[2])
  for(j in 1:dim(Y)[1]){
    output_Y <- output_Y + Huber_loss_prime(lambda,eigen(Y)$values[j])*
                           eigen(Y)$vectors[,j]%*%t(eigen(Y)$vectors[,j])
  }
  return(output_Y)
}

Huber_gamma <- function(lambda,u){
  return(sign(u)*max(abs(u)-lambda,0))
}

Huber_gamma_matrix <- function(lambda,Y){# M is d x d matrix
  #Apply Huber_loss function for each entries
  output_Y <- matrix(0,nrow = dim(Y)[1], ncol = dim(Y)[2])
  for(j in 1:dim(Y)[1]){
    output_Y <- output_Y + Huber_gamma(lambda,eigen(Y)$values[j])*
                           eigen(Y)$vectors[,j]%*%t(eigen(Y)$vectors[,j])
  }
  return(output_Y)
}

Minsker <- function(Y,lambda1,lambda2,Number_of_iterations,B,S_initial){
  n <- dim(Y)[1]
  d <- dim(Y)[2]
  lambda = (sqrt(n*(n-1))*lambda2)/2
  stepsize <- rep(0,Number_of_iterations)
  for(t in 1: Number_of_iterations){
    if(B==0){stepsize[t] <- 1
    }else{stepsize[t] <- t^(-2/3)}
  }

  S_trend <- array(0,dim=c(d,d,Number_of_iterations))
  G_trend <- array(0,dim=c(d,d,Number_of_iterations))
```

```
  huber_trend <- array(0,dim=c(d,d,Number_of_iterations))
  Y_tilde <- array(0,dim = c(d,n,n))
  Y_tilde_Y_tilde_T <- array(0,dim = c(d,d,n,n))

  for(i in 1:n){
    for(j in 1:n){
      Y_tilde[ ,i,j] <- (Y[i,] - Y[j,])/(sqrt(2))
      Y_tilde_Y_tilde_T[ , ,i,j] <- Y_tilde[ ,i,j]%*%(t(Y_tilde[ ,i,j]))
    }
  }
  S <- S_initial
  for(t in 1:Number_of_iterations){ #Step 1
    G = 0
    if(B == 0){#counting up all the combinations
      for(ii in 1:(n-1)){
        for(jj in (ii+1):n){
          G  <- G -  Huber_loss_prime_matrix(lambda,
                      Y_tilde_Y_tilde_T[,,ii,jj]-S)/(n*(n-1)*0.5)
        }
      }
    }else{
      for(b in 1:B){
        i <- as.integer( runif(1, min = 1, max = n-1) )
        j <- as.integer( runif(1, min = i+1, max = n) )

        G <- G - (Huber_loss_prime_matrix(lambda, Y_tilde_Y_tilde_T[,,i,j]
                  - S))/B
      }
    }
    #Step 5,6
    S <- Huber_gamma_matrix((lambda1)/2,S - stepsize[t]*G)
    S_trend[,,t] <-S
    G_trend[,,t] <-G
  }
  return(list(S=S,S_trend = S_trend, G_trend = G_trend))
}
```

7.3 Score matching method

The following is the function to estimate the parameters from a given observations X by score matching method.

```
  # This R-code works only in case h = 1
  Score_Matching_3D <- function(X,n,flag){
  #Input : X:samples data n:number of Samples, flag : flag of Robust or Non-robust
  #Output: Estimates for mean and covariance by SME
  # flag is 1 for robust, 2 for non-robust
    d <- 3
    # set variables and vectors
    theta_SM <- matrix(0, nrow = 1, ncol = d*(d+1)/2 + d)
    Gamma11 <- array(0,dim=c(d,d,d))
    Gamma12 <- array(0,dim=c(d,1,d))
```

```
Gamma21 <- array(0,dim=c(1,d,d))
Gamma22 <- array(0,dim=c(1,1,d))
Gamma <- array(0, dim=c(d+1,d+1,d))
temp <- c(0,0)

if(flag == "Robust"){
  if(robust_covariance == "Minsker"){
    Sigma_SM <- Minsker(X,lambda1,lambda2,Number_of_iterations,B,S_initial)$S
  }else{
    Sigma_SM <- covariance_estimation(X)
  }
}
# Calculate Gamma
for(j in 1:d){
  if(flag == "Robust"){
    for(w in 1:d){
      Gamma12[w,1,j] =  - hatt_mu(X[,w])
      Gamma21[1,w,j] =  - hatt_mu(X[,w])
    }
    Gamma11[ , ,j] = Sigma_SM + t(t(Gamma12[ ,1,j])) %*% Gamma12[ ,1,j]

  }else if(flag == "Non-Robust"){
  for(i in 1:n){
    Gamma11[,,j] = Gamma11[,,j] + (X[i,]) %*% t(X[i,])/n
    Gamma12[,1,j] = Gamma12[,1,j] - (t(1 %*% ( X[i,])))/n
    Gamma21[1,,j] = Gamma21[1,,j] - t(X[i,])/n
  }
  }
  Gamma22[,,j] = 1
  Gamma[,,j] = rbind(cbind(Gamma11[,,j], Gamma12[,,j]),
                     abind(Gamma21[1,,j],Gamma22[,,j]))
}
zero <- matrix(0, nrow=d+1,ncol=d+1)
Gamma_12 = rbind(cbind(Gamma[,,1],zero,zero),cbind(zero, Gamma[,,2], zero),
                 cbind(zero, zero,Gamma[,,3]))
transition = rbind( # This is "R" in the thesis
  c(1,0,0,0,0,0,0,0,0),
  c(0,1,0,0,0,0,0,0,0),
  c(0,0,1,0,0,0,0,0,0),
  c(0,0,0,1,0,0,0,0,0),
  c(0,1,0,0,0,0,0,0,0),
  c(0,0,0,0,1,0,0,0,0),
  c(0,0,0,0,0,1,0,0,0),
  c(0,0,0,0,0,0,1,0,0),
  c(0,0,1,0,0,0,0,0,0),
  c(0,0,0,0,0,1,0,0,0),
  c(0,0,0,0,0,0,0,1,0),
  c(0,0,0,0,0,0,0,0,1)
)
Gamma_12t =  t(transition) %*% Gamma_12 %*% transition
```

```
    # Calculate the g
    e <- array(rbind(
      c(1,0,0),
      c(0,1,0),
      c(0,0,1)
    ),dim=c(3,3))
    g1 <- array(0, dim=c(d,d))
    g2 <- array(0, dim=c(1,d))
    for(j in 1:d){
      g1[,j] = e[j,]
      g2[,j] = 0
    }
    g <- abind(g1[,1],g2[,1],g1[,2],g2[,2],g1[,3],g2[,3])
    theta_SM[1, ] <- solve(Gamma_12t)%*% t(transition)%*% g

    K_SM <- matrix(c(theta_SM[1,1],theta_SM[1,2],theta_SM[1,3],
                     theta_SM[1,2],theta_SM[1,5],theta_SM[1,6],
                     theta_SM[1,3],theta_SM[1,6],theta_SM[1,8])
                     ,nrow = d, ncol = d)
    eta_SM <- matrix(c(theta_SM[1,4],
                       theta_SM[1,7],
                       theta_SM[1,9])
                     ,nrow =d, ncol =1)
    mu_SM <- solve(K_SM) %*% eta_SM
    Sigma_SM <- solve(K_SM)
    return(list(mu_SM=mu_SM,Sigma_SM=Sigma_SM))
  }
```

7.4 Maximum likelihood method

The following is the function to estimate the parameters from a given observations X by most likelihood method.

```
  Most_Likelihood <- function(X,n,flag){
#Input : X:samples data n:number of Samples, flag : flag of Robust or Non-robust
#Output: Estimates for mean and covariance by MLE
# flag is 1 for robust, 2 for non-robust

  Sigma_ML <- matrix(0,nrow = 3,ncol=3)
  if(flag == "Non-Robust"){
    mu_ML <- c(sum(X[,1])/n,sum(X[,2])/n,sum(X[,3])/n)
    for(i in 1:3){
      for(j in 1:3){
        Sigma_ML[i,j] <- sum((X[,i]-mu_ML[i])*(X[,j]-mu_ML[j]))/n
      }
    }
  }else if(flag == "Robust" ){
    mu_ML <- c(hatt_mu(X[,1]),hatt_mu(X[,2]),hatt_mu(X[,3]))
    if(robust_covariance == "Minsker"){
      Sigma_ML <- Minsker(X,lambda1,lambda2,Number_of_iterations,B,S_initial)$S
                  #See the Robust estimation R code
    }else{
```

```
      Sigma_ML <- covariance_estimation(X) #See the Robust estimation R code
    }
  }
  return(list(mu_ML=mu_ML,Sigma_ML=Sigma_ML))
}
```

7.5 TPR and FPR

This is the function to calculate TPR and FPR of the estimation result.

```
  TPR_FPR <- function(hatt_Sigma,Sigma0,K){
#Input : Estimates of covariance, true covariance, number of estimates
#Output: TPR and FPR for each epsilon
  TPR <- rep(0,100)
  FPR <- rep(0,100)
  d <-3
  for(e in 1:100){
    epsilon <- e/100
    S0_off <- 0
    hatt_S0_off <-  rep(0,100)
    hatt_S_off_and_S0_off <- rep(0,100)
    for(i in 1:(d-1)){
      for(j in (i+1):(d)){
        if(Sigma0[i,j] != 0){
          S0_off <- S0_off + 2
        }
      }
    }
    for(sample in 1:K){
      for(i in 1:(d-1)){
        for(j in (i+1):(d)){
          if((hatt_Sigma[sample,i,j])^2 > epsilon^2){
            hatt_S0_off[sample] <- hatt_S0_off[sample] + 2
          }
        }
      }
    }
    for(sample in 1:K){
      for(i in 1:(d-1)){
        for(j in (i+1):(d)){
          if((Sigma0[i,j] != 0)&((hatt_Sigma[sample,i,j])^2 > epsilon^2)){
            hatt_S_off_and_S0_off[sample]  <- hatt_S_off_and_S0_off[sample]  + 2
          }
        }
      }
    }
    FPR[e] <- (mean(hatt_S0_off - hatt_S_off_and_S0_off))/(d*(d-1) - S0_off )
    TPR[e] <- (mean(hatt_S_off_and_S0_off))/(S0_off )
  }
  return(rbind(TPR,FPR))
}
```

Bibliography

[1] Pinelis, I. (1994). *Optimum Bounds for the Distributions of Martingales in Banach Spaces*. The Annals of Probability, 22(4), 1679-1706.

[2] Pinelis, I. and Sakhanenko, A. I. (1986). *Remarks on Inequalities for Large Deviation Probabilities*. Theory of Probability & Its Applications, 30(1), 143-148.

[3] Kalisch, M. and Peter, Bühlmann. (2008). *Robustification of the PC-Algorithm for Directed Acyclic Graphs*. Journal of Computational and Graphical Statistics, 17(4), 773-789.

[4] Tropp, Joel A. (2015). *An Introduction to Matrix Concentration Inequalities*. Foundations and Trends in Machine Learning: Vol. 8: No. 1-2, pp 1-230.

[5] Lugosi, G., Mendelson, S. (2019). *Mean Estimation and Regression Under Heavy-Tailed Distributions: A Survey*. Found Comput Math 19, 1145-1190

[6] Rousseeuw, P.J. and Croux, C. (2013). *Alternatives to the median absolute deviation*. Journal of the American Statistical Association, 88(424): 1273-1283

[7] Hyvärinen, A. (2005) *Estimation of non-normalized statistical models by Score matching*. Jornal of Machine Learning Reserarch, Vol. 6: 695 -709

[8] Yu, S., Drton, M. and Shojaie, A. (2019) *Generalized Score Matching for Non-Negative Data*. Jornal of Machine Learning Reserarch, Vol. 20: 1 -70

[9] Lin, L., Drton, M. and Shojaie, A. (2016) *Estimation of high-dimensional graphical models using regularized score matching*. Electron. J. Stat., Vol. 10 No.1: 806 -854

[10] Bernstein, S. (1946) *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow

[11] Bennett, G. (1962) *Probability inequalities for the sum of independent random variables*. Journal of the American Statistical Association vol. 57 : 33-45

[12] Buldygin, V. V. and Kozachenko, Yu. V.(1980) *Subgaussian random variables* Ukrainian Mathematical Journal vol. 32 : 483 - 489

[13] Hoeffding, W.(1963) *Probability Inequalities for Sums of Bounded Random Variables*. Journal of the American Statistical Association Vol. 58, No. 301, 13-30

[14] Huber, P. J. and Romchetti, E. M.(2009) *Robust Statistics*. Wiley Probability and Statistics

[15] Berger, R. L. and Casella, G.(1990) *Statistical Inference*. Duxbury Press

[16] Minsker, S. and Wang, L.(2022) *Robust Estimation of Covariance Matrices: Adversarial Contamination and Beyond*, arXiv https://arxiv.org/abs/2203.02880

[17] Bhatia, R.(1997) *Matrix Analysis*, Springer

[18] Vershynin, R.(2010) *Introduction to the non-asymptotic analysis of random matrices*, arXiv https://arxiv.org/abs/1011.3027