About Stein's estimators: the original result and extensions

Auteur : Demaret, Tom

Promoteur(s) : Swan, Yvik

Faculté : Faculté des Sciences

Diplôme : Master en sciences mathématiques, à finalité spécialisée en statistique

Année académique : 2018-2019

URI/URL : http://hdl.handle.net/2268.2/6981

University of Liège

Master's thesis

About Stein's estimators

The original result and extensions

Author:
Tom Demaret

Supervisor:
Yvik Swan

A thesis submitted in fulfillment of the requirements for the Master's degree in mathematics

Department of mathematics

Faculty of Sciences

Academic year 2018-2019

Acknowledgments

I would like to thank Pr. Yvik Swan for proposing this topic to me and for his help and patience along the way.

Thanks also to the readers: in the hope they will have an enjoyable time and find this work interesting.

Contents

1 Stein's original result 6

2 First generalization: elliptically symmetric laws 14

3 Second generalization: high dimension 24

3.1 Reminders on the Moore-Penrose inverse 25

3.2 The main result 26

3.3 Technical results 29

4 Simulations 41

A Appendix: Technical results 55

A.1 Results on Wishart matrices 55

A.2 Results on the Moore-Penrose Inverse 57

B Appendix `R` code 59

Introduction

Suppose you are interested in estimating the mean θ of a Gaussian law and dispose, to this end, of a single random Gaussian variable X (which, as the Gaussian law is sum-stable, could itself be the sample mean of a large number of independent Gaussian random variables). Given the symmetry of the normal distribution, our intuition tells us that looking at X itself is the best way to estimate θ. To assess the quality of our intuitive estimator, the square error loss function and its associated risk, the mean square error, are the most common evaluation criterion. There are many reasons behind this popularity: convenience, elegance, mathematical tractability... In our case, an additional reason is its intrinsic link with the Gaussian law. Indeed, for the normal distribution, taking the maximum likelihood estimator is the same as minimizing the mean square error, and the estimator obtained this way is X, confirming our intuition. To further assess the quality of an estimator, there exists other ways: two important ones to set up the premises of this work are given in the following definitions.

Definition. *An estimator δ^* of a parameter $\theta \in \Theta$ is minimax with respect to a risk function $R(\theta, \delta)$ if*

$$\sup_{\theta \in \Theta} R(\theta, \delta^*) = \inf_{\delta} \sup_{\theta \in \Theta} R(\theta, \delta)$$

Intuitively, an estimator is minimax if it is "the best in the worst case".

Definition. *An estimator δ^* of a parameter $\theta \in \Theta$ is admissible with respect to a risk function $R(\theta, \delta)$ if no other estimator dominates it, meaning there does not exist an estimator δ such that $R(\theta, \delta) \leq R(\theta, \delta^*)$ for all θ and $R(\theta, \delta) < R(\theta, \delta^*)$ for at least some θ.*

In one dimension, the estimator X is minimax and admissible (see [12] and references in it), and at first glance, nothing seems to indicate that it should be any different in higher dimensions. In fact, the estimator X, for a variable whose mean is the vector θ, is minimax in any dimension and if the dimension is p and the covariance is $\sigma^2 I$ (we look at this case for simplicity as it is equivalent to looking at p independent one dimensional Gaussian variables and it makes computations easy), then the mean square error, $\mathbb{E}\left[\|X - \theta\|^2\right]$, is equal to $p\sigma^2$. It has also been shown to be admissible for $p = 2$ (see [12]), but in 1956, Charles Stein found estimators that dominate X as soon as p is greater or equal to 3. More precisely, he exhibited a biased estimator of the form $X + g(X)$, with g a certain function, that is also minimax: its risk is p minus a term proportional to $\frac{1}{\|X\|^2}$ so it becomes equal to that of X when the norm of θ goes to infinity, but with lower risk for all finite θ. One could say it exploits the weakness in the definition of a minimax estimator. This somewhat paradoxical result, stating in essence that combined information on unrelated events could bring better results overall than looking at each individually, came as a big surprise when first presented, but soon led to a plethora of research. This thesis aims to explore some of this research.

In Chapter 1, the first result by Stein is presented: starting from the simple case of identity covariance matrix and working our way step by step to the most direct generalizations. This is mostly taken from [12, 13, 14].

In Chapter 2, the estimators are extended to different probability laws. The link between mean square error and the Gaussian law comes from the presence of a $(x-\theta)^2$ term in the density function. This term appears in other laws, called elliptically symmetric laws, so it is natural that those laws behave similarly. This is taken from [6].

In our world filled with data, the case of high dimension and low sample size is getting more and more common and brings new kinds of problems. This is explored in Chapter 3. The main result is from [2].

Finally, the results are illustrated through simulations in Chapter 4 and some technical side results used are proved in the Appendix.

Chapter 1

Stein's original result

The James-Stein estimator is an estimator for the mean θ of a normal distribution which, at the price of a bias, dominates the usual estimator under mean square error: $\mathbb{E}\left[\|\hat{\theta}-\theta\|^2\right]$, where $\hat{\theta}$ is our estimator.

The result is based on a simple lemma, proved using Fubini's theorem.

Lemma 1.1. *Let X be a real random variable following a standard Gaussian law $\mathcal{N}(0,1)$ and $g:\mathbb{R}\to\mathbb{R}$ an absolutely continuous function such that g' is integrable. Then, if $\mathbb{E}\,|g'(X)|<\infty$,*

$$\mathbb{E}\left[g'(X)\right]=\mathbb{E}\left[Xg(X)\right].$$

Proof. The density function $\phi(x)$ of the standard Gaussian law, as it will be noted from now on, is such that $\phi'(x)=-x\phi(x)$.

Note also how, using $\int_{\mathbb{R}} y\phi(y)\mathrm{d}y=0$,

$$\int_{-\infty}^{x} -y\phi(y)\mathrm{d}y=\int_{x}^{+\infty} y\phi(y)\mathrm{d}y \quad \forall x\in\mathbb{R}$$

We then have

$$\begin{aligned}
\mathbb{E}\left[g'(X)\right] &= \int_{\mathbb{R}} g'(x)\phi(x)\mathrm{d}x \\
&= \int_{\mathbb{R}} g'(x)\left(\int_{-\infty}^{x}(-y\phi(y))\mathrm{d}y\right)\mathrm{d}x \\
&= \int_{0}^{+\infty} g'(x)\left(\int_{x}^{+\infty} y\phi(y)\mathrm{d}y\right)\mathrm{d}x-\int_{-\infty}^{0} g'(x)\left(\int_{-\infty}^{x} y\phi(y)\mathrm{d}y\right)\mathrm{d}x \\
&= \int_{0}^{+\infty} y\phi(y)\left(\int_{0}^{y} g'(x)\mathrm{d}x\right)dy-\int_{-\infty}^{0} y\phi(y)\left(\int_{y}^{0} g'(x)\mathrm{d}x\right)\mathrm{d}y \\
&= \int_{\mathbb{R}} y\phi(y)(g(y)-g(0))\mathrm{d}y \\
&= \mathbb{E}\left[Xg(X)\right]-g(0)\,\mathbb{E}[X]=\mathbb{E}\left[Xg(X)\right]
\end{aligned}$$

□

The result can be extended to any Gaussian random variable $Y \sim \mathcal{N}(\mu, \sigma^2)$ by writing $Y = \sigma X + \mu$ where $X \sim \mathcal{N}(0,1)$ and $h(y) = g\left(\frac{y-\mu}{\sigma}\right)$. We then obtain

$$\begin{aligned}\mathbb{E}\left[h'(Y)\right] &= \frac{1}{\sigma}\,\mathbb{E}\left[g'\left(\frac{Y-\mu}{\sigma}\right)\right] \\ &= \frac{1}{\sigma}\,\mathbb{E}\left[g'(X)\right] = \frac{1}{\sigma}\,\mathbb{E}\left[Xg(X)\right] \\ &= \frac{1}{\sigma}\,\mathbb{E}\left[\frac{Y-\mu}{\sigma}g\left(\frac{Y-\mu}{\sigma}\right)\right] \\ &= \mathbb{E}\left[\frac{Y-\mu}{\sigma^2}h(Y)\right].\end{aligned}$$

The result then needs to be extended to any dimension. To remain as general as possible, we will use the following definition.

Definition 1.1. A function $h : \mathbb{R}^p \to \mathbb{R}$ is called *almost differentiable* if there exists a function $\nabla h : \mathbb{R}^p \to \mathbb{R}^p$ such that, for all $a \in \mathbb{R}^p$,

$$h(x+a) = h(x) + \int_0^1 a \cdot \nabla h(x+ta)\mathrm{d}t.$$

The function ∇h is essentially the vector of partial derivatives, which is why we will use the following notations from now on. For a function $f : \mathbb{R}^p \to \mathbb{R}$,

- $\nabla f = (\partial_1 f, ..., \partial_p f)'$
- $\mathrm{div}(f) = \sum_{i=1}^p \partial_i f(X)$
- $\Delta f = \sum_{i=1}^p \partial_i^2 f(X)$.

Lemma 1.2. *Let X be a p-dimensional random variable following a standard Gaussian law with mean θ and the identity as covariance matrix, and $h : \mathbb{R}^p \to \mathbb{R}$ an almost differentiable function. If $\mathbb{E}\left|\nabla h(X)\right| < \infty$, then*

$$\mathbb{E}\left[\nabla h(X)\right] = \mathbb{E}\left[(X-\theta)h(X)\right].$$

Proof. For all $i \in \{1, ..., p\}$, write $X = (X_i, X_{-i})$, meaning that we decompose X between its ith component and all the others. Because X is normal, X_i and X_{-i} are independent and therefore, we find that, using Lemma 1.1,

$$\mathbb{E}\left[\partial_i h(X)|X_{-i}\right] = \mathbb{E}\left[(X_i - \theta_i)h(X)|X_{-i}\right],$$

and then, taking the expectation,

$$\mathbb{E}\left[\partial_i h(X)\right] = \mathbb{E}\left[(X_i - \theta_i)h(X)\right],$$

whence the conclusion. □

For what follows, we first consider X to be a p-dimensional random variable following a standard Gaussian law with mean θ and the identity as covariance matrix.

Theorem 1.1. *For an almost differentiable function $g : \mathbb{R}^p \to \mathbb{R}^p$ (meaning that all its components are almost differentiable), such that*

$$\mathbb{E}\left[\sum_{i=1}^{p} |\partial_i g_i(X)|\right] < \infty, \tag{A}$$

we have

$$\mathbb{E}\left[\|X + g(X) - \theta\|^2\right] = p + \mathbb{E}\left[\|g(X)\|^2 + 2\text{div}(g(X))\right]. \tag{1.1}$$

Proof. Using Lemma 1.2, we have

$$\begin{aligned}\mathbb{E}\left[(X_i + g_i(X) - \theta_i)^2\right] &= \mathbb{E}\left[(X_i - \theta_i)^2 + 2(X_i - \theta_i)g_i(X) + g_i^2(X)\right] \\ &= 1 + 2\,\mathbb{E}\left[\partial_i g_i(X)\right] + \mathbb{E}\left[g_i^2(X)\right]\end{aligned}$$

and we get the result by summing over i. □

This theorem gives an expression for the MSE of any estimator of the form $X + g(X)$ by decomposing it between the error of the usual estimator $\mathbb{E}\left[\|X - \theta\|^2\right] = p$ and a term that depends on the function g. The goal will be to make this term negative. With this in mind, we concentrate in (1.1) on functions $g : \mathbb{R}^p \to \mathbb{R}^p$ of the form

$$g = \nabla \log f = \frac{\nabla f}{f}$$

with f such that this is well defined. This leads to a modified version of Theorem 1.1.

Theorem 1.2. *Let $f : \mathbb{R}^p \to \mathbb{R}_0^+$ be a almost differentiable function such that ∇f is also almost differentiable,*

$$\mathbb{E}\left[\frac{1}{f(X)}\sum_{i=1}^{p} |\partial_i^2 f(X)|\right] < \infty \tag{A'}$$

and

$$\mathbb{E}\left[\|\nabla \text{log} f(X)\|^2\right] < \infty.$$

Then

$$\begin{aligned}\mathbb{E}\left[\|X + \nabla\text{log}f(X) - \theta\|^2\right] &= p + \mathbb{E}\left[2\frac{\nabla^2 f(X)}{f(X)} - \frac{\|\nabla f(X)\|^2}{f^2(X)}\right] \\ &= p + 4\,\mathbb{E}\left[\frac{\nabla^2\sqrt{f(X)}}{\sqrt{f(X)}}\right].\end{aligned}$$

Remark. Note that condition A' is simply the rewriting of condition A with our particular choice of g and the other conditions ensure that the final expression is finite.

Proof. Using

$$\mathrm{div}(\nabla \mathrm{log} f) = \frac{\Delta f}{f} - \frac{\|\nabla f\|^2}{f^2}.$$

we directly get from 1.1

$$\begin{aligned}\mathbb{E}\left[\|X + \nabla \mathrm{log} f(X) - \theta\|^2\right] &= p + \mathbb{E}\left[\frac{\|\nabla f\|^2}{f^2} + 2\left(\frac{\Delta f}{f} - \frac{\|\nabla f\|^2}{f^2}\right)\right] \\ &= p + \mathbb{E}\left[2\frac{\Delta f}{f} - \frac{\|\nabla f\|^2}{f^2}\right].\end{aligned}$$

Finally, as

$$\Delta\left(\sqrt{f}\right) = \mathrm{div}(\nabla f) = \mathrm{div}\left(\frac{\nabla f}{2\sqrt{f}}\right) = \frac{\Delta f}{2\sqrt{f}} - \frac{\|\nabla f\|^2}{4f^{\frac{3}{2}}},$$

we can replace $\|\nabla f\|^2$ by

$$2f(\Delta f) - 4f^{\frac{3}{2}}(\Delta\sqrt{f})$$

to get the desired result. □

From this, it follows that if we can find a function f satisfying the assumptions of the theorem and such that $\Delta\left(\sqrt{f(x)}\right) \leq 0$, then we have an estimator, $X + \nabla \mathrm{log} f(X)$, that dominates the usual estimator X. Indeed, in that case,

$$\mathbb{E}\left[\|X + \nabla \mathrm{log} f(X) - \theta\|^2\right] \leq p = \mathbb{E}\left[\|X - \theta\|^2\right]$$

The famous James-Stein estimator is obtained by choosing $f(x) = \left(\frac{1}{\|X\|^2}\right)^b$ (see [15] for more details).

We have

$$\nabla f(X) = -b\left(\frac{1}{\|X\|^2}\right)^{-(b+1)} 2X$$

so

$$\nabla \mathrm{log} f(X) = \frac{\nabla f(X)}{f(X)} = \frac{-2b}{\|X\|^2}X$$

and the estimator is $\left(1 - \frac{2b}{\|X\|^2}\right)X$. As we have

$$\Delta\left(\sqrt{f(X)}\right) = -\frac{b(p-2-b)}{\|X\|^{b+2}},$$

this estimator dominates X for $0 \leq b \leq (p-2)$ (when $p > 2$). Its risk is equal to $p - 4\,\mathbb{E}\left[\frac{b(p-2-b)}{\|X\|^2}\right]$. The improvement is therefore maximal for $b = \frac{p-2}{2}$, whence the following definition.

Definition 1.2. The usual James-Stein estimator for $X \sim \mathcal{N}_p(\theta, I)$ is

$$\left(1 - \frac{p-2}{\|X\|^2}\right) X. \tag{1.2}$$

Its risk is equal to $p - (p-2)^2 \, \mathbb{E}\left[\frac{1}{\|X\|^2}\right]$.

It is interesting to note that the degree of the improvement depends on the value of $\|X\|^2$: the closer X is to 0, the bigger the improvement will be. On the other hand, if X is big, the James-Stein estimator will be very close to the actual value X. This means that the estimator is much more useful when $X \sim \mathcal{N}_p(\theta, I)$ with $\|\theta\|$ small. It may therefore be useful to modify the estimator, by centering the data first. This will be studied by simulations, in Section 4.

We will now work by steps to generalize this result to any covariance matrix.

Let's first look at the case where the covariance matrix is a multiple of the identity, i.e. $X \sim \mathcal{N}_p(\theta, \sigma^2 I)$. This is the case if we consider a sample of p independent normally distributed variables as a vector of size p.

If σ^2 is known, looking at $\frac{X}{\sigma}$ and applying previous results, the James-Stein estimator for X takes the form

$$\left(1 - \frac{(p-2)\sigma^2}{\|X\|^2}\right) X \tag{1.3}$$

and its risk is

$$\sigma^2 \left(p - (p-2)^2 \, \mathbb{E}\left[\frac{1}{\|X\|^2}\right]\right).$$

If σ^2 is unknown, it needs to be estimated and we assume we have at our disposal a variable $s \sim \sigma^2 \chi^2_n$, independent of X. This would typically be obtained through a sample of $n+1$ independent normally distributed variables v_i ($v_i \sim \mathcal{N}(\mu, \sigma^2)$), by taking $s = \sum_{i=1}^{n+1} (v_i - \bar{v})^2$.

Remark. The notation s is used for readability purposes, but be careful that it does not denote the standard deviation. In our example, it denotes n times the sample variance.

Setting

$$Y = \frac{X}{\sigma}, \quad \eta = \frac{\theta}{\sigma}, \quad s^* = \frac{s}{\sigma^2} \tag{1.4}$$

and looking at estimators of the form $\left(1 - \frac{c(p-2)s}{\|X\|^2}\right) X$, with c a constant to be determined, we get

$$\mathbb{E}\left[\left\| X - \frac{c(p-2)s}{\|X\|^2} X - \theta \right\|^2\right]$$

$$
\begin{aligned}
&= \sigma^2\,\mathbb{E}\left[\left\|Y - \frac{c(p-2)s^*}{\|Y\|^2}Y - \eta\right\|^2\right]\\
&= \sigma^2\,\mathbb{E}\left[\left((Y-\eta) - \frac{c(p-2)s^*}{\|Y\|^2}Y\right)'\left((Y-\eta) - \frac{c(p-2)s^*}{\|Y\|^2}Y\right)\right]\\
&= \sigma^2\,\mathbb{E}\left[\|Y-\eta\|^2 - 2c(p-2)s^*\frac{(Y-\eta)'Y}{\|Y\|^2} + c^2(p-2)^2 s^{*2}\frac{1}{\|Y\|^2}\right]\\
&= \sigma^2\left(p - 2c(p-2)\,\mathbb{E}\left[s^*\right]\mathbb{E}\left[\frac{(Y-\eta)'Y}{\|Y\|^2}\right] + c^2(p-2)^2\,\mathbb{E}\left[s^{*2}\right]\mathbb{E}\left[\frac{1}{\|Y\|^2}\right]\right)\\
&= \sigma^2\left(p - 2c(p-2)n\,\mathbb{E}\left[\frac{(Y-\eta)'Y}{\|Y\|^2}\right] + c^2(p-2)^2 n(n+2)\,\mathbb{E}\left[\frac{1}{\|Y\|^2}\right]\right)
\end{aligned}
$$

using the independence of Y and s^* and the expression for the moment of a χ^2_n law.

The middle term can easily be computed by integration by parts, but, because $Y \sim \mathcal{N}_p(\eta, I)$, we can also write

$$
\begin{aligned}
\mathbb{E}\left[\left\|Y - \frac{(p-2)}{\|Y\|^2}Y - \eta\right\|^2\right] &= \mathbb{E}\left[\|Y-\eta\|^2 - 2(p-2)\frac{(Y-\eta)'Y}{\|Y\|^2} + (p-2)^2\frac{1}{\|Y\|^2}\right]\\
&= p + (p-2)^2\,\mathbb{E}\left[\frac{1}{\|Y\|^2}\right] - 2(p-2)\,\mathbb{E}\left[\frac{(Y-\eta)'Y}{\|Y\|^2}\right]
\end{aligned}
$$

and, as we know that the first term is also equal to $p - (p-2)^2\,\mathbb{E}\left[\frac{1}{\|Y\|^2}\right]$, we find that

$$
\mathbb{E}\left[\frac{(Y-\eta)'Y}{\|Y\|^2}\right] = (p-2)\,\mathbb{E}\left[\frac{1}{\|Y\|^2}\right]
$$

Wrapping up, this gives us

$$
\begin{aligned}
&\mathbb{E}\left[\left\|X - \frac{c(p-2)s}{\|X\|^2}X - \theta\right\|^2\right]\\
&= \sigma^2\left(p - n(p-2)^2(2c - c^2(n+2))\,\mathbb{E}\left[\frac{1}{\|Y\|^2}\right]\right).
\end{aligned}
$$

Since c is arbitrary, we can choose it so as to minimize $2c - c^2(n+2)$, which is done by taking $c = \frac{1}{n+2}$.

Definition 1.3. The James-Stein estimator for $X \sim \mathcal{N}_p(\theta, \sigma^2 I)$, with σ^2 unknown estimated through $s \sim \sigma^2\chi^2_n$, is

$$
\left(1 - \frac{(p-2)s}{(n+2)\|X\|^2}\right)X. \tag{1.5}
$$

Its risk is equal to $\sigma^2\left(p - \frac{n}{n+2}(p-2)^2\,\mathbb{E}\left[\frac{1}{\|X\|^2}\right]\right)$.

Note how estimating σ^2 only caused a loss of precision by a proportion of $\frac{2}{n+2}$ compared to the case of σ^2 known.

Taking $c = \dfrac{1}{n}$ is another possibility. While less optimal, it yields the estimator

$$\left(1 - \frac{(p-2)s}{n\|X\|^2}\right) X, \tag{1.6}$$

where $\dfrac{s}{n}$ appears. As $\dfrac{s}{n}$ is such that $\mathbb{E}\left[\dfrac{s}{n}\right] = \sigma^2$, this is simply the estimator (1.3) where σ^2 has been replaced by its estimation.

The same kind of transformation as in (1.4) can be done if the covariance is of the form $\Sigma = \mathrm{diag}(\sigma_1^2, ..., \sigma_p^2)$. If $X \sim \mathcal{N}_p(\theta, \Sigma)$ and $S = \mathrm{diag}(s_1, ..., s_p)$ with $s_i \sim \sigma_i^2 \chi^2_{n_i}$ for $i \in \{1, ..., p\}$, denoting $\sqrt{\Sigma} = \mathrm{diag}(\sigma_1, ..., \sigma_p)$, we would define

$$Y = \sqrt{\Sigma}^{-1} X, \quad \eta = \sqrt{\Sigma}^{-1} \theta, \quad s^* = \Sigma^{-1} S$$

Each component of our James-Stein estimator would then be $\left(1 - \dfrac{(p-2)s_i}{n_i\|X\|^2}\right) X_i$.

As an example, imagine p independent samples $V_{i1}, ..., V_{i(n_i+1)}$ are available, where $V_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2)$, for $i \in \{1, ..., p\}$ and $j \in \{1, ..., (n_i+1)\}$. Compute $\bar{V}_i = \frac{1}{n_i+1} \sum_{j=1}^{n_i+1} V_{ij}$ and $s_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i+1} (V_{ij} - \bar{V}_i)^2$, the sample means and variances of each sample. As we are in the normal case, $\bar{V}_i$ and s_i^2 are independent for all i and $s_i^2 \sim \frac{\sigma_i^2}{n_i} \chi^2_{n_i}$. We now consider $X = (\bar{V}_1, ..., \bar{V}_p)'$ and $S = \mathrm{diag}(s_1^2, ..., s_p^2)$. Clearly, $X \sim \mathcal{N}_p(\theta, \Sigma)$ with $\theta = (\mu_1, ..., \mu_p)'$ and $\Sigma = \mathrm{diag}(\frac{\sigma_1^2}{n_1}, ..., \frac{\sigma_p^2}{n_p})$. So if we take the estimator $\hat{X}_{JS}$ with

$$\left(\hat{X}_{JS}\right)_i = \left(1 - \frac{(p-2)s_i^2}{n_i\|X\|^2}\right) \bar{V}_i, \quad \text{for } i \in \{1, ..., p\},$$

we know that

$$\mathbb{E}\left[\|\hat{X}_{JS} - \theta\|^2\right] \leq \mathbb{E}\left[\|X - \theta\|^2\right].$$

For more applied examples, the article [4], "Stein's Paradox in Statistics" by Efron and Morris is a great read.

Finally, we consider the general case $X \sim \mathcal{N}_p(\theta, \Sigma)$ with Σ an unknown positive semi-definite matrix. Here we assume having at our disposal S, a Wishart matrix with n degrees of freedom ($S \sim \mathcal{W}_p(n, \Sigma)$). Similarly as before, this means for example a sample of size n, $V_1, ..., V_n$, ($V_i \sim \mathcal{N}_p(0, \Sigma)$) is available and $S = \sum V_i V_i'$. The loss function we use in this case takes the slightly different form of

$$\mathbb{E}\left[(\hat{\theta} - \theta)' \Sigma^{-1} (\hat{\theta} - \theta)\right].$$

We assume S is invertible and consider estimators of the form

$$\left(1 - \frac{c(p-2)}{X' S^{-1} X}\right) X. \tag{1.7}$$

We have

$$\begin{aligned}&\mathbb{E}_{\theta,\Sigma}\left[\left(\left(1-\frac{c(p-2)}{X'S^{-1}X}\right)X-\theta\right)'\Sigma^{-1}\left(\left(1-\frac{c(p-2)}{X'S^{-1}X}\right)X-\theta\right)\right]\\ &=\mathbb{E}_{\theta^*,I}\left[\left(\left(1-\frac{c(p-2)}{Y'S^{*-1}Y}\right)Y-\theta^*\right)'\left(\left(1-\frac{c(p-2)}{Y'S^{*-1}Y}\right)Y-\theta^*\right)\right],\end{aligned}$$

by applying the transformation $X \to PDX = Y$, where D is a matrix such that $D\Sigma D' = I$, and P is an orthogonal matrix with its first row proportional to $D\theta$, so that $PDX = Y \sim \mathcal{N}_p(\theta^*, I)$, with $\theta^* = (\sqrt{\theta'\Sigma^{-1}\theta}, 0, ..., 0)'$ and $S^* = PDSD'P'$ (so that $S^{*-1} = P'^{-1}D'^{-1}S^{-1}D^{-1}P^{-1}$) following a Wishart $\mathcal{W}_p(n, I)$.

As the conditional distribution of $Y'S^{*-1}Y$ given Y is that of $\dfrac{Y'Y}{Q}$ where $Q \sim \chi^2_{n-p+1}$ (see the appendix for more details), we find ourselves in the same situation as for the case of σ^2 unknown. Thus, the optimal choice for c is $\frac{1}{n-p+3}$ and the James-Stein estimator is

$$\left(1-\frac{p-2}{(n-p+3)X'S^{-1}X}\right)X, \tag{1.8}$$

its risk being

$$p-\frac{n-p+1}{n-p+3}(p-2)^2\,\mathbb{E}\left[\frac{1}{\|X\|^2}\right]. \tag{1.9}$$

Chapter 2

First generalization: elliptically symmetric laws

The next natural step is to see if similar estimators exist for other probability laws than the Gaussian. We will imagine having at our disposal n random p-dimensional variables, X and $V_1, ..., V_{n-1}$, whose joint density is of the form

$$f\left((X-\theta)'\Sigma^{-1}(X-\theta)+\sum_{j=1}^{n-1} V_j'\Sigma^{-1}V_j\right), \tag{2.1}$$

with $f : \mathbb{R} \to \mathbb{R}^+$ a Lebesgue integrable function.

Both θ and Σ are unknown. The V_j will be used to estimate Σ through

$$S = \sum_{j=1}^{n-1} V_j V_j',$$

so that we can look at estimators of θ of the form $\delta(X,S) = X + g(X,S)$ under the loss

$$\mathbb{E}\left[(\delta(X,S)-\theta)'\Sigma^{-1}(\delta(X,S)-\theta)\right].$$

The matrix S is again assumed to be invertible.

All the distributions defined this way are elliptically symmetric. Taking $f(R^2)$ proportional to $\exp(-\frac{1}{2}R^2)$ yields the multivariate normal, while taking it proportional to $\left(1+\frac{1}{\nu}R^2\right)^{-(\nu+p)/2}$ gives the multivariate t-distribution.

As before, we will look at

$$\begin{aligned}&\mathbb{E}\left[(X+g(X,S)-\theta)'\Sigma^{-1}(X+g(X,S)-\theta)\right]\\&\qquad=\mathbb{E}\left[(X-\theta)'\Sigma^{-1}(X-\theta)\right]+\mathbb{E}\left[(2g(X,S)'\Sigma^{-1}(X-\theta)\right]+\mathbb{E}\left[g(X,S)'\Sigma^{-1}g(X,S)\right],\end{aligned}$$

and make it so that the last two terms are negative for the estimator to dominate X.

The result depends on two lemmas allowing us to express the two terms we're interested in using expectation $\mathbb{E}^*$, which will denote the expectation with respect to the distribution

$$C^{-1}F\left((X-\theta)'\Sigma^{-1}(X-\theta)+\sum_{j=1}^{n-1}V_j'\Sigma^{-1}V_j\right)$$

where F is defined by

$$F(t)=\int_t^{+\infty}f(s)\mathrm{d}s \tag{2.2}$$

and C^{-1} is a normalizing constant, i.e.

$$C=\int_{\mathbb{R}^p\times\ldots\times\mathbb{R}^p}F\left((x-\theta)'\Sigma^{-1}(x-\theta)+\sum_{j=1}^{n-1}v_j'\Sigma^{-1}v_j\right)\mathrm{d}x\mathrm{d}v_1\ldots\mathrm{d}v_{n-1}.$$

This is in contrast to the standard expectation $\mathbb{E}$, with respect to the f in (2.1).

Before jumping into the results, we can try to get some intuition as to where this F comes from. In the multinormal case, i.e. $f(R^2)\propto\exp(-\frac{1}{2}R^2)$, we have $F=f$ and therefore $\mathbb{E}^*=\mathbb{E}$. Considering dimension 1 for simplicity, the result in Lemma 1.1 in Chapter 1 is based on the fact that $\phi'(x)=-x\phi(x)$ for the normal density function. For variables with density of the form $f(x^2)$, this can be generalized by writing

$$(F(x^2))'=-2xf(x^2).$$

The following lemma is therefore simply a generalization of Stein's Lemma 1.2 in Chapter 1, coinciding with it in the multinormal case.

Lemma 2.1. *If $g(x,.)$ is a differentiable function, then*

$$\mathbb{E}\left[g(X,S)'\Sigma^{-1}(X-\theta)\right]=C\,\mathbb{E}^*\left[\mathrm{div}_X(g(X,S))\right]. \tag{2.3}$$

Lemma 2.2. *Let $T(X,S)$ be a function from $\mathbb{R}^p\times\mathbb{R}^{p\times p}$ to $\mathbb{R}^{p\times p}$. Then we have*

$$\mathbb{E}\left[\mathrm{tr}\left(T(X,S)\Sigma^{-1}\right)\right]=2C\,\mathbb{E}^*\left[D^*_{1/2}(T(X,S))\right]+C(n-p-2)\,\mathbb{E}^*\left[\mathrm{tr(S^{-1}T(X,S)}\right]$$

with

$$D^*_{1/2}(T(X,S))=\sum_{i=1}^{p}\frac{\partial T_{ii}(X,S)}{\partial S_{ii}}+\frac{1}{2}\sum_{i\neq j}\frac{\partial T_{ij}(X,S)}{\partial S_{ij}}.$$

This $D^*_{1/2}$ can be seen as a generalization of divergence for matrices, with the $\dfrac{1}{2}$ being a symptom of the symmetry of S.

Remark. Notation $\mathrm{tr}(A)$ is used for the trace of a $p \times p$ matrix,

$$\mathrm{tr}(A) = \sum_{i=1}^{p} A_{ii},$$

and $\dfrac{\partial}{\partial A_{ij}}$ denotes the derivative with respect to the component (i, j) of the matrix A (or vector in some cases).

Before proving these lemmas, we will see how they can be used, as it is quite direct.

Theorem 2.1. *If* $\mathbb{E}\,[X'X] < \infty$ *and* $\mathbb{E}\,[g'(X,S)g(X,S)] < \infty$*, then*

$$\begin{aligned} &\mathbb{E}\left[(X + g(X,S) - \theta)'\Sigma^{-1}(X + g(X,S) - \theta)\right] - \mathbb{E}\left[(X-\theta)'\Sigma^{-1}(X-\theta)\right] \\ &= \mathbb{E}\left[(2g(X,S)'\Sigma^{-1}(X-\theta)\right] + \mathbb{E}\left[g(X,S)'\Sigma^{-1}g(X,S)\right] \\ &= C\,\mathbb{E}^*\left[2\mathrm{div}_X(g(X,S)) + (n-p-2)g'(X,S)S^{-1}g(X,S) + 2D^*_{1/2}(g(X,S)g'(X,S))\right]. \end{aligned} \tag{2.4}$$

Proof. Using Lemma 2.1 on the first term of the difference gives

$$\mathbb{E}\left[2g(X,S)'\Sigma^{-1}(X-\theta)\right] = C\,\mathbb{E}^*\left[2\mathrm{div}_X(g(X,S))\right].$$

Lemma 2.2, with $T(X,S) = g(X,S)g'(X,S)$, on the second gives

$$\begin{aligned} \mathbb{E}\left[g(X,S)'\Sigma^{-1}g(X,S)\right] &= \mathbb{E}\left[\mathrm{tr}\left(g(X,S)'\Sigma^{-1}g(X,S)\right)\right] \\ &= \mathbb{E}\left[\mathrm{tr}\left(\Sigma^{-1}g(X,S)g(X,S)'\right)\right] \\ &= 2C\,\mathbb{E}^*\left[D^*_{1/2}(g(X,S)g(X,S)')\right] \\ &\quad + C(n-p-2)\,\mathbb{E}^*\left[\mathrm{tr(S^{-1}g(X,S)g(X,S)'}\right]. \end{aligned}$$

The results follows immediately. □

In the light of equation (2.4), we are in a similar position as after Theorem 1.1 in Chapter 1. Following the same heuristic, it will be our objective to identify functions $g : \mathbb{R}^p \times \mathbb{R}^{p\times p} \to \mathbb{R}^p$ such that (2.4) is negative, hereby yielding domination in terms of MSE. This will be performed in Corollary 2.2, at the end of this chapter.

The proofs of the two lemmas rely on an integration by slice result, as well as a corollary involving Stokes Theorem, that can be derived from [5](Theorem 3.2.12). This divides $\mathbb{R}^p$ into ellipsoids and allows the same kind of "integration by parts" generalization as in Stein's Lemma 1.1.

Lemma 2.3. *For any $r \in \mathbb{R}$ and any continuously differentiable function ϕ defined on $\mathbb{R}^p$, let $[\phi = r]$ be the submanifold in $\mathbb{R}^p$ associated with ϕ. Then, for any Lebesgue integrable function f, we have*

$$\int_{\mathbb{R}^p} f(x)\mathrm{d}x = \int_{\{r\in\mathbb{R}\,|[\phi=r]\neq\emptyset\}} \int_{[\phi=r]} \frac{f(x)}{\|\nabla\phi(x)\|}\mathrm{d}\sigma_r\mathrm{d}r,$$

where σ_r is the area measure defined on $[\phi = r]$.

Corollary 2.1. *If g is a function defined on $\mathbb{R}^p$ such that $\nabla\phi \cdot g$ is integrable, then*

$$\int_{\mathbb{R}^p} \nabla\phi(x) \cdot g(x)\mathrm{d}x = \int_{\{r \in \mathbb{R}\,|[\phi = r] \neq \emptyset\}} \int_{\mathcal{B}_r} \mathrm{div}(g(x))\mathrm{d}x\mathrm{d}r,$$

where $\mathcal{B}_r$ is the set with boundary $[\phi = r]$ corresponding, for any $x \in [\phi = r]$, to the outward normal vector $\nabla\phi(x)$.

We now prove the two lemmas.

Proof of Lemma 2.1. We want to compute

$$\begin{aligned}
&\mathbb{E}\left[(g(X,S)'\Sigma^{-1}(X-\theta)\right] \\
&= \int_{\mathbb{R}^p \times \ldots \times \mathbb{R}^p} \int_{\mathbb{R}^p} g(x,s)'\Sigma^{-1}(x-\theta) f\left((x-\theta)'\Sigma^{-1}(x-\theta) + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j\right) \mathrm{d}x\mathrm{d}v_1 \ldots \mathrm{d}v_{n-1}
\end{aligned}$$

We define $\phi(x) = \sqrt{(x-\theta)'\Sigma^{-1}(x-\theta)}$ to use Lemma 2.3 and Corollary 2.1 on the inner integral. We have

$$\nabla\phi(x) = \frac{\Sigma^{-1}(x-\theta)}{\sqrt{(x-\theta)'\Sigma^{-1}(x-\theta)}}$$

and therefore, setting $R = \sqrt{(x-\theta)'\Sigma^{-1}(x-\theta)}$,

$$\begin{aligned}
&\int_{\mathbb{R}^p} g(x,s)'\Sigma^{-1}(x-\theta) f\left((x-\theta)'\Sigma^{-1}(x-\theta) + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j\right)\mathrm{d}x \\
&= \int_0^{+\infty} f\left(R^2 + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j\right) \int_{[\phi=R]} \frac{g(x,s)'\Sigma^{-1}(x-\theta)}{\|\nabla\phi(x)\|}\mathrm{d}\sigma_r\mathrm{d}R \\
&= \int_0^{+\infty} f\left(R^2 + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j\right) \int_{[\phi=R]} g(x,s)'\sqrt{(x-\theta)'\Sigma^{-1}(x-\theta)}\frac{\nabla\phi(x)}{\|\nabla\phi(x)\|}\mathrm{d}\sigma_r\mathrm{d}R \\
&= \int_0^{+\infty} Rf\left(R^2 + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j\right) \int_{[\phi=R]} g(x,s)'\frac{\nabla\phi(x)}{\|\nabla\phi(x)\|}\mathrm{d}\sigma_r\mathrm{d}R \\
&= \int_{\mathbb{R}^p} Rf\left(R^2 + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j\right) \nabla\phi(x) \cdot g(x)\mathrm{d}x
\end{aligned}$$

(using Lemma 2.3 backwards so that we can now use Corollary 2.1)

$$= \int_0^{+\infty} Rf\left(R^2 + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j\right) \int_{[\phi \leq R]} \mathrm{div}_x(g(x,s))\mathrm{d}x\mathrm{d}R$$

$$
\begin{aligned}
&= \int_{\mathbb{R}^p} \operatorname{div}_x(g(x,s)) \int_{\sqrt{(x-\theta)'\Sigma^{-1}(x-\theta)}}^{+\infty} Rf\left(R^2 + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j\right) \mathrm{d}R\mathrm{d}x \\
&= \int_{\mathbb{R}^p} \operatorname{div}_x(g(x,s)) \frac{1}{2} \int_{(x-\theta)'\Sigma^{-1}(x-\theta)}^{+\infty} f\left(r + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j\right) \mathrm{d}r\mathrm{d}x \\
&= \int_{\mathbb{R}^p} \operatorname{div}_x(g(x,s)) F\left((x-\theta)'\Sigma^{-1}(x-\theta) + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j\right) \mathrm{d}x.
\end{aligned}
$$

Replacing this into the first expression gives the desired result in (2.3). □

Proof of Lemma 2.2. First develop the term in the expectation:

$$
\begin{aligned}
\operatorname{tr}(T(X,S)\Sigma^{-1}) &= \operatorname{tr}(T(X,S)\Sigma^{-1}SS^{-1} \\
&= \operatorname{tr}(T(X,S)\Sigma^{-1}\sum_{i=1}^{n-1} V_iV_i'S^{-1} \\
&= \sum_{i=1}^{n-1} \operatorname{tr}(V_i'S^{-1}T(X,S)\Sigma^{-1}V_i) \\
&= \sum_{i=1}^{n-1} V_i'S^{-1}T(X,S)\Sigma^{-1}V_i.
\end{aligned}
$$

Then, mimicking the proof of Lemma 2.1 done just before, with V_i instead of $X-\theta$ and $g(V_i,S) = T(X,S)S^{-1}V_i$, we get (with S^{lm} denoting S^{-1}_{lm})

$$
\begin{aligned}
\mathbb{E}\left[\operatorname{tr}(T(X,S)\Sigma^{-1})\right] &= C\sum_{i=1}^{n-1} \mathbb{E}^*\left[\operatorname{div}_{V_i}(T(X,S)S^{-1}V_i)\right] \\
&= C\sum_{i=1}^{n-1} \mathbb{E}^*\left[\sum_{j=1}^{p} \frac{\partial}{\partial V_{ij}}\left(\sum_{m=1}^{p}\sum_{l=1}^{p} T_{jl}(X,S)S^{lm}V_{im}\right)\right] \\
&= C\,\mathbb{E}^*\left[A_1 + A_2 + A_3\right], \qquad (2.5)
\end{aligned}
$$

and

$$
\begin{aligned}
A_1 &= \sum_{i=1}^{n-1}\sum_{j=1}^{p}\sum_{m=1}^{p}\sum_{l=1}^{p}\left(\frac{\partial}{\partial V_{ij}}V_{im}\right)T_{jl}(X,S)S^{lm}, \\
A_2 &= \sum_{i=1}^{n-1}\sum_{j=1}^{p}\sum_{m=1}^{p}\sum_{l=1}^{p} V_{im}\left(\frac{\partial}{\partial V_{ij}}T_{jl}(X,S)\right)S^{lm}, \\
\text{and } A_3 &= \sum_{i=1}^{n-1}\sum_{j=1}^{p}\sum_{m=1}^{p}\sum_{l=1}^{p} V_{im}T_{jl}(X,S)\left(\frac{\partial}{\partial V_{ij}}S^{lm}\right).
\end{aligned}
$$

We then compute A_1, A_2 and A_3 separately. First,

$$\begin{aligned} A_1 &= \sum_{i=1}^{n-1}\sum_{j=1}^{p}\sum_{m=1}^{p}\sum_{l=1}^{p} \delta_{jm} T_{jl}(X,S) S^{lm} \\ &= (n-1)\sum_{j=1}^{p}\sum_{l=1}^{p} T_{jl}(X,S) S^{lj} \\ &= (n-1)\operatorname{tr}(T(X,S)S^{-1}). \end{aligned}$$

For A_2, because S is symmetric, we first get

$$A_2 = \sum_{i=1}^{n-1}\sum_{j=1}^{p}\sum_{m=1}^{p}\sum_{l=1}^{p} V_{im} S^{lm} \left(\sum_{q\le r} \frac{\partial T_{jl}(X,S)}{\partial S_{qr}} \frac{\partial S_{qr}}{\partial V_{ij}} \right).$$

Using the definition of $S = \sum_{i=1}^{n-1} V_i V_i'$, we have

$$\frac{\partial S_{qr}}{\partial V_{ij}} = \frac{\partial}{\partial V_{ij}} (V_{iq} V_{ir}) = V_{iq}\delta_{jr} + V_{ir}\delta_{jq}$$

and then

$$\sum_{i=1}^{n-1} (V_{iq}\delta_{jr} + V_{ir}\delta_{jq}) V_{im} = S_{mq}\delta_{jr} + S_{mr}\delta_{jq}.$$

The following is immediate from the definition of the inverse, but it is reminded because it will be used several times:

$$\sum_{m=1}^{p} S^{am} S_{mb} = \delta_{ab}. \tag{2.6}$$

We now get

$$\begin{aligned} A_2 &= \sum_{j=1}^{p}\sum_{m=1}^{p}\sum_{l=1}^{p} S^{lm} \left(\sum_{q\le r} \frac{\partial T_{jl}(X,S)}{\partial S_{qr}} S_{mq}\delta_{jr} + S_{mr}\delta_{jq} \right) \\ &= \sum_{j=1}^{p}\sum_{l=1}^{p}\sum_{q\le r} \frac{\partial T_{jl}(X,S)}{\partial S_{qr}} (\delta_{lq}\delta_{jr} + \delta_{lr}\delta_{jq}) \quad \text{(thanks to (2.6))} \\ &= \sum_{j=1}^{p}\sum_{l=1}^{p}\sum_{q\le r} \left(\frac{\partial T_{jl}(X,S)}{\partial S_{lj}} \delta_{lq}\delta_{jr} + \frac{\partial T_{jl}(X,S)}{\partial S_{jl}} \delta_{lr}\delta_{jq} \right) \\ &= \sum_{j=1}^{p}\sum_{l=1}^{p}\sum_{q\le r} \frac{\partial T_{jl}(X,S)}{\partial S_{lj}} (\delta_{lq}\delta_{jr} + \delta_{lr}\delta_{jq}). \end{aligned}$$

Finally noting that

$$\sum_{q\le r} (\delta_{lq}\delta_{jr} + \delta_{lr}\delta_{jq}) = \begin{cases} 2 \text{ if } j = l, \\ 1 \text{ if } j \ne l, \end{cases}$$

we get

$$\begin{aligned} A_2 &= 2\left(\sum_{j=1}^{p} \frac{\partial T_{jj}(X,S)}{\partial S_{jj}} + \frac{1}{2}\sum_{j\neq p} \frac{\partial T_{jl}(X,S)}{\partial S_{jl}}\right) \\ &= 2D^*_{1/2}(T(X,S)). \end{aligned}$$

The last term left is A_3. Starting the same way as for A_2, we can get

$$A_3 = \sum_{j=1}^{p}\sum_{l=1}^{p} T_{jl}(X,S) \sum_{m=1}^{p}\sum_{q\leq r} \frac{\partial S^{lm}}{\partial S_{qr}} (S_{mq}\delta_{jr} + S_{mr}\delta_{jq}). \tag{2.7}$$

To compute $\frac{\partial S^{lm}}{\partial S_{qr}}$, we use (2.6) again and take the derivative on both sides to find

$$\sum_{j=1}^{p} \left(\frac{\partial S^{lj}}{\partial S_{st}} S_{jn} + S^{lj} \frac{\partial S_{jn}}{\partial S_{st}} \right) = 0.$$

As S is symmetric, if $q \neq r$, $\frac{\partial S_{jn}}{\partial S_{qr}} = \delta_{jq}\delta_{nr} + \delta_{jr}\delta_{nq}$, which leads to

$$\begin{aligned} \sum_{j=1}^{p} \frac{\partial S^{lj}}{\partial S_{qr}} S_{jn} &= -\sum_{j=1}^{p} S^{lj}(\delta_{jq}\delta_{nr} + \delta_{jr}\delta_{nq}) \\ &= -S^{lq}\delta_{nr} - S^{lr}\delta_{nq}. \end{aligned}$$

We now multiply this by S^{nm} and sum over n to get

$$\sum_{n=1}^{p}\sum_{j=1}^{p} \frac{\partial S^{lj}}{\partial S_{qr}} S_{jn} S^{nm} = \sum_{n=1}^{p} \left(-S^{lq}\delta_{nr} - S^{lr}\delta_{nq}\right) S^{nm}.$$

Killing a sum thanks to (2.6) again, this gives us

$$\sum_{j=1}^{p} \frac{\partial S^{lj}}{\partial S_{qr}} \delta_{jm} = \frac{\partial S^{lm}}{\partial S_{qr}} = -S^{lq}S^{rm} - S^{lr}S^{qm}.$$

The case $q = r$ is treated similarly, as $\frac{\partial S_{jn}}{\partial S_{qq}} = \delta_{jq}\delta_{nq}$, to get

$$\frac{\partial S^{lm}}{\partial S_{qr}} = \begin{cases} -S^{lq}S^{rm} - S^{lr}S^{qm} & \text{if } q \neq r, \\ -S^{lq}S^{qm} & \text{if } q = r. \end{cases} \tag{2.8}$$

In (2.7), the second half becomes

$$\sum_{m=1}^{p}\sum_{q\leq r} \frac{\partial S^{lm}}{\partial S_{qr}} (S_{mq}\delta_{jr} + S_{mr}\delta_{jq})$$

$$
\begin{aligned}
&= -\sum_{m=1}^{p}\left(\sum_{q=r} 2S_{mq}\delta_{jq}S^{lq}S^{qm} + \sum_{q<r}(S_{mq}\delta_{jr} + S_{mr}\delta_{jq})(S^{lq}S^{rm} + S^{lr}S^{qm})\right)\\
&= -\sum_{m=1}^{p}\left(2S_{mj}S^{lj}S^{jm} + \sum_{q<r}(S_{mq}\delta_{jr} + S_{mr}\delta_{jq})(S^{lq}S^{rm} + S^{lr}S^{qm})\right).
\end{aligned}
$$

Placing this last expression into the one for A_3 in (2.7), after using once again that S and S^{-1} are symmetric and (2.6), we obtain

$$
\begin{aligned}
A_3 &= -\sum_{j=1}^{p}\sum_{l=1}^{p} T_{jl}(X,S)\left(2S^{lj} + \sum_{q<r}(\delta_{jr}S^{lr} + \delta_{jq}S^{lq})\right)\\
&= -\sum_{j=1}^{p}\sum_{l=1}^{p} T_{jl}(X,S)S^{lj}\left(2 + \sum_{q<r}(\delta_{jr} + \delta_{jq})\right)\\
&= -\sum_{j=1}^{p}\sum_{l=1}^{p} T_{jl}(X,S)S^{lj}(2 + (p-1))\\
&= -(p+1)\mathrm{tr}\left(T(X,S)S^{-1}\right).
\end{aligned}
$$

Substituting the expressions found for A_1,A_2 and A_3 in (2.5), we get the announced result. □

As anticipated, we can now find estimators that dominate X.

Corollary 2.2. *Let $r : \mathbb{R} \to \mathbb{R}$ be a nondecreasing positive function bounded by $\dfrac{2(p-2)}{n-p+2}$. If $\mathbb{E}\left[X'X\right] < \infty$ and $\mathbb{E}\left[\dfrac{X'X}{(X'S^{-1}X)^2}\right] < \infty$, then the estimator*

$$
\left(1 - \frac{r(X'S^{-1}X)}{X'S^{-1}X}\right)X \tag{2.9}
$$

dominates X under mean square error.

Remark. The function r replaces the constants in the estimators from Chapter 1 (the c in (1.7) more specifically) and the bounds on r are reminiscent of those on those constants.

Proof. We apply Theorem 2.1 with $g(X,S) = -\dfrac{r(X'S^{-1}X)}{X'S^{-1}X}X$ and show that

$$
C\,\mathbb{E}^*\left[2\mathrm{div}_X(g(X,S)) + (n-p-2)g(X,S)'S^{-1}g(X,S) + 2D^*_{1/2}(g(X,S)g(X,S)')\right] \leq 0. \tag{2.10}
$$

For the first term of (2.10), we need to compute

$$
\mathrm{div}_X(g(X,S)) = -\mathrm{div}_X\left(\frac{r(X'S^{-1}X)}{X'S^{-1}X}X\right)
$$

$$= -\left(\mathrm{div}_X(X)\frac{r(X'S^{-1}X)}{X'S^{-1}X} + X'\nabla_X\left(\frac{r(X'S^{-1}X)}{X'S^{-1}X}\right)\right).$$

As $\mathrm{div}_X(X) = p$ and

$$\nabla_X\left(\frac{r(X'S^{-1}X)}{X'S^{-1}X}\right) = \left(2\frac{r'(X'S^{-1}X)X'S^{-1}X - r(X'S^{-1}X)}{(X'S^{-1}X)^2}S^{-1}X\right),$$

this gives the expression

$$\mathrm{div}_X(g(X,S)) = -\left((p-2)\frac{r(X'S^{-1}X)}{X'S^{-1}X} + 2r'(X'S^{-1}X)\right).$$

The second term in (2.10) is direct:

$$g(X,S)'S^{-1}g(X,S) = \frac{r^2(X'S^{-1}X)}{X'S^{-1}X}.$$

Finally, for the last term, we have

$$\begin{aligned}
&D^*_{1/2}(g(X,S)g(X,S)') \\
&= D^*_{1/2}\left(\frac{r^2(X'S^{-1}X)}{(X'S^{-1}X)^2}X'X\right) \\
&= \sum_{i=1}^{p}\frac{\partial}{\partial S_{ii}}\left(\frac{r^2(X'S^{-1}X)}{(X'S^{-1}X)^2}\right)X_i^2 + \frac{1}{2}\sum_{i\neq j}\frac{\partial}{\partial S_{ij}}\left(\frac{r^2(X'S^{-1}X)}{(X'S^{-1}X)^2}\right)X_iX_j.
\end{aligned}$$

As

$$\begin{aligned}
&\frac{\partial}{\partial S_{ij}}\left(\frac{r^2(X'S^{-1}X)}{(X'S^{-1}X)^2}\right) \\
&= \left(\frac{2(X'S^{-1}X)^2r(X'S^{-1}X)r'(X'S^{-1}X) - 2(X'S^{-1}X)r^2(X'S^{-1}X)}{(X'S^{-1}X)^4}\right)\frac{\partial}{\partial S_{ij}}(X'S^{-1}X)
\end{aligned}$$

and, using (2.8),

$$\begin{aligned}
\frac{\partial}{\partial S_{ij}}(X'S^{-1}X) &= \sum_{l,m}X'_l\frac{\partial S^{lm}}{\partial S_{ij}}X_m \\
&= -(2-\delta_{ij})(X'S^{-1})_i(X'S^{-1})_j,
\end{aligned}$$

it follows that

$$\begin{aligned}
&D^*_{1/2}(g(X,S)g'(X,S)) \\
&= \left(\frac{2(X'S^{-1}X)^2r(X'S^{-1}X)r'(X'S^{-1}X) - 2(X'S^{-1}X)r^2(X'S^{-1}X)}{(X'S^{-1}X)^4}\right)
\end{aligned}$$

$$
\begin{aligned}
&\times \left(\sum_{i=1}^{p} \frac{\partial}{\partial S_{ii}}(X'S^{-1}X)X_i^2 + \frac{1}{2}\sum_{i\neq j}\frac{\partial}{\partial S_{ij}}(X'S^{-1}X)X_iX_j\right)\\
&= -2\left(\frac{(X'S^{-1}X)^2 r(X'S^{-1}X)r'(X'S^{-1}X) - (X'S^{-1}X)r^2(X'S^{-1}X)}{(X'S^{-1}X)^4}\right)\\
&\quad \times \underbrace{\left(\sum_{i=1}^{p}(X'S^{-1})_i^2X_i^2 + \frac{1}{2}\sum_{i\neq j}2(X'S^{-1})_i(X'S^{-1})_jX_iX_j\right)}_{=(X'S^{-1}X)^2}\\
&= -2\left(r(X'S^{-1}X)r'(X'S^{-1}X) - \frac{r^2(X'S^{-1}X)}{X'S^{-1}X}\right).
\end{aligned}
$$

Finally putting everything back together in the expression (2.10), we obtain

$$
\begin{aligned}
&C\,\mathbb{E}^*\left[2\mathrm{div}_X(g(X,S)) + (n-p-2)g(X,S)'S^{-1}g(X,S) + 2D^*_{1/2}(g(X,S)g(X,S)')\right]\\
=&C\,\mathbb{E}^*\Bigg[2\left((p-2)\frac{r(X'S^{-1}X)}{X'S^{-1}X} + 2r'(X'S^{-1}X)\right)\\
&\qquad + (n-p-2)\frac{r^2(X'S^{-1}X)}{X'S^{-1}X}\\
&\qquad\quad - 4\left(r(X'S^{-1}X)r'(X'S^{-1}X) - \frac{r^2(X'S^{-1}X)}{X'S^{-1}X}\right)\Bigg]\\
=&C\,\mathbb{E}^*\Bigg[\frac{r(X'S^{-1}X)}{X'S^{-1}X}\left(-2(p-2) + (n-p-2)r(X'S^{-1}X)\right)\\
&\qquad - 4r'(X'S^{-1}X)\left(1 + r(X'S^{-1}X)\right)\Bigg]
\end{aligned}
$$

which, given the definition of r, is negative as required. □

Choosing r constant and equal to $\frac{(p-2)}{n-p+2}$ (n became $n-1$ given how we considered the problem in this section) gives the James-Stein estimator. This means that the estimator (1.8) found in the normal case also works in the multivariate Student case: simulations will be used to compare them in Chapter 4.

Chapter 3

Second generalization: high dimension

An implicit assumption for the previous results was for the dimension p to be smaller than the sample size n in order for the estimated covariance matrix to be non-singular. Indeed, if, given a sample $Y_1, ..., Y_n$, we compute a matrix S as before by $S = \sum_{j=1}^{n} Y_j Y_j'$, then $\text{rk}(S) \leq \min(n, p)$. It is easier to see by writing $S = Y'Y$, with Y the $n \times p$ matrix whose rows are our observations Y_i. As the Y_i are usually independent and, in the cases that interest us, from a continuous distribution, the rank of S is actually equal to $\min(n, p)$ almost surely. Therefore, when $p > n$, S is a $p \times p$ matrix of rank n and is almost surely singular. To illustrate this, let's look at 3 observations (rounded to the nearest hundredth for readability) from a 4-dimensional normal with mean 0 and covariance matrix the diagonal matrix with $1, 2, 3, 4$ on the diagonal. The 3 vectors are

$$\begin{aligned} X_1 &= (0.43, -0.45, -1.59, -2.96)' \\ X_2 &= (-0.78, -0.89, -3.46, 3.15)' \\ X_3 &= (-1.29, -0.15, -0.47, -1.91)' \end{aligned}$$

and the matrix S obtained, along with its singular value decomposition to show its singularity and also anticipating on what is to come, is

$$S = \begin{pmatrix} 1.56 & -0.12 & -0.41 & -2.19 \\ -0.12 & 0.28 & 1.12 & -2.05 \\ -0.41 & 1.12 & 4.56 & -8.46 \\ -2.19 & -2.05 & -8.46 & 21.39 \end{pmatrix} = U \begin{pmatrix} 25.27 & 0 & 0 & 0 \\ 0 & 2.52 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} V'$$

where U and V are two unitary matrices. For the exact values, the `R` code used is in the appendix.

When matrices are singular like here, generalization of the inverse exists, with diverse properties. One of the most popular ones, due to its uniqueness, is the Moore-Penrose pseudoinverse and the focus of this section will be to show how to use it to obtain new forms of the James-Stein estimator.

We put ourselves back in the situation where X is a p-dimensional variable normally distributed with mean θ and covariance matrix Σ and we would like to estimate θ, knowing that we have at our disposal a random matrix S, independent of X, following a Wishart distribution $\mathcal{W}_p(n, \Sigma)$. By definition of the Wishart, this means $S = Y'Y$, where $Y = (Y_1, ..., Y_n)$ with $Y_i \overset{iid}{\sim} \mathcal{N}_p(0, \Sigma)$.

When $p \leq n$, we obtained earlier the estimators (1.8) and (2.9). As they involve S^{-1}, they can not be used when $p > n$ because S is singular. Instead, the Moore-Penrose pseudoinverse of S, noted S^+, will be used to form the estimator

$$\left(I - \frac{r(X'S^+X)SS^+}{X'S^+X}\right) X \tag{3.1}$$

with $r : \mathbb{R} \to \mathbb{R}$, a positive bounded differentiable real function. This will be proved to dominate the usual estimator X under the same loss as in Chapter 2, $\mathbb{E}\left[(\hat{\theta} - \theta)'\Sigma^{-1}(\hat{\theta} - \theta)\right]$.

This estimator is a direct extension of (2.9), as a property of the Moore-Penrose inverse is to coincide with the actual inverse when the matrix is invertible.

3.1 Reminders on the Moore-Penrose inverse

Before jumping into the proof of the domination of the estimator (3.1), the definition of the Moore-Penrose generalized inverse, along with some simple properties that will be useful, are reminded here. Some proofs are skipped and more details can be found in the appendix.

Definition 3.1. For a matrix $A \in \mathbb{R}^{m \times n}$, the Moore-Penrose inverse is the matrix $A^+ \in \mathbb{R}^{n \times m}$ such that

- $AA^+A = A$
- $A^+AA^+ = A^+$
- $(AA^+)' = AA^+$
- $(A^+A)' = A^+A$

This matrix exists and is unique.

It can be computed using the singular value decomposition. If $A = UMV'$ where U and V are two square unitary matrices of sizes m and n respectively, and M is a matrix the same dimensions as A with non negative real numbers on the diagonal and zeros everywhere else, then $A^+ = VM^+U'$. The Moore-Penrose inverse of M is simply its transpose where every diagonal element is replaced by its inverse. Looking back at our earlier example, the Moore-Penrose inverse of S is

$$S^+ = \begin{pmatrix} 0.22 & -0.05 & -0.18 & -0.06 \\ -0.05 & 0.01 & 0.04 & 0.01 \\ -0.18 & 0.04 & 0.15 & 0.04 \\ -0.06 & 0.01 & 0.04 & 0.05 \end{pmatrix} = V \begin{pmatrix} \frac{1}{25.27} & 0 & 0 & 0 \\ 0 & \frac{1}{2.52} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} U'$$

with U and V the matrices from the singular value decomposition of S.

Looking at symmetric matrices, as will be the case with S, a few additional properties will be useful.

If S is symmetric, its pseudoinverse is also symmetric, as we can see on the example. This observation comes naturally by looking at the transpose of the conditions in the definition: $(S^+)'$ is also a pseudoinverse of S and it follows from uniqueness that S^+ is symmetric. It then follows from the definition that $SS^+ = S^+S$ and that

$$S(I - SS^+) = (I - SS^+)S = S^+(I - SS^+) = (I - SS^+)S^+ = 0. \tag{3.2}$$

The formula for the derivative will also be used, which can be found in [7] (Theorem 4.3).

Proposition 3.1. *For $A(t)$ a differentiable matrix function of constant rank, we have*

$$\frac{\partial A^+}{\partial t} = -A^+\frac{\partial A}{\partial t}A^+ + (I - A^+A)\frac{\partial A'}{\partial t}(A^+)'A^+ + A^+(A^+)'\frac{\partial A'}{\partial t}(I - AA^+).$$

In the symmetric case, this becomes

$$\frac{\partial S^+}{\partial t} = -S^+\frac{\partial S}{\partial t}S^+ + (I - SS^+)\frac{\partial S}{\partial t}S^+S^+ + S^+S^+\frac{\partial S}{\partial t}(I - SS^+) \tag{3.3}$$

Finally, it will be interesting to note that, as $S = Y'Y$, the pseudoinverse of Y is S^+Y'. A proof of this is in the appendix. From this, looking at

$$\begin{aligned} SS^+Y' = Y'YY^+ = \big((Y'YY^+)'\big)' &= \big((YY^+)'Y\big)' \\ &= \big((YS^+Y')'Y\big)' \\ &= \big(YS^+Y'Y\big)' && \text{(because } S^+ \text{ is symmetric)} \\ &= \big(YY^+Y\big)' = Y', \end{aligned}$$

we can show that

$$SS^+Y' = Y' \quad \Leftrightarrow \quad (I - SS^+)Y' = 0. \tag{3.4}$$

We can now move onto the main result.

3.2 The main result

Theorem 3.1. *Let $\min(p, n) \geq 3$ and $r : \mathbb{R} \to \mathbb{R}$ be a differentiable function. The estimator (3.1) dominates X if*

- *r is between 0 and $\dfrac{2(\min(n,p) - 2)}{n + p - 2\min(n,p) + 3}$,*
- *r is nondecreasing,*

- *r' is bounded.*

This covers both the case $p > n$ and $p \leq n$, as when $p \leq n$, then $S^+ = S^{-1}$ and the estimator (3.1) is equal to the estimator in the previous section.

Proof. Several technical results will be used along the way, whose proofs will be given later, in Section 3.3, for clarity.

Without originality, we will again look at the difference of risk between our estimator, which will be written $X + g(X, S)$ with g defined appropriately, and X:

$$\begin{aligned}\Delta_\theta &= \mathbb{E}\left[(X + g(X,S) - \theta)'\Sigma^{-1}(X + g(X,S) - \theta)\right] - \mathbb{E}\left[(X-\theta)'\Sigma^{-1}(X-\theta)\right] \\ &= 2\,\mathbb{E}\left[g(X,S)'\Sigma^{-1}(X-\theta)\right] + \mathbb{E}\left[g(X,S)'\Sigma^{-1}g(X,S)\right], \qquad (3.5)\end{aligned}$$

and show that this is nonpositive. The function g is defined as

$$g(X,S) = -\frac{r(X'S^+X)SS^+X}{X'S^+X}.$$

Using Lemma (2.1) from the previous section, the first part in (3.5) is equal to

$$2\,\mathbb{E}\left[\operatorname{div}_X(g(X,S))\right],$$

which is in turn, using Lemma 3.3, equal to

$$2\,\mathbb{E}\left[2r'(X'S^+X) + r(X'S^+X)\frac{\operatorname{tr}(SS^+) - 2}{X'S^+X}\right].$$

Thanks to Lemma 3.4, for the second term on the right hand of (3.5), we find

$$\begin{aligned}&\mathbb{E}\left[g(X,S)'\Sigma^{-1}g(X,S)\right] \\ &= \mathbb{E}\left[\operatorname{tr}\left(\Sigma^{-1}Sr^2(X'S^+X)\frac{S^+XX'S^+S}{(X'S^+X)^2}\right)\right] \\ &= \mathbb{E}\left[n\operatorname{tr}\left(r^2(X'S^+X)\frac{S^+XX'S^+S}{(X'S^+X)^2}\right)\right. \\ &\qquad \left. + \operatorname{tr}\left(Y'\nabla_Y\left(r^2(X'S^+X)\frac{SS^+XX'S^+}{(X'S^+X)^2}\right)\right)\right].\end{aligned}$$

The assumption needed for Lemma 3.4 is proved in Theorem 3.2.

Pursuing our developments, with the help of Lemma 3.2, we reap

$$\begin{aligned}&\mathbb{E}\left[g(X,S)'\Sigma^{-1}g(X,S)\right] \\ &= \mathbb{E}\left[n\frac{r^2(X'S^+X)}{X'S^+X} - 4r(X'S^+X)r'(X'S^+X) + r^2(X'S^+X)\frac{p - 2\operatorname{tr}(SS^+) + 3}{X'S^+X}\right] \\ &= \mathbb{E}\left[r^2(X'S^+X)\frac{n + p - 2\operatorname{tr}(SS^+) + 3}{X'S^+X} - 4r(X'S^+X)r'(X'S^+X)\right].\end{aligned}$$

Plugging these expressions in (3.5), we obtain

$$\begin{aligned}\Delta_\theta = \mathbb{E}\Bigg[& r^2(X'S^+X)\frac{n+p-2\operatorname{tr}(SS^+)+3}{X'S^+X} \\ & - 2r(X'S^+X)\frac{\operatorname{tr}(SS^+)-2}{X'S^+X} \\ & - 4r'(X'S^+X)\big(1+r(X'S^+X)\big)\Bigg]\end{aligned} \tag{3.6}$$

As r is nonnegative and nondecreasing, we know that $-4r'(X'S^+X)\big(1+r(X'S^+X) \leq 0$. To finish, if $r(X'S^+X) \neq 0$, we have

$$\begin{aligned} r^2(X'S^+X)\frac{n+p-2\operatorname{tr}(SS^+)+3}{X'S^+X} - 2r(X'S^+X)\frac{\operatorname{tr}(SS^+)-2}{X'S^+X} \leq 0 \\ \Leftrightarrow r(X'S^+X) \leq \frac{2\operatorname{tr}(SS^+)-2}{n+p-2\operatorname{tr}(SS^+)+3} = \frac{2(\min(n,p)-2)}{n+p-2(\min(n,p)-2)+3},\end{aligned}$$

which corresponds to our assumptions on r. We therefore have, as desired, that $\Delta_\theta \leq 0$. □

We have already stated that in the $n \geq p$ case, the estimator (3.1) coincides with (2.9). In the $n < p$ case that interests us, the roles of p and n are simply reversed in the bound on the function r. The estimator still has the form

$$\left(I - \frac{r(X'S^+X)SS^+}{X'S^+X}\right)X$$

with r a nondecreasing differentiable function, but the bounds on r are

$$0 \leq r \leq \frac{2(n-2)}{p-n+3}.$$

The generalization of the standard James-Stein estimator (1.8), which corresponds to a choice of r constant, is therefore

$$\left(I - \frac{aSS^+}{X'S^+X}\right)X$$

with

$$0 \leq a \leq \frac{2(n-2)}{p-n+3}.$$

To stay consistent with previous results, a will be chosen equal to $\dfrac{n-2}{p-n+3}$ for simulations.

3.3 Technical results

In this section, we will write $F = X'S^+X$ to shorten notations. For (Y_{ij}) $(i \in \{1, ..., n\}, j \in \{1, ..., p\})$, we will also denote ∇_Y the matrix with components $(\nabla_Y)_{ij} = \frac{\partial}{\partial Y_{ij}}$.

The first result simply looks at derivatives that will be useful in other proofs.

Lemma 3.1. *Reinstating the notations used previously, Y is an $n \times p$ matrix, $S = Y'Y$ and X is a vector of size p. We have*

(i) $\left(\frac{\partial S}{\partial Y_{\alpha\beta}}\right)_{kl} = \delta_{\beta k} Y_{\alpha l} + \delta_{\beta l} Y_{\alpha k}$;

(ii) $\frac{\partial F}{\partial Y_{\alpha\beta}} = -2(X'S^+Y')_\alpha (S^+X)_\beta + 2(X'S^+S^+Y')_\alpha((I - SS^+)X)_\beta$;

(iii) $\frac{\partial (S^+XX'SS^+)_{kl}}{\partial Y_{\alpha\beta}}$

$$\begin{aligned}
=&(S^+S^+Y')_{k\alpha}((I - SS^+)XX'SS^+)_{\beta l}\\
&- S^+_{k\beta}(YS^+XX'SS^+)_{\alpha l} - (S^+Y')_{k\alpha}(S^+XX'SS^+)_{\beta l}\\
&+ (I - SS^+)_{k\beta}(YS^+S^+XX'SS^+)_{\alpha l}\\
&+ (S^+XX')_{k\beta}(YS^+)_{\alpha l} + (S^+XX'Y')_{k\alpha}(S^+)_{\beta l}\\
&+ (S^+XX'S^+Y')_{k\alpha}(I - SS^+)_{\beta l}\\
&- (S^+XX'SS^+)_{k\beta}(YS^+)_{\alpha l} - (S^+XX'SS^+Y')_{k\alpha}(S^+)_{\beta l}.
\end{aligned}$$

Proof. Showing (i) is quick:

$$\left(\frac{\partial S}{\partial Y_{\alpha\beta}}\right)_{kl} = \frac{\partial}{\partial Y_{\alpha\beta}} S_{kl} = \frac{\partial}{\partial Y_{\alpha\beta}} \sum_j Y_{jk}Y_{jl} = \delta_{\beta k} Y_{\alpha l} + \delta_{\beta l} Y_{\alpha k}.$$

We can now use it, along with what (3.3) and (3.4), to prove (ii). We have

$$\begin{aligned}
\frac{\partial F}{\partial Y_{\alpha\beta}} &= X' \frac{\partial S^+}{\partial Y_{\alpha\beta}} X\\
&= -\sum_{k,l} (X'S^+)_k \left(\frac{\partial S}{\partial Y_{\alpha\beta}}\right)_{kl} (S^+X)_l\\
&\quad + \sum_{k,l} (X'S^+S^+)_k \left(\frac{\partial S}{\partial Y_{\alpha\beta}}\right)_{kl} ((I - SS^+)X)_l\\
&\quad + \sum_{k,l} (X'(I - SS^+))k \left(\frac{\partial S}{\partial Y_{\alpha\beta}}\right)_{kl} (S^+S^+X)_l\\
&= -\sum_{k,l} (X'S^+)_k \left(\delta_{\beta k} Y_{\alpha l} + \delta_{\beta l} Y_{\alpha k}\right) (S^+X)_l
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{k,l}(X'S^+S^+)_k\left(\delta_{\beta k}Y_{\alpha l}+\delta_{\beta l}Y_{\alpha k}\right)((I-SS^+)X)_l\\
&+\sum_{k,l}(X'(I-SS^+))k\left(\delta_{\beta k}Y_{\alpha l}+\delta_{\beta l}Y_{\alpha k}\right)(S^+S^+X)_l\\
=&-\sum_{l}(X'S^+)_\beta Y_{\alpha l}(S^+X)_l-\sum_{k}(X'S^+)_k Y_{\alpha k}(S^+X)_\beta\\
&+\sum_{l}(X'S^+S^+)_\beta Y_{\alpha l}((I-SS^+)X)_l+\sum_{k}(X'S^+S^+)_k Y_{\alpha k}((I-SS^+)X)_\beta\\
&+\sum_{l}(X'(I-SS^+))_\beta Y_{\alpha l}(S^+S^+X)_l+\sum_{k}(X'(I-SS^+))_k Y_{\alpha k}(S^+S^+X)_\beta\\
=&-2(X'S^+Y')_\alpha(S^+X)_\beta+2(X'S^+S^+Y')_\alpha((I-SS^+)X)_\beta.
\end{aligned}
$$

Before proving (iii), see that thanks to (i), for any matrices A and B with appropriate dimensions, we have

$$
\begin{aligned}
\left(A\frac{\partial S}{\partial Y_{\alpha\beta}}B\right)_{kl}&=\sum_{i,j}A_{ki}\left(\frac{\partial S}{\partial Y_{\alpha\beta}}\right)_{ij}B_{jl}\\
&=\sum_{i,j}A_{ki}(\delta_{\beta i}Y_{\alpha j}+\delta_{\beta j}Y_{\alpha i})B_{jl}\\
&=A_{k\beta}(YB)_{\alpha l}+(AY')_{k\alpha}B_{\beta l}.
\end{aligned}
$$

Now for (iii), recalling (3.4) several times, we have

$$
\begin{aligned}
&\frac{\partial(S^+XX'SS^+)_{kl}}{\partial Y_{\alpha\beta}}\\
&\quad=\Bigg(S^+S^+\frac{\partial S}{\partial Y_{\alpha\beta}}(I-SS^+)XX'SS^+\\
&\qquad-S^+\frac{\partial S}{\partial Y_{\alpha\beta}}S^+XX'SS^++(I-SS^+)\frac{\partial S}{\partial Y_{\alpha\beta}}S^+S^+XX'SS^+\\
&\qquad+S^+XX'\frac{\partial S}{\partial Y_{\alpha\beta}}S^++S^+XX'SS^+S^+\frac{\partial S}{\partial Y_{\alpha\beta}}(I-SS^+)\\
&\qquad-S^+XX'SS^+\frac{\partial S}{\partial Y_{\alpha\beta}}S^++S^+XX'S(I-SS^+)\frac{\partial S}{\partial Y_{\alpha\beta}}S^+S^+\Bigg)_{kl}\\
&\quad=(S^+S^+Y')_{k\alpha}((I-SS^+)XX'SS^+)_{\beta l}\\
&\qquad-S^+_{k\beta}(YS^+XX'SS^+)_{\alpha l}-(S^+Y')_{k\alpha}(S^+XX'SS^+)_{\beta l}\\
&\qquad+(I-SS^+)_{k\beta}(YS^+S^+XX'SS^+)_{\alpha l}\\
&\qquad+(S^+XX')_{k\beta}(YS^+)_{\alpha l}+(S^+XX'Y')_{k\alpha}(S^+)_{\beta l}\\
&\qquad+(S^+XX'S^+Y')_{k\alpha}(I-SS^+)_{\beta l}\\
&\qquad-(S^+XX'SS^+)_{k\beta}(YS^+)_{\alpha l}-(S^+XX'SS^+Y')_{k\alpha}(S^+)_{\beta l}.
\end{aligned}
$$

□

Given these premises, we are now in position to state the three lemmas that were used in the proof of Theorem 3.1.

Lemma 3.2. *Under the assumptions of Theorem 3.1, we have*

$$\operatorname{tr}\left(Y\nabla_Y\left(r^2(X'S^+X)\frac{SS^+XX'S^+}{(X'S^+X)^2}\right)\right)$$
$$= -4r(X'S^+X)r'(X'S^+X) + r^2(X'S^+X)\frac{p-2\operatorname{tr}(SS^+)+3}{X'S^+X}.$$

Proof. We have

$$\left(Y\nabla_Y\left(r^2(F)\frac{SS^+XX'S^+}{F^2}\right)\right)_{ij}$$
$$= \sum_{\alpha,\beta}(Y')_{i\alpha}\frac{\partial}{\partial Y_{\alpha\beta}}\left(r^2(F)\frac{(SS^+XX'S^+)_{\beta j}}{F^2}\right)$$
$$= 2\sum_{\alpha,\beta}(Y')_{i\alpha}r(F)r'(F)\frac{\partial F}{\partial Y_{\alpha\beta}}\frac{(SS^+XX'S^+)_{\beta j}}{F^2} \tag{3.7}$$
$$+\sum_{\alpha,\beta}(Y')_{i\alpha}r^2(F)\frac{\partial(SS^+XX'S^+)_{\beta j}}{\partial Y_{\alpha\beta}}\frac{1}{F^2} \tag{3.8}$$
$$+\sum_{\alpha,\beta}(Y')_{i\alpha}r^2(F)\frac{(-2)}{F^3}\frac{\partial F}{\partial Y_{\alpha\beta}}(SS^+XX'S^+)_{\beta j}. \tag{3.9}$$

Lemma 3.1(ii) gives us

$$\sum_{\alpha,\beta}(Y')_{i\alpha}\frac{\partial F}{\partial Y_{\alpha\beta}}(SS^+XX'S^+)_{\beta j}$$
$$= -2\sum_{\alpha,\beta}(X'S^+Y')_{\alpha}Y_{\alpha i}(S^+X)_{\beta}(SS^+XX'S^+)_{\beta j}$$
$$+2\sum_{\alpha,\beta}(X'S^+S^+Y')_{\alpha}Y_{\alpha i}(S^+XX'SS^+)_{\beta j}((I-SS^+)X)_{\beta}$$
$$= -2(X'S^+Y'Y)_i(S^+XSS^+XX'S^+)_j + 2(X'S^+S^+Y'Y)_i((S^+XX'S\underbrace{S^+(I-SS^+)}_{=0}X)_j$$
$$= -2X'S^+X(SS^+XX'S^+)_{ij} = -2F(SS^+XX'S^+)_{ij}.$$

This simplifies (3.7) into

$$2\sum_{\alpha,\beta}(Y')_{i\alpha}r(F)r'(F)\frac{\partial F}{\partial Y_{\alpha\beta}}\frac{(SS^+XX'S^+)_{\beta j}}{F^2}$$
$$= -4r(F)r'(F)\frac{(SS^+XX'S^+)_{ij}}{F}$$

and (3.9) into

$$\begin{aligned}\sum_{\alpha,\beta}(Y')_{i\alpha}r^2(F)\frac{(-2)}{F^3}\frac{\partial F}{\partial Y_{\alpha\beta}}(SS^+XX'S^+)_{\beta j}\\ = 4r^2(F)\frac{(SS^+XX'S^+)_{ij}}{F^2}.\end{aligned}$$

For (3.8), Lemma 3.1(iii) is first used to get

$$\begin{aligned}&\sum_{\alpha,\beta}(Y')_{i\alpha}\frac{\partial(SS^+XX'S^+)_{\beta j}}{\partial Y_{\alpha\beta}}\\
&\quad=\sum_{\alpha,\beta}(Y')_{i\alpha}\frac{\partial(S^+XX'SS^+)_{j\beta}}{\partial Y_{\alpha\beta}}\\
&\quad=\sum_{\alpha,\beta}\Big((S^+S^+Y')_{j\alpha}Y_{\alpha i}((I-SS^+)XX'SS^+)_{\beta\beta}\\
&\qquad\quad-S^+_{j\beta}(Y')_{i\alpha}(YS^+XX'SS^+)_{\alpha\beta}\\
&\qquad\quad-(S^+Y')_{j\alpha}Y_{\alpha i}(S^+XX'SS^+)_{\beta\beta}\\
&\qquad\quad+(I-SS^+)_{j\beta}(Y')_{i\alpha}(YS^+S^+XX'SS^+)_{\alpha\beta}\\
&\qquad\quad+(S^+XX')_{j\beta}(Y')_{i\alpha}(YS^+)_{\alpha\beta}+(S^+XX'Y')_{k\alpha}(S^+)_{\beta\beta}\\
&\qquad\quad+(S^+XX'S^+Y')_{j\alpha}Y_{\alpha i}(I-SS^+)_{\beta\beta}\\
&\qquad\quad-(S^+XX'SS^+)_{j\beta}(Y')_{i\alpha}(YS^+)_{\alpha\beta}\\
&\qquad\qquad\quad-(S^+XX'SS^+Y')_{j\alpha}Y_{\alpha i}(S^+)_{\beta\beta}\Big)\\
&\quad=(S^+XX'SS^+(I-SS^+))_{ij}\\
&\qquad-(SS^+XX'S^+)_{ij}-\operatorname{tr}(S^+XX'SS^+)(SS^+)_{ij}\\
&\qquad+\operatorname{tr}((I-SS^+)XX'SS^+)(S^+)_{ij}\\
&\qquad+(SS^+XX'S^+)_{ij}+\operatorname{tr}(S^+)(SXX'S^+)_{ij}\\
&\qquad+\operatorname{tr}(I-SS^+)(SS^+XX'S^+)_{ij}\\
&\qquad-(SS^+XX'S^+)_{ij}-\operatorname{tr}(S^+)(SXX'S^+)_{ij}\\
&\quad=\big(p-\operatorname{tr}(SS^+)-1\big)(SS^+XX'S^+)_{ij}-(X'S^+X)(SS^+)_{ij}.\end{aligned}$$

That gives us the following expression

$$\begin{aligned}&\sum_{\alpha,\beta}(Y')_{i\alpha}r^2(F)\frac{\partial(SS^+XX'S^+)_{\beta j}}{\partial Y_{\alpha\beta}}\frac{1}{F^2}\\
&\qquad=\big(p-\operatorname{tr}(SS^+)-1\big)r^2(F)\frac{(SS^+XX'S^+)_{ij}}{F^2}-r^2\frac{(SS^+)_{ij}}{F}\end{aligned}$$

Combining the three expressions obtained completes the proof, as we have

$$\operatorname{tr}\left(Y\nabla_Y\left(r^2(X'S^+X)\frac{SS^+XX'S^+}{(X'S^+X)^2}\right)\right)$$

$$
\begin{aligned}
&= \sum_i \Big(-4r(F)r'(F)\frac{(SS^+XX'S^+)_{ii}}{F} \\
&\qquad + \big(p - \operatorname{tr}(SS^+) - 1\big)r^2(F)\frac{(SS^+XX'S^+)_{ij}}{F^2} - r^2\frac{(SS^+)_{ii}}{F} \\
&\qquad + 4r^2(F)\frac{(SS^+XX'S^+)_{ii}}{F^2}\Big) \\
&= -4r(F)r'(F) + r^2(F)\frac{p - 2\operatorname{tr}(SS^+) + 3}{F}
\end{aligned}
$$

because $\operatorname{tr}(SS^+XX'S^+) = \operatorname{tr}(X'S^+X) = X'S^+X = F$. □

Lemma 3.3. *Under the assumptions of Theorem 3.1, we have*

$$
\operatorname{div}_X \frac{r(X'S^+X)SS^+X}{X'S^+X} = 2r'(X'S^+X) + r(X'S^+X)\frac{\operatorname{tr}(SS^+) - 2}{X'S^+X}.
$$

Proof. We simply compute

$$
\begin{aligned}
&\operatorname{div}_X \left(r(F)\frac{SS^+X}{F} \right) \\
&= \sum_i \frac{\partial}{\partial X_i}\left(r(F)\frac{(SS^+X)_i}{F} \right) \\
&= \sum_i \Bigg(r'(F)\frac{\partial F}{\partial X_i}\frac{(SS^+X)_i}{F} \\
&\qquad + r(F)\frac{\partial (SS^+X)_i}{\partial X_i}\frac{1}{F} - r(F)\frac{1}{F^2}\frac{\partial F}{\partial X_i}(SS^+X)_i \Bigg) \\
&= \sum_i \Bigg(r'(F)\left(\frac{\partial}{\partial X_i}\sum_{k,l} X_k X_l S^+_{kl} \right)\frac{(SS^+X)_i}{F} \\
&\qquad + \frac{r(F)}{F}\frac{\partial\left(\sum_k (SS^+)_{ik}X_k\right)}{\partial X_i} \\
&\qquad - \frac{r(F)}{F^2}\frac{\partial\left(\sum_{k,l} X_k X_l S^+_{kl}\right)}{\partial X_i}(SS^+X)_i \Bigg) \\
&= \sum_i \Bigg(r'(F)\big((X'S^+)_i + (X'S^+)_i\big)\frac{(SS^+X)_i}{F} \\
&\qquad + r(F)\frac{(SS^+)_{ii}}{F} - r(F)\frac{\big((X'S^+)_i + (X'S^+)_i\big)(SS^+X)_i}{F} \Bigg) \\
&= 2r'(F) + r(F)\frac{\operatorname{tr}(SS^+) - 2}{F}
\end{aligned}
$$

which is what was announced.

□

Lemma 3.4. *Using the notations from the beginning of section* 3*, let* Y *be a normal matrix* $\mathcal{N}_{n\times p}(0, I\otimes\Sigma)$*, so that* $S = Y'Y$ *follows a Wishart* $\mathcal{W}_p(n,\Sigma)$*. Let* $G(S)$ *be a* $p\times p$ *random matrix depending on* S *and, with* A *the symmetric positive definite square root of* Σ*, define* $\tilde{Y} = YA^{-1}$ *and* $H = AGA^{-1}$*. Then under the condition*

$$\mathbb{E}\left[|\operatorname{div}_{\operatorname{vec}(\tilde{Y})}\operatorname{vec}(\tilde{Y}H)|\right] < \infty, \tag{3.10}$$

we have

$$\mathbb{E}\left[\operatorname{tr}(\Sigma^{-1}SG)\right] = \mathbb{E}\left[n\operatorname{tr}(G) + \operatorname{tr}(Y'\nabla_Y G')\right].$$

Proof. Notice that $\tilde{Y} \sim \mathcal{N}_{n\times p}(0, I_n\otimes I_p)$, which means that $\operatorname{vec}(\tilde{Y}) \sim \mathcal{N}_{np}(0, I_{np})$. Denoting $\tilde{S} = \tilde{Y}'\tilde{Y} = A^{-1}SA^{-1}$, we can write

$$\begin{aligned}\mathbb{E}\left[\operatorname{tr}(\tilde{S}H)\right] &= \mathbb{E}\left[\sum_{\alpha,i,j}\tilde{Y}_{\alpha i}\tilde{Y}_{\alpha j}H_{ji}\right]\\ &= \mathbb{E}\left[\operatorname{vec}(\tilde{Y})\cdot\operatorname{vec}(\tilde{Y}H)\right]\end{aligned}$$

Now using Lemma 2.1, we find that

$$\begin{aligned}\mathbb{E}\left[\operatorname{vec}(\tilde{Y})\cdot\operatorname{vec}(\tilde{Y}H)\right] &= \mathbb{E}\left[\operatorname{div}_{\operatorname{vec}(\tilde{Y})}\operatorname{vec}(\tilde{Y}H)\right]\\ &= \mathbb{E}\left[\sum_{\alpha,i,j}\frac{\partial}{\partial\tilde{Y}_{\alpha i}}\tilde{Y}_{\alpha j}H_{ji}\right]\\ &= \mathbb{E}\left[\sum_{\alpha,i,j}\left(\delta_{ij}H_{ji} + \tilde{Y}_{\alpha j}\frac{\partial H_{ji}}{\partial\tilde{Y}_{\alpha i}}\right)\right]\\ &= \mathbb{E}\left[n\sum_i H_{ii} + \tilde{Y}_{\alpha j}\frac{\partial H_{ji}}{\partial\tilde{Y}_{\alpha i}}\right].\end{aligned}$$

In matrix notation, this means that

$$\mathbb{E}\left[\operatorname{tr}(\tilde{S}H)\right] = \mathbb{E}\left[n\operatorname{tr}(H) + \operatorname{tr}((\tilde{Y}'\nabla_{\tilde{Y}})'H)\right].$$

This concludes the proof, as

$$\begin{aligned}\mathbb{E}\left[\operatorname{tr}(H)\right] &= \mathbb{E}\left[\operatorname{tr}(AGA^{-1})\right] = \mathbb{E}\left[\operatorname{tr}(G)\right],\\ \mathbb{E}\left[\operatorname{tr}(\tilde{S}H)\right] &= \mathbb{E}\left[\operatorname{tr}(A^{-1}SGA^{-1})\right] = \mathbb{E}\left[\operatorname{tr}(\Sigma^{-1}SG)\right],\\ \mathbb{E}\left[\operatorname{tr}((\tilde{Y}'\nabla_{\tilde{Y}})'H)\right] &= \mathbb{E}\left[\operatorname{tr}(A(Y'\nabla_Y)'GA^{-1})\right] = \mathbb{E}\left[\operatorname{tr}((Y'\nabla_Y)'G)\right].\end{aligned}$$

□

Finally, we look at the theorem needed to verify the condition for Lemma 3.4.

Theorem 3.2. *Let* $Y \sim \mathcal{N}_{n\times p}(0, I \otimes \Sigma)$ *and* $\tilde{Y} = YA^{-1}$ *with* A *the symmetric positive definite square root of* Σ. *With a differentiable function* $r : \mathbb{R} \to [0, C_1]$ *such that* $|r'| \leq C_2$ *(with* $C_1, C_2 \in \mathbb{R}_0^+$*), define*

$$G = r^2(X'S^+X)\frac{S^+XX'S^+S}{(X'S^+X)^2}$$

and $H = AGA^{-1}$. *Then, for all* p *and* n,

$$\mathbb{E}\left[|\operatorname{div}_{\mathrm{vec}(\tilde{Y})}\operatorname{vec}(\tilde{Y}H)|\right] < \infty. \tag{3.11}$$

Proof. We begin by looking at $\operatorname{div}_{\mathrm{vec}(\tilde{Y})}\operatorname{vec}(\tilde{Y}H)$. We have

$$\begin{aligned}
&\operatorname{div}_{\mathrm{vec}(\tilde{Y})}\operatorname{vec}(\tilde{Y}H)\\
&\quad= \sum_{\alpha,i,j}\frac{\partial}{\partial\tilde{Y}_{\alpha i}}(\tilde{Y}_{\alpha j}H_{j,i})\\
&\quad= n\sum_i H_{ii} + \sum_{\alpha,i,j}\tilde{Y}_{\alpha j}\frac{\partial H_{j,i}}{\partial\tilde{Y}_{\alpha i}}\\
&\quad= n\sum_i H_{ii} + \sum_{\alpha,i,j}\tilde{Y}_{\alpha j}\sum_\beta\frac{\partial Y_{\alpha\beta}}{\partial\tilde{Y}_{\alpha i}}\frac{\partial(AGA^{-1})_{j,i}}{\partial Y_{\alpha\beta}}\\
&\quad= n\sum_i H_{ii} + \sum_{\alpha,\beta,i,j}\tilde{Y}_{\alpha j}A_{\beta i}\frac{\partial}{\partial Y_{\alpha\beta}}\left(r^2(F)\frac{(AS^+XX'SS^+A^{-1})_{ji}}{(F)^2}\right)\\
&\quad= n\sum_i H_{ii} + \sum_{\alpha,\beta,i,j}\tilde{Y}_{\alpha j}A_{\beta i}\\
&\qquad\qquad\times\left(2r(F)r'(F)\frac{\partial F}{\partial Y_{\alpha\beta}}\frac{(AS^+XX'SS^+A^{-1})_{ji}}{(F)^2}\right. && (3.12)\\
&\qquad\qquad\qquad+\frac{r^2(F)}{F^2}\sum_{k,l}A_{jk}\frac{\partial(S^+XX'SS^+)_{kl}}{\partial Y_{\alpha\beta}}A^{-1}_{li} && (3.13)\\
&\qquad\qquad\qquad\left.-\,r^2(F)(AS^+XX'SS^+A^{-1})_{ji}2\frac{1}{F^3}\frac{\partial F}{\partial Y_{\alpha\beta}}\right). && (3.14)
\end{aligned}$$

We now develop each part separately.

First, for (3.12), we use Lemma 3.1(2) to get

$$\begin{aligned}
&2\sum_{\alpha,\beta,i,j}\tilde{Y}_{\alpha j}A_{\beta i}r(F)r'(F)\frac{\partial F}{\partial Y_{\alpha\beta}}\frac{(AS^+XX'SS^+A^{-1})_{ji}}{(F)^2}\\
&\quad= 4\frac{r(F)r'(F)}{F^2}\\
&\qquad\times\sum_{\alpha,\beta,i,j}\big(-(X'S^+Y')_\alpha\tilde{Y}_{\alpha j}(AS^+XX'SS^+A^{-1})_{ji}A_{i\beta}(S^+X)_\beta
\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad + (X'S^+S^+Y')_\alpha \tilde{Y}_{\alpha j}(AS^+XX'SS^+A^{-1})_{ji}A_{i\beta}((I-SS^+)X)_\beta\Big) \\
&= -4\frac{r(F)r'(F)}{F^2}(X'S^+Y'YA^{-1}AS^+XX'SS^+A^{-1}AS^+X) \\
&\quad + 4\frac{r(F)r'(F)}{F^2}(X'S^+S^+Y'YA^{-1}AS^+XX'SS^+A^{-1}A(I-SS^+)X) \\
&= -4\frac{r(F)r'(F)}{F^2}(X'S^+SS^+XX'S^+SS^+X) \\
&\quad + 4\frac{r(F)r'(F)}{F^2}(X'S^+S^+SS^+XX'S\,\underbrace{S^+(I-SS^+)}_{=0}X) \\
&= -4r(F)r'(F).
\end{aligned}
$$

Similarly, for (3.14), we get

$$
\begin{aligned}
&\sum_{\alpha,\beta,i,j} \tilde{Y}_{\alpha j}A_{\beta i}r^2(F)(AS^+XX'SS^+A^{-1})_{ji}2\frac{1}{F^3}\frac{\partial F}{\partial Y_{\alpha\beta}} \\
&\qquad = 4\frac{r^2(F)}{F^3}\sum_{\alpha,\beta,i,j}(X'S^+Y')_\alpha\tilde{Y}_{\alpha j}(AS^+XX'SS^+A^{-1})_{ji}A_{i\beta}(S^+X)_\beta \\
&\qquad = 4\frac{r^2(F)}{F^3}(X'S^+Y'YA^{-1}AS^+XX'SS^+A^{-1}AS^+X) \\
&\qquad = 4\frac{r^2(F)}{F^3}(X'S^+SS^+XX'S^+SS^+X) \\
&\qquad = 4\frac{r^2(F)}{F}.
\end{aligned}
$$

The last part left is (3.13). Using Lemma 3.1(3) this time, we have

$$
\begin{aligned}
&\sum_{\alpha,\beta,i,j}\tilde{Y}_{\alpha j}A_{\beta i}\frac{r^2(F)}{F^2}\sum_{k,l}A_{jk}\frac{\partial(S^+XX'SS^+)_{kl}}{\partial Y_{\alpha\beta}}A^{-1}_{li} \\
&\quad = \frac{r^2(F)}{F^2}\sum_{\alpha,\beta,i,j,k,l}\tilde{Y}_{\alpha j}A_{\beta i}A_{jk}A^{-1}_{li} \\
&\qquad\qquad \times\Big((S^+S^+Y')_{k\alpha}((I-SS^+)XX'SS^+)_{\beta l} \\
&\qquad\qquad - S^+_{k\beta}(YS^+XX'SS^+)_{\alpha l} \\
&\qquad\qquad - (S^+Y')_{k\alpha}(S^+XX'SS^+)_{\beta l} \\
&\qquad\qquad + (I-SS^+)_{k\beta}(YS^+S^+XX'SS^+)_{\alpha l} \\
&\qquad\qquad + (S^+XX')_{k\beta}(YS^+)_{\alpha l} \\
&\qquad\qquad + (S^+XX'Y')_{k\alpha}(S^+)_{\beta l} \\
&\qquad\qquad + (S^+XX'S^+Y')_{k\alpha}(I-SS^+)_{\beta l} \\
&\qquad\qquad - (S^+XX'SS^+)_{k\beta}(YS^+)_{\alpha l}
\end{aligned}
$$

$$
\begin{aligned}
&\qquad\qquad - (S^+XX'SS^+Y')_{k\alpha}(S^+)_{\beta l}\Big) \\
&= \frac{r^2(F)}{F^2} \sum_{\alpha,\beta,i,j,k,l} \Big(A_{jk}(S^+S^+Y')_{k\alpha}\tilde{Y}_{\alpha j}A_{i\beta}((I-SS^+)XX'SS^+)_{\beta l}A^{-1}_{li} \\
&\qquad - \tilde{Y}'_{j\alpha}(YS^+XX'SS^+)_{\alpha l}A^{-1}_{li}A_{i\beta}S^+_{\beta k}A_{kj} \\
&\qquad - A_{jk}(S^+Y')_{k\alpha}\tilde{Y}_{\alpha j}A_{i\beta}(S^+XX'SS^+)_{\beta l}A^{-1}_{li} \\
&\qquad + \tilde{Y}'_{j\alpha}(YS^+S^+XX'SS^+)_{\alpha l}A^{-1}_{li}A_{i\beta}(I-SS^+)_{\beta k}A_{kj} \\
&\qquad + \tilde{Y}'_{j\alpha}(YS^+)_{\alpha l}A^{-1}_{li}A_{i\beta}(XX'S^+)_{\beta k}A_{kj} \\
&\qquad + A_{jk}(S^+XX'Y')_{k\alpha}\tilde{Y}_{\alpha j}A_{i\beta}(S^+)_{\beta l}A^{-1}_{li} \\
&\qquad + A_{jk}(S^+XX'S^+Y')_{k\alpha}\tilde{Y}_{\alpha j}A_{i\beta}(I-SS^+)_{\beta l}A^{-1}_{li} \\
&\qquad - \tilde{Y}'_{j\alpha}(YS^+)_{\alpha l}A^{-1}_{li}A_{i\beta}(SS^+XX'S^+)_{\beta k}A_{kj} \\
&\qquad\qquad - A_{jk}(S^+XX'SS^+Y')_{k\alpha}\tilde{Y}_{\alpha j}A_{i\beta}(S^+)_{\beta l}A^{-1}_{li}\Big) \\
&= \frac{r^2(F)}{F^2}\Big(\operatorname{tr}(AS^+S^+Y'YA^{-1})\operatorname{tr}(A(I-SS^+)XX'SS^+A^{-1}) \\
&\qquad - \operatorname{tr}(A^{-1}Y'YS^+XX'SS^+A^{-1}AS^+A) \\
&\qquad - \operatorname{tr}(AS^+Y'YA^{-1})\operatorname{tr}(AS^+XX'SS^+A^{-1}) \\
&\qquad + \operatorname{tr}(A^{-1}Y'YS^+S^+XX'SS^+A^{-1}A(I-SS^+)A) \\
&\qquad + \operatorname{tr}(A^{-1}Y'YS^+A^{-1}AXX'S^+A) \\
&\qquad + \operatorname{tr}(AS^+XX'Y'YA^{-1})\operatorname{tr}(AS^+A^{-1}) \\
&\qquad + \operatorname{tr}(AS^+XX'S^+Y'YA^{-1})\operatorname{tr}(A(I-SS^+)A^{-1}) \\
&\qquad - \operatorname{tr}(A^{-1}Y'YS^+A^{-1}ASS^+XX'S^+A) \\
&\qquad\qquad - \operatorname{tr}(AS^+XX'SS^+Y'YA^{-1})\operatorname{tr}(AS^+A^{-1})\Big) \\
&= \frac{r^2(F)}{F^2}\Big(\operatorname{tr}(S^+SS^+)\operatorname{tr}(S^+(I-SS^+)XX'S) \\
&\qquad - \operatorname{tr}(X'\underbrace{SS^+S^+SS^+}_{=S^+}X) \\
&\qquad - \operatorname{tr}(S^+S)\operatorname{tr}(X'S^+SS^+X) \\
&\qquad + \operatorname{tr}(SS^+S^+XX'SS^+(I-SS^+)) \\
&\qquad + \operatorname{tr}(X'S^+SS^+X) \\
&\qquad + \operatorname{tr}(X'SS^+X)\operatorname{tr}(S^+) \\
&\qquad + \operatorname{tr}(S^+XX'S^+Y'Y)\operatorname{tr}(A(I-SS^+)A^{-1}) \\
&\qquad - \operatorname{tr}(X'\underbrace{S^+SS^+SS^+}_{=S^+}X) \\
&\qquad\qquad - \operatorname{tr}(X'SS^+SS^+X)\operatorname{tr}(S^+)\Big)
\end{aligned}
$$

$$\begin{aligned}
&= \frac{r^2(F)}{F^2}\Big(- X'S^+X - \operatorname{tr}(SS^+)X'S^+X + X'S^+X + X'SS^+X\operatorname{tr}(S^+) \\
&\qquad\qquad + X'S^+X(p - tr(SS^+)) - X'S^+X - X'SS^+X\operatorname{tr}(S^+)\Big) \\
&= \frac{r^2(F)}{F}(p - 2\operatorname{tr}(SS^+) - 1).
\end{aligned}$$

We can finally put everything back into the expression for $\operatorname{div}_{\operatorname{vec}(\tilde{Y})}\operatorname{vec}(\tilde{Y}H)$, that we need to bound:

$$\begin{aligned}
&\mathbb{E}\left[|\operatorname{div}_{\operatorname{vec}(\tilde{Y})}\operatorname{vec}(\tilde{Y}H)|\right] \\
&= \mathbb{E}\left[\left|n\operatorname{tr}(H) + 4\frac{r^2(F)}{F} + (p - 2\operatorname{tr}(SS^+) - 1)\frac{r^2(F)}{F} - 4r(F)r'(F)\right|\right] \\
&\leq C_1^2\left|3 + p - 2\operatorname{tr}(SS^+) + n\right|\mathbb{E}\left[\frac{1}{F}\right] + 4C_1C_2. \qquad (3.15)
\end{aligned}$$

For the bound to be finite, we still need to show that $\mathbb{E}\left[\frac{1}{F}\right] < \infty$. As $S \sim \mathcal{W}_p(n, \Sigma)$, we can define $T \sim \mathcal{W}_p(n, I)$ such that $S = ATA$. Let the spectral decomposition of T be $T = H'DH$ with $D = \operatorname{diag}(\lambda_i)$. Writing the eigenvalues of T^+ as λ_i^+, and $\lambda_{\min}^+$ for the smallest nonzero one, we have that $D^{-1} = \operatorname{diag}(\lambda_i^+)$. We will need the following identity on Moore-Penrose inverses of products, following from [16] (Thm 1.1, equations (1.2) and (1.4)) and symmetry of T (more details in the appendix):

$$(ATA)^+ = (T^+TA)^+T^+(AT^+T)^+.$$

With this, we find

$$\begin{aligned}
X'S^+X = X'(ATA)^+X &= X'(T^+TA)^+T^+(AT^+T)^+X \\
&= \sum_k \left(X'(T^+TA)^+H'\right)_k^2\lambda_k^+ \\
&\geq \lambda_{\min}^+ X'(T^+TA)^+H'H(AT^+T)^+X \\
&= \lambda_{\min}^+ X'(T^+TA)^+(AT^+T)^+X
\end{aligned}$$

This can be bounded using the Cauchy-Schwarz inequality:

$$\begin{aligned}
& X'(T^+TA)^+(AT^+T)^+X \leq X'(T^+TA)^+(AT^+T)^+XX'(AT^+T)(T^+TA)X. \\
\Leftrightarrow\quad & \frac{X'AT^+TAX}{X'(T^+TA)^+(AT^+T)^+X} \geq \frac{1}{X'(T^+TA)^+(AT^+T)^+X}
\end{aligned}$$

We then have

$$\frac{1}{F} = \frac{1}{X'S^+X} \leq \frac{1}{\lambda_{\min}^+}\frac{1}{X'(T^+TA)^+(AT^+T)^+X}$$

$$\leq \frac{1}{\lambda_{\min}^{+}} \frac{X'AT^{+}TAX}{X'(T^{+}TA)^{+}(AT^{+}T)^{+}X}.$$

Writing $Q = AT^{+}TA$ and $R = (T^{+}TA)^{+}(AT^{+}T)^{+}$ to shorten notations, the bound in (3.15) becomes

$$C_1^2 \left|3 + p - 2\operatorname{tr}(SS^{+}) + n\right| \mathbb{E}\left[\frac{1}{\lambda_{\min}^{+}} \frac{X'QX}{X'RX}\right] + 4C_1C_2. \tag{3.16}$$

To finish, we will split what is inside the expectation using some independence results.

When $n \geq p$, we are in the standard Wishart case, T has full rank and is invertible and therefore $Q = R = I$. The independence between $\lambda_{\min}^{+}$ and $\dfrac{X'QX}{X'RX}$ then follows directly from the independence between S and X. When $n < p$, T is a singular Wishart matrix. In this case, results [11] show that T can be written as $T = H_1'D_1H_1$, where H_1 is semi-orthogonal ($H_1H_1' = I$) and D is a diagonal matrix with only the positive eigenvalues of T, and provide the joint density of H_1 and $D_1 = \operatorname{diag}(d_i)$:

$$\begin{aligned} &f_{H_1,D_1}(H_1, D_1) \\ &\quad = K(p,n)|D_1|^{(n-p-1)/2}\left[\operatorname{etr}\left(-\frac{1}{2}D_1\right)\right]\left[\prod_{i<j}(d_i - d_j)\right] g_{n,p}(H_1) \end{aligned}$$

with $K(p,n)$ a constant, $g_{p,n}$ a function and where $\operatorname{etr}(A)$ denotes the exponential of the trace of A. This means that H_1 and D_1 are independent. As $\lambda_{\min}^{+}$ is a function of D_1 and we can write $T^{+}T = H_1'H_1$ (because $T^{+} = H_1'D_1^{-1}H_1$), $\lambda_{\min}^{+}$ and $\dfrac{X'QX}{X'RX}$ are independent in this case too.

From how the Moore-Penrose pseudoinverse can be obtained through singular value decomposition, as was explained in Section 3.1, it is easy to see that the nonzero eigenvalues of T^{+} are the inverses of the nonzero eigenvalues of T, as the spectral decomposition is a singular value decomposition in our case. Therefore, denoting $\lambda_{\max}$ the largest eigenvalue of T, we have $\lambda_{\max} = \dfrac{1}{\lambda_{\min}^{+}}$ and (3.16) now gives the bound

$$C_1^2|3 + p - 2\operatorname{tr}(SS^{+}) + n|\,\mathbb{E}\left[\lambda_{\max}\right]\mathbb{E}\left[\frac{X'QX}{X'RX}\right] + 4C_1C_2. \tag{3.17}$$

As T is positive semi-definite, $\mathbb{E}\left[\lambda_{\max}\right] \leq \mathbb{E}\left[\operatorname{tr}(T)\right]$. When $n \geq p$, the trace of a Wishart matrix follows a Chi-square distribution. More precisely, $\operatorname{tr}(T) \sim \chi^2_{pn}$ (See in the appendix for results about Wishart matrices). When $n < p$, by definition of Wishart matrices, we can still write $T = Z'Z$ with $Z \sim \mathcal{N}_{n \times p}(0, I_n \otimes I_p)$. We can then switch their places and $ZZ' \sim \mathcal{W}_n(p, I_n)$ so that we also get $\operatorname{tr}(T) = \operatorname{tr}(ZZ') \sim \chi^2_{pn}$. So, in both cases, $\mathbb{E}\left[\lambda_{\max}\right] \leq \mathbb{E}\left[\operatorname{tr}(T)\right] = pn < \infty$.

We only need to check that the other expectation in (3.17) is also finite. Let $r = \operatorname{rk}(R) = \operatorname{rk}(Q) = \operatorname{rk}(S)$. We write the spectral decomposition of $(T^{+}TA)$ as $U\Lambda U'$ with $\Lambda =$

$\mathrm{diag}(L, 0_{(p-r)})$, where L is a vector containing the r nonzero eigenvalues of (T^+TA). We get that $R = (T^+TA)^+(AT^+T)^+ = U\mathrm{diag}(1_{(r)}, 0_{(p-r)})$. We define $E = U\left[0_{(p-r)\times r} I_{(p-r)}\right]'$, a $p \times (p-r)$ matrix such that $RE = 0$ and E has full column rank $p-r$. We also have that $QE = AT^+TAU\left[0_{(p-r)\times r} I_{(p-r)}\right]' = AU\Lambda U'U\left[0_{(p-r)\times r} I_{(p-r)}\right]' = 0$. Along with the fact that R and Q are symmetric and positive semidefinite, this allows us to use a result in [10](Theorem 1(i)) to conclude that

$$\mathbb{E}\left[\frac{X'QX}{X'RX}\right] < \infty.$$

The proof is now complete. □

Chapter 4

Simulations

The simulations were done with `R` (the code is in the appendix) and explore the different situations studied up to here. Before presenting the results, we introduce a simple modification that can be made to the James-Stein estimators we've encountered. For example, looking at the first form of the estimator (1.2) from Chapter 1,

$$\left(1 - \frac{p-2}{\|X\|^2}\right) X, \tag{4.1}$$

when $\|X\|^2$ is small, the term that multiply X can be negative which seems to cause the estimator to behave badly. A simple workaround is to force the coefficient to be positive, by restricting it to its positive part. This is further justified by the following result.

Proposition 4.1. *Suppose $\theta = \mathbb{E}[X]$ is estimated by $h(X)X$, where $P(h(X) < 0) > 0$. Under the additional assumption that $\mathbb{E}[X_k | h(X) < 0]$ has the same sign as θ_k for all $k \in \{1, ..., p\}$ ($p = \dim(X)$), taking instead the estimator $h_+(X)X$, where $h_+(X) = \max(0, h(X))$, we have $\mathrm{MSE}(h_+(X)X) \leq \mathrm{MSE}(h(X)X)$.*

Proof. To see this, let us look at the contribution of $h(X)$ in the MSE. We have, looking at components individually,

$$\mathbb{E}\left[(h(X)X_k - \theta_k)^2\right] = \mathbb{E}\left[h^2(X)X_k^2\right] - 2\theta_k \mathbb{E}\left[h(X)X_k\right] + \theta_k^2. \tag{4.2}$$

We can ignore the θ_k^2 term and focus on the rest. Conditioning with respect to the sign of $h(X)$, we have

$$\begin{aligned} &\mathbb{E}\left[h^2(X)X_k^2\right] - 2\theta_k \mathbb{E}\left[h(X)X_k\right] \\ &\quad = \left(\mathbb{E}\left[h^2(X)X_k^2 | h(X) \geq 0\right] - 2\theta_k \mathbb{E}\left[h(X)X_k | h(X) \geq 0\right]\right) P(h(X) \geq 0) \\ &\qquad + \left(\mathbb{E}\left[h^2(X)X_k^2 | h(X) < 0\right] - 2\theta_k \mathbb{E}\left[h(X)X_k | h(X) < 0\right]\right) P(h(X) < 0). \end{aligned}$$

As we propose to replace $h(X)$ by $h_+(X)$, the only difference occurs when $h(X) < 0$, so we only have to look at the second part. First, we have

$$\mathbb{E}\left[h^2(X)X_k^2 | h(X) < 0\right] \geq 0. \tag{4.3}$$

Now we use the additional assumption that $\mathbb{E}\left[X_k|h(X)<0\right]$ has the same sign as θ_k, to see, by conditioning, that

$$-2\theta_k\,\mathbb{E}\left[h(X)X_k|h(X)<0\right]>0,$$

which means that (4.2) would be smaller if we replaced $h(X)$ by 0 whenever it is negative, whence the domination of the estimator $h_+(X)X$. □

In the case of our simple James-Stein estimator (4.1), $h(X)=\left(1-\frac{b}{\|X\|^2}\right)$ and the additional assumption is verified. Indeed,

$$\begin{aligned}\mathbb{E}\left[X_k|h(X)<0\right]&=\mathbb{E}\left[X_k\big|\|X\|^2<b\right]\\&=\mathbb{E}\left[\mathbb{E}\left[X_k|X_k^2\right]\Big|\sum_{i\neq k}X_i^2<b-X_k^2\right]\\&=\mathbb{E}\left[|X_k|P\left(X_k>0|X_k^2\right)-|X_k|P\left(X_k<0|X_k^2\right)\Big|\sum_{i\neq k}X_i^2<b-X_k^2\right].\end{aligned}\tag{4.4}$$

This has the same sign as θ_k because, if $\theta_k>0$, then $P\left(X_k>0|X_k^2\right)>P\left(X_k<0|X_k^2\right)$ and therefore (4.4) is also positive. Similarly, if $\theta_k<0$, (4.4) is also negative.

This yields the following estimator, commonly called *positive-part* James-Stein estimator,

$$\left(1-\frac{p-2}{\|X\|^2}\right)_+X.\tag{4.5}$$

Simulations will show that the improvement it brings can be quite substantial.

To follow the same order as the presented results, we start by looking at the multinormal case with the identity as the covariance matrix. For each graphs, a sample of size 50000 was generated in order to estimate the MSE.

The first figure represents the case where the mean vector is equal to 0, with p that varies. As expected, the observed value X has an MSE equal to the dimension. In this case, the James-Stein estimator completely cancels the linear increase of the MSE to give a constant value of about 2.5, and even slightly better for the positive-part version.

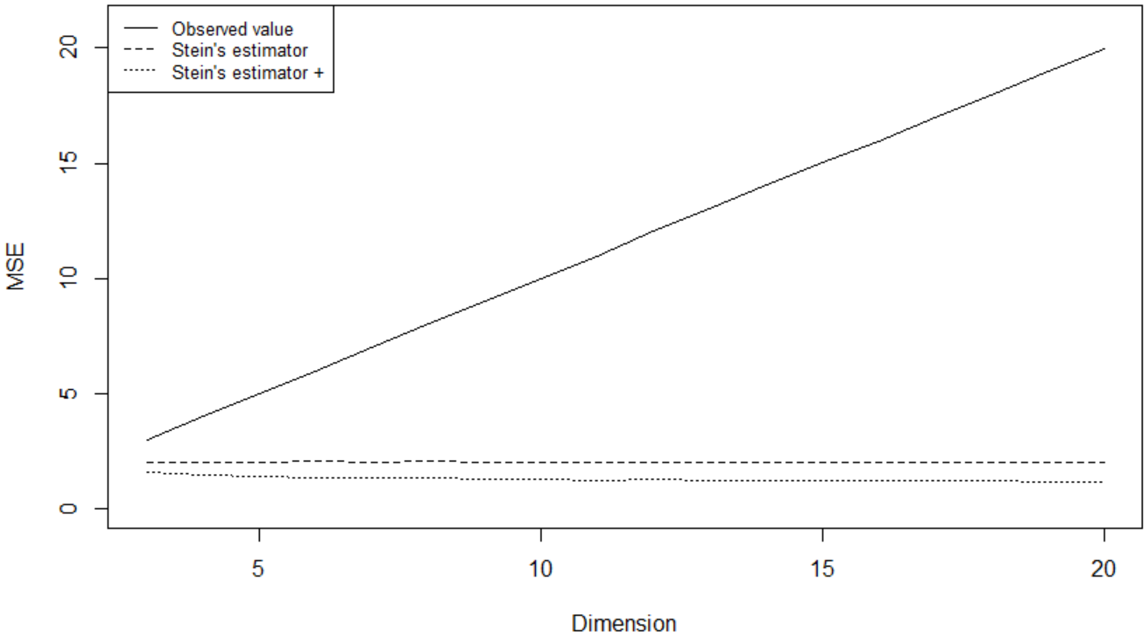

Figure 4.1: MSE when the dimension increases, with the mean equal to $(0, ..., 0)$

While the improvement is quite amazing, it is also pretty natural, as the James-Stein estimator can be seen as a "shrinkage" towards the origin, so when the mean is also at the origin, this improvement was to be expected. Figure 4.2 illustrates this: a vector of size 20 (above) is compared to its James-Stein estimator (below). The reasons for the improvement are clear. This also explains in parts the bad behaviour when the multiplicative term is negative: if $\|X\|^2$ gets too small, it can become greater than 1 in absolute value, and we no longer have a shrinkage.

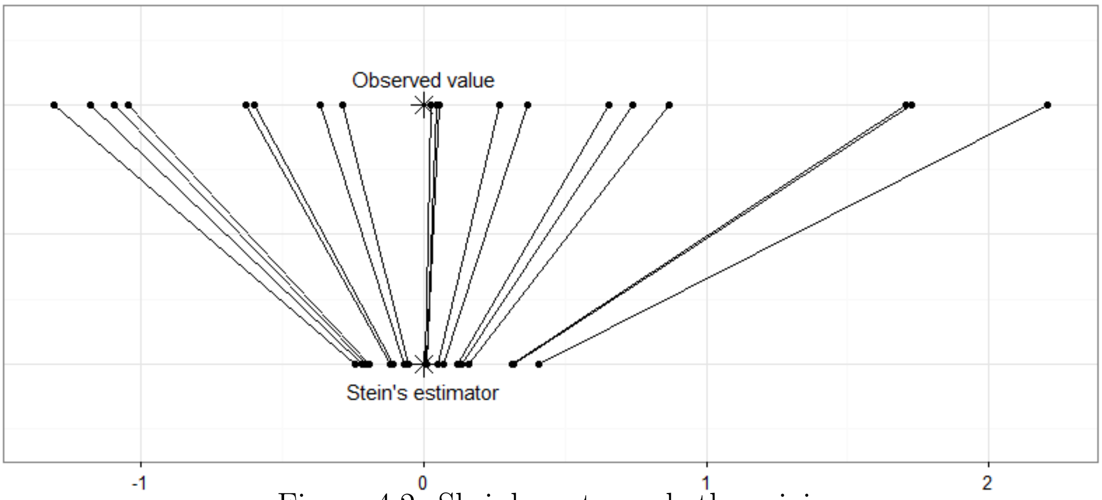

Figure 4.2: Shrinkage towards the origin

The result become more paradoxical when the mean does not coincide with the origin because, as the Gaussian law is distributed symmetrically, shrinking everything towards the origin feels unnatural, but the following graph shows that when the mean is $(1, ..., 1)$, the improvement is still substantial, even if less important.

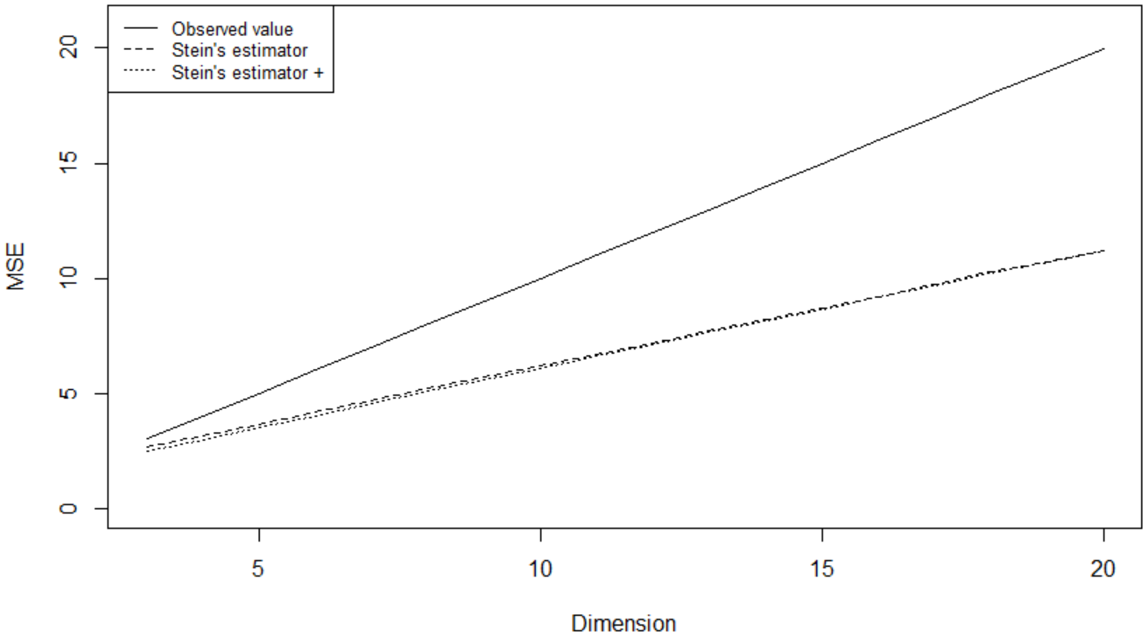

Figure 4.3: MSE when the dimension increases, with the mean equal to $(1, ..., 1)$

Doing the same as in Figure 4.2 in this case gives Figure 4.4, where we can see that almost all the values in the James-Stein estimator are below the actual mean.

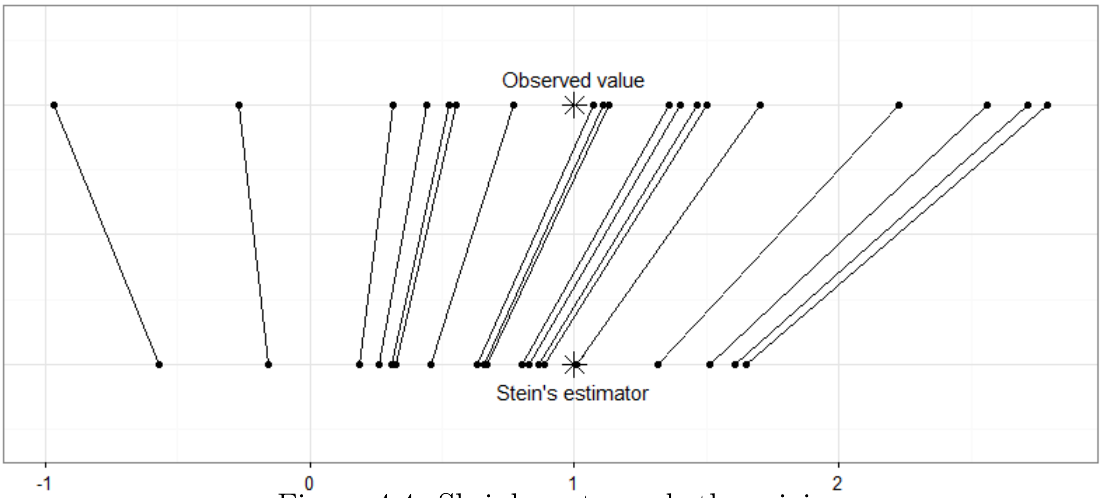

Figure 4.4: Shrinkage towards the origin

On Figure 4.3, we also see that the positive-part and normal estimator almost coincide because differences only occur when the norm of X is really small, which is less often the case here. The greater improvement when X is close to the origin can still be exploited, as it was evoked in Chapter 1, by "centering" the data. Typically, if you have a sample of n observations,$x_1, ..., x_n$, that you treat as a vector X, you could subtract the mean of the sample, $\bar{x}$, to each component, then taking the James-Stein estimator. Doing so, we lose one degree of freedom and the estimator becomes

$$\left(1 - \frac{p-3}{\|X-\nu\|^2}\right)(X-\nu)+\nu$$

where ν denotes the vector $(\bar{x}, ..., \bar{x})'$. Instead of shrinking towards the origin, we are now shrinking towards ν. This estimator will be called the *centered* James-Stein estimator. This will bring improvement when the mean is close to a multiple of $(1, ..., 1)'$, or to stay in the situation we just described, when all the observations in our sample have similar means. However, depending on the form of the mean, this can sometimes give worse results.

We will now fix the dimension to be equal to 20 to focus on the influence of the norm and the form of the mean vector.

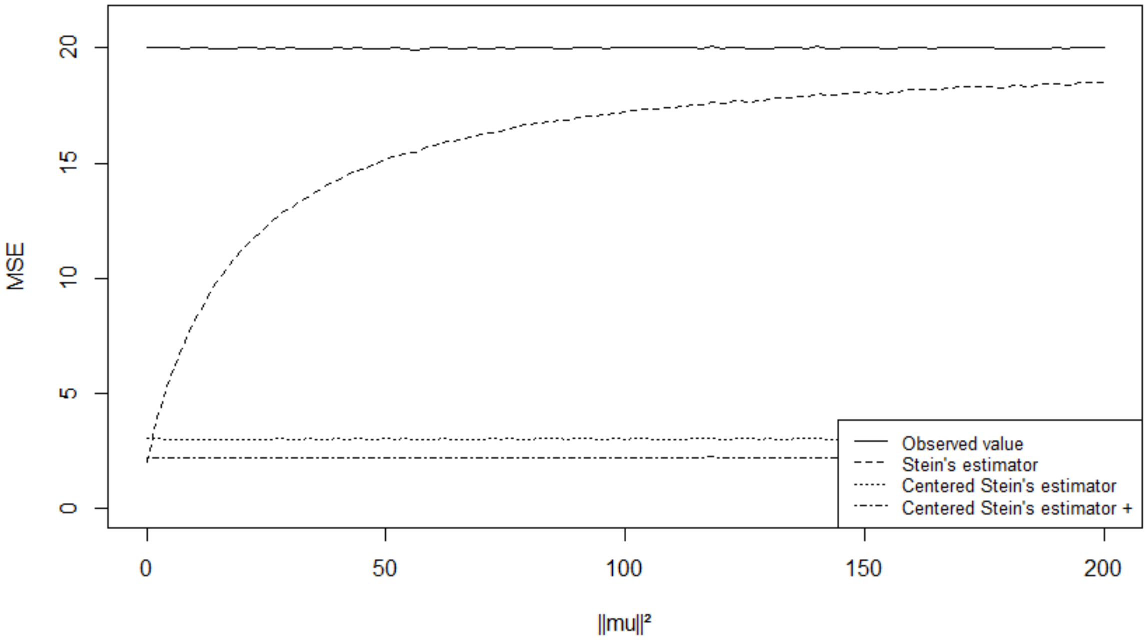

Figure 4.5: MSE when the norm of the mean increases, with the mean a multiple of $(1, ..., 1)$

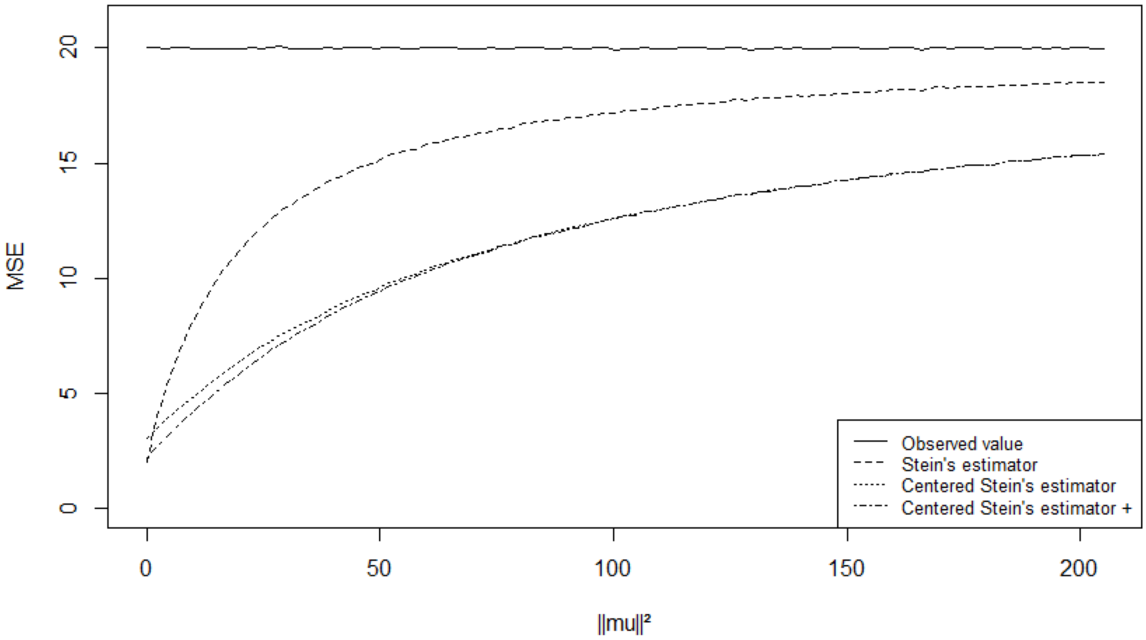

Figure 4.6: MSE when the norm of the mean increases, with the mean a multiple of $(1, 2, ..., 20)$

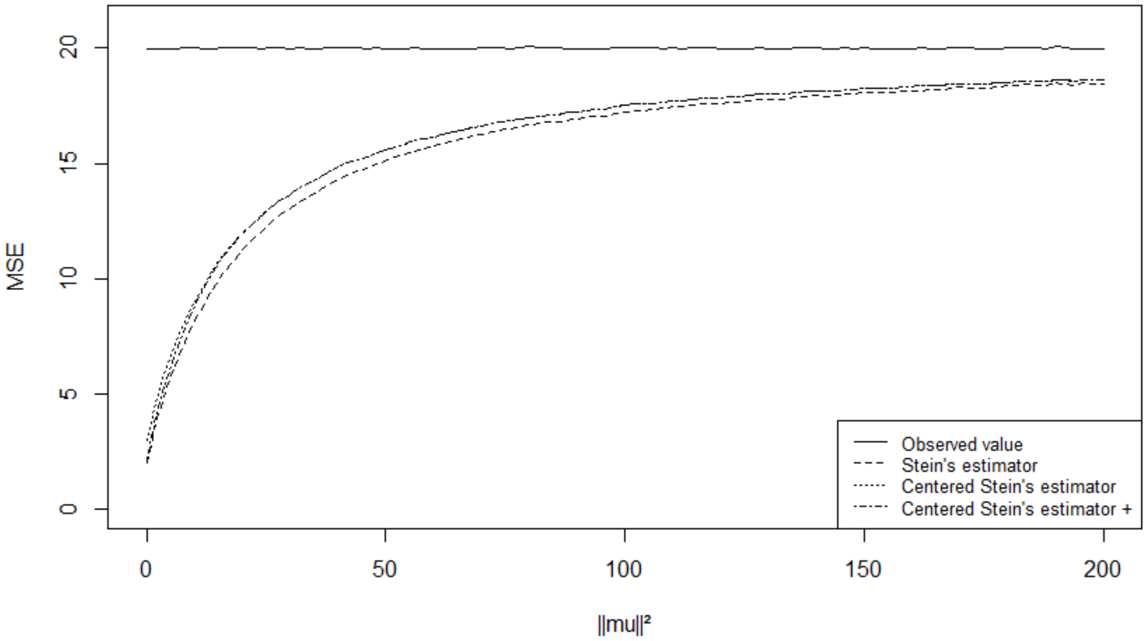

Figure 4.7: MSE when the norm of the mean increases, with the mean a multiple of $(-1, ..., -1, 1, ..., 1)$

In each graph, the MSE of the usual estimator gets closer and closer to 20, the MSE of X, as the norm of mean gets larger, but the evolution is the same no matter the type of mean. The form of the mean, only influences the centered estimator: when the mean is a multiple of $(1, ..., 1)'$, the centered version gives the same results as if the mean was 0, which was to be expected, and the positive-part version is even better ; when the mean is a multiple of $(1, 2, ..., 20)'$ the centered version still gives better results by a significant margin. The next figure shows that the centered estimator is not always a good solution to the mean being different from 0: the means is a multiple of $(0, ..., 0, 1)$ and, as the difference appears more clearly in this case, the dimension is lowered to only 5.

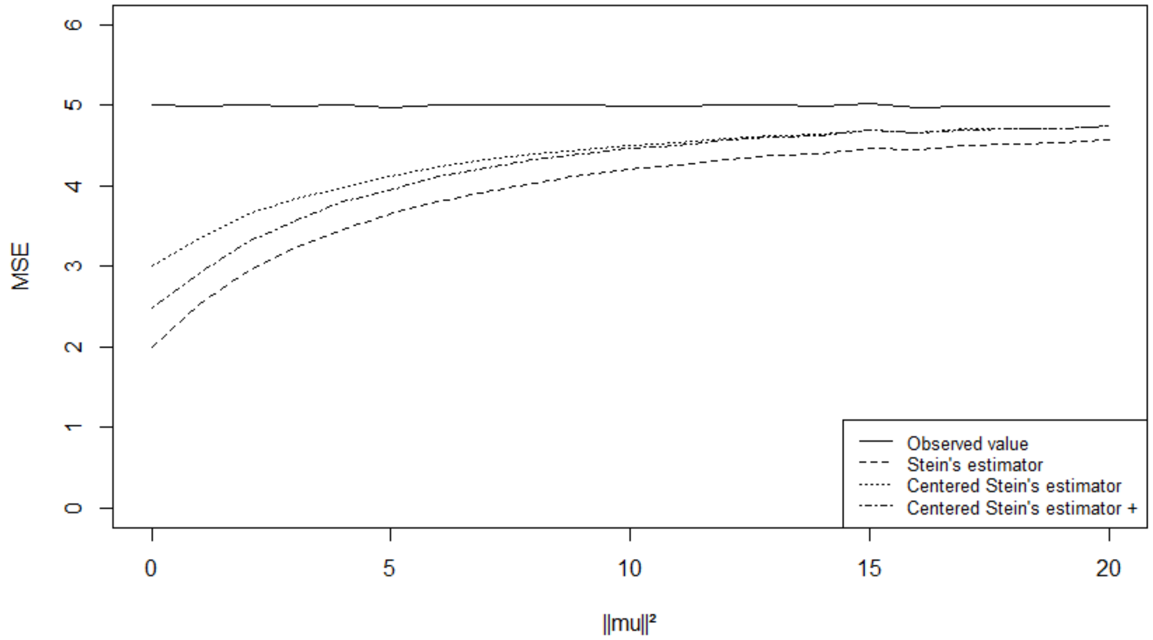

Figure 4.8: MSE when the norm of the mean increases, with the mean a multiple of $(0, ..., 0, 1)$

Let us now get to the case where the covariance matrix is an unknown multiple of the identity: $X \sim \mathcal{N}_p(\theta, \sigma^2 I)$. We will look at the estimators (1.5) and (1.6) from Chapter 1. A low value of $n = 10$ was chosen to show that choosing the denominator equal to $n+2$ was indeed better than n, but even here, there is only a small improvement. If n gets bigger, the difference becomes negligible. The dimension is still set to 20, the mean is equal to $(1, ..., 1)$ and it is σ^2 that varies. The s found in the estimators studied would usually be obtained through a sample, but for easier simulations, it was directly generated as a $\sigma^2\chi^2_n$ variable.

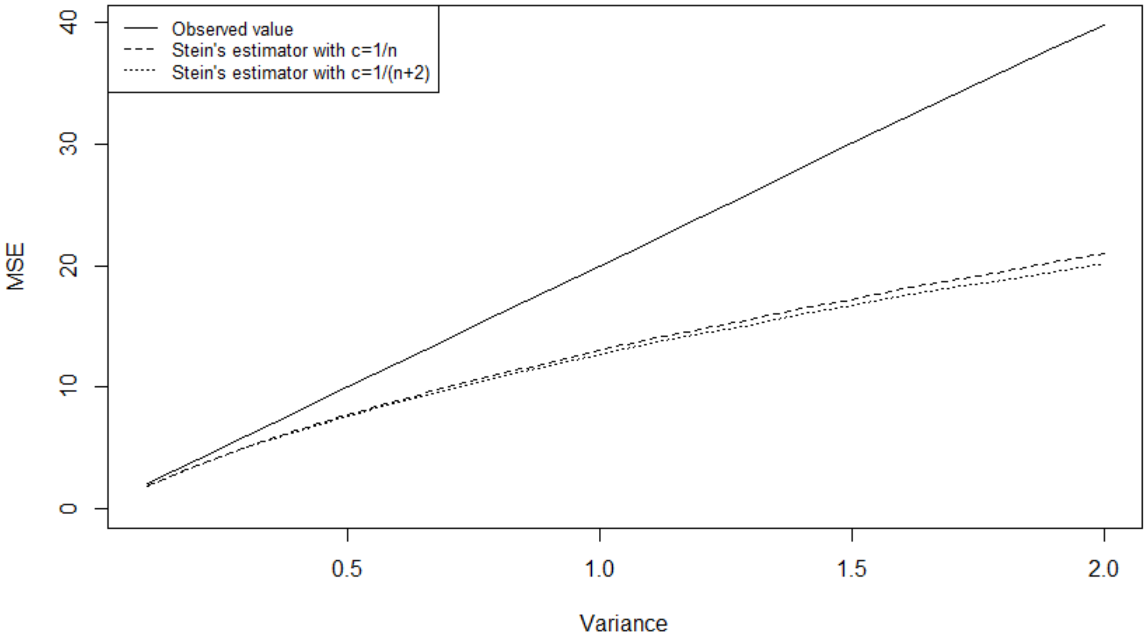

Figure 4.9: MSE when the variance increases

In Chapter 2, we saw that the James-Stein estimator was also usable for the multivariate Student distribution. The next two figures (4.10 and 4.11) look at the differences compared to the normal case. The mean was set to 0 for Figure 4.10 and to a multiple of $(1, ..., 1)'$ for Figure 4.11. The covariance is the identity matrix in both cases. The results are very similar. The MSE of the James-Stein estimator follows the same evolution as the MSE of the true value when the degrees of freedom increases. The degree of improvement is also about the same as in the normal case.

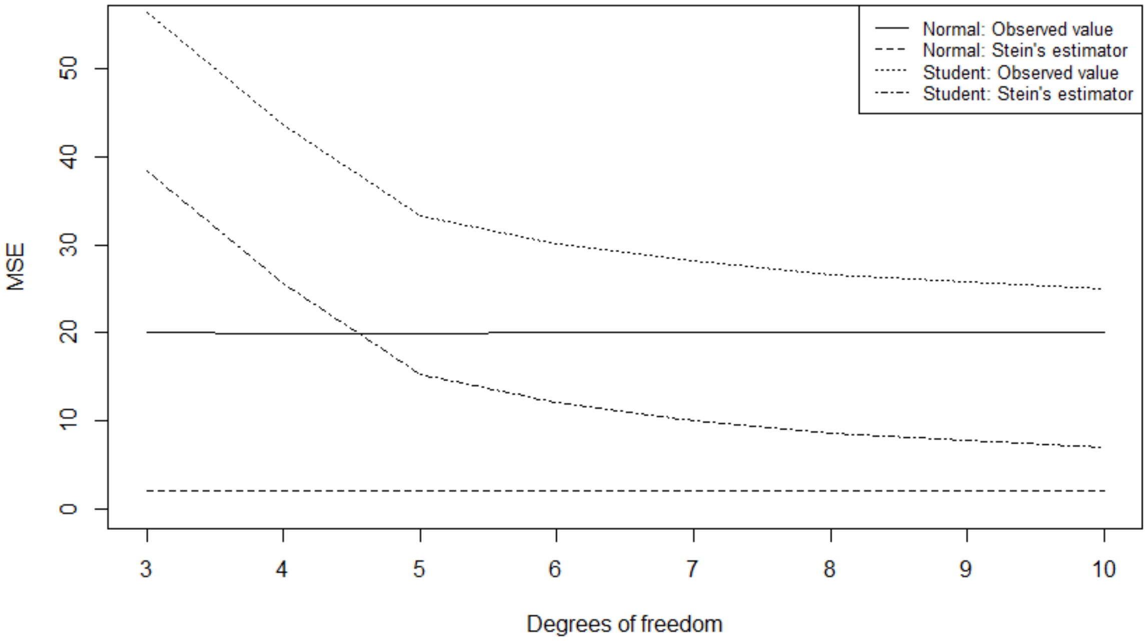

Figure 4.10: MSE in the Student and normal cases as the degrees of freedom of the Student increase

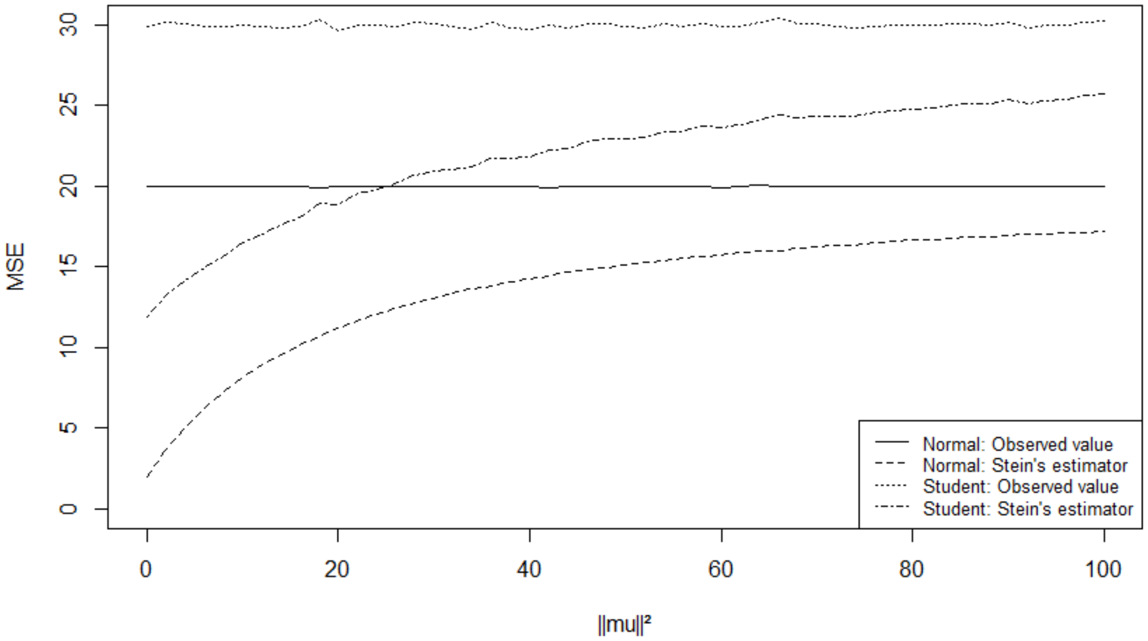

Figure 4.11: MSE in the Student and normal cases as the norm of the mean increases

Finally, we go back to the normal case to look at the case of unknown covariance and compare when $p > n$ and $p \leq n$. The studied estimators are

$$\left(1 - \frac{p-2}{(n-p+3)X'S^{-1}X}\right)X$$

when $p \leq n$, and

$$\left(I - \frac{(n-2)SS^{+}}{(p-n+3)X'S^{+}X}\right)X$$

when $p > n$. The MSE is estimated on a sample of size 5000 and for each observation in this sample, a sample of size $n+1$ is generated to compute S, so that this one follows a Wishart with n degrees of freedom. At first, the same matrix S was used for each of the 5000 observations, but the estimator was found to be quite unstable, which was fixed by generating a new one each time, the downside being a huge time loss on the simulations due to the drastic increase in computations required.

We begin by looking at the effect of the form of the covariance matrix. The dimension is fixed to 50 and the mean is $(1, ..., 1)'$. three types of covariance matrix are studied:

- Spiked: a diagonal matrix with 1 in the first half and 5 in the second.
- Block: a block diagonal matrix formed of 2 × 2 blocks

$$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

- Regressive: with $\rho = 0.5$, a matrix of the form

$$\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots \\ \rho & 1 & \rho & \rho^2 & \\ \rho^2 & \rho & 1 & \rho & \\ \rho^3 & \rho^2 & \rho & 1 & \\ \vdots & & & & \ddots \end{pmatrix}$$

As a reminder, the loss used here is $\mathbb{E}\left[(\hat{\theta}-\theta)'\Sigma^{-1}(\hat{\theta}-\theta)\right]$ where $\hat{\theta}$ is the estimator, so the MSE of the trivial estimator X will be equal to the dimension, no matter the form of the covariance matrix.

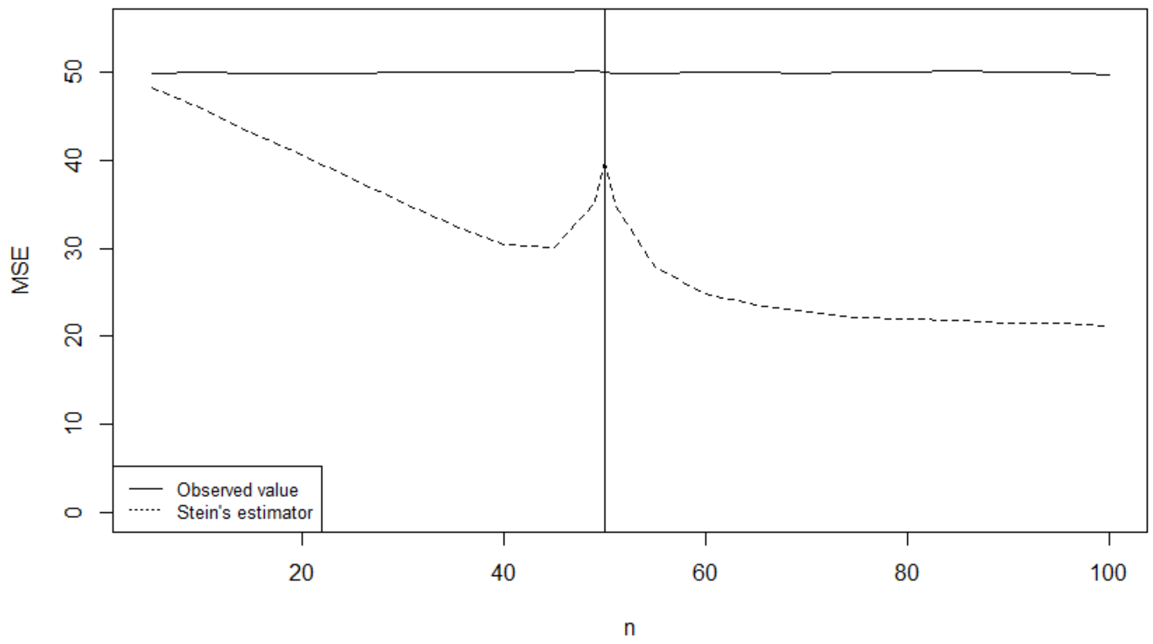

Figure 4.12: MSE when n increases for spiked covariance

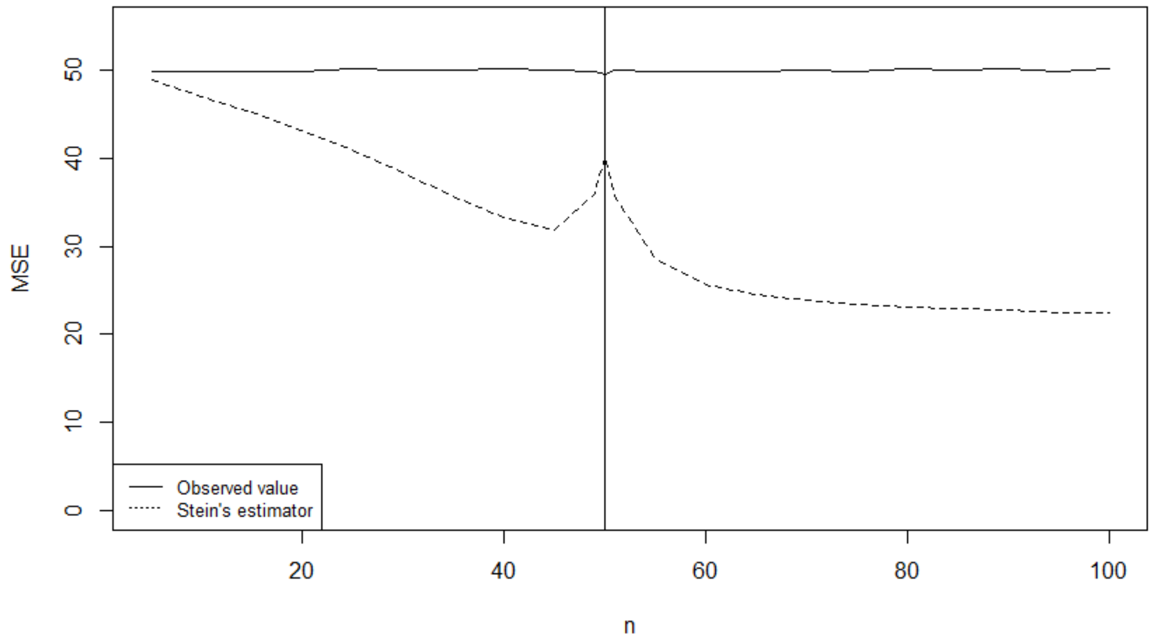

Figure 4.13: MSE when n increases for block covariance

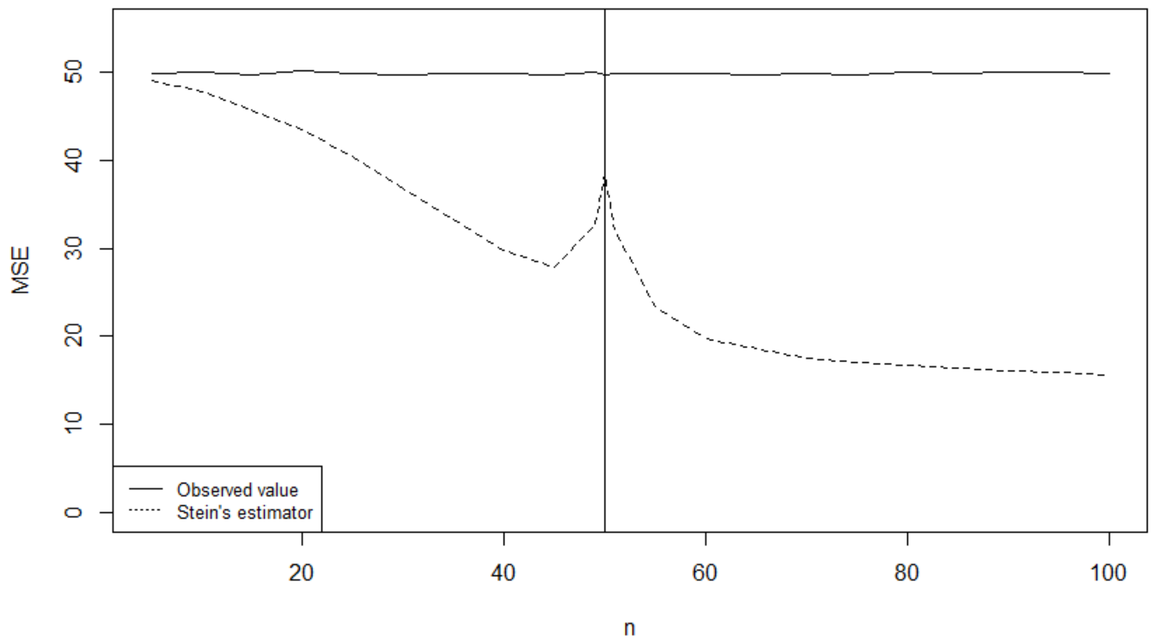

Figure 4.14: MSE when n increases for regressive covariance

There are no significant difference depending on the matrix, apart from slightly better results in the regressive covariance case. What is interesting is the peak when n gets close to p. This is not too surprising looking at the expression for the risk when $p \leq n$ (1.9), where, when $n-p$ is really small, the term $\frac{n-p+1}{n-p+3}$ gets small and the improvement is worse. If we ignore this peak, the improvement gets better and better as the size of the sample used to estimate S increases, which does not come as a surprise. The last two graphs will look at the effect of the dimension and of the norm of the mean, when $n = p-1$ and $n = \frac{p}{2}$, with block covariance.

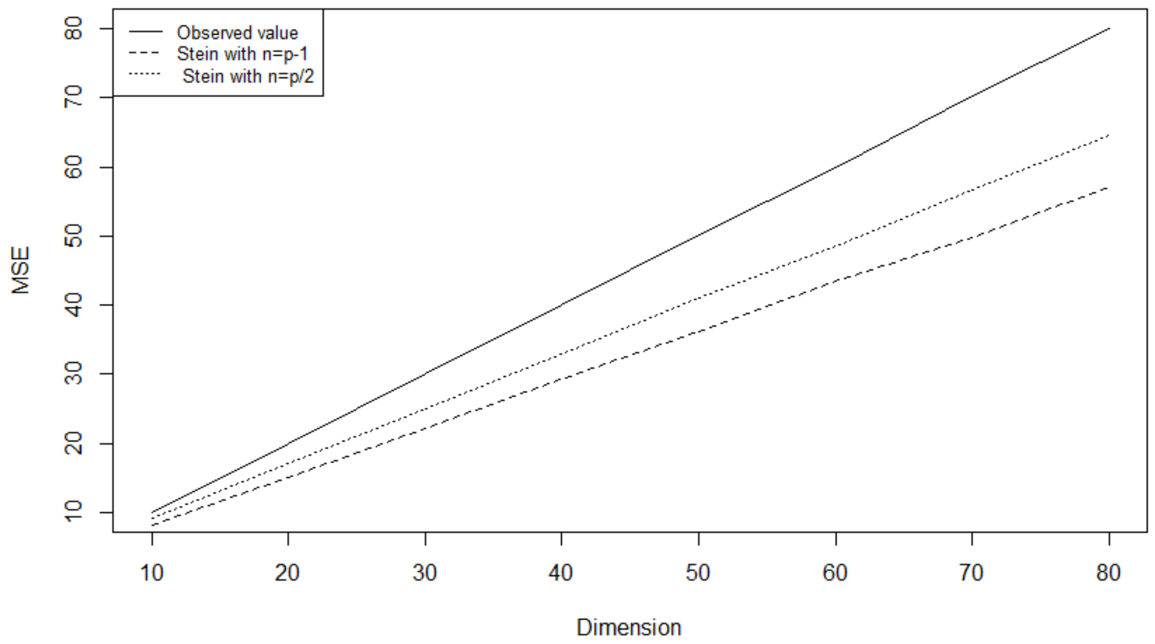

Figure 4.15: MSE when the dimension p increases, with the mean equal to $(1, ..., 1)'$

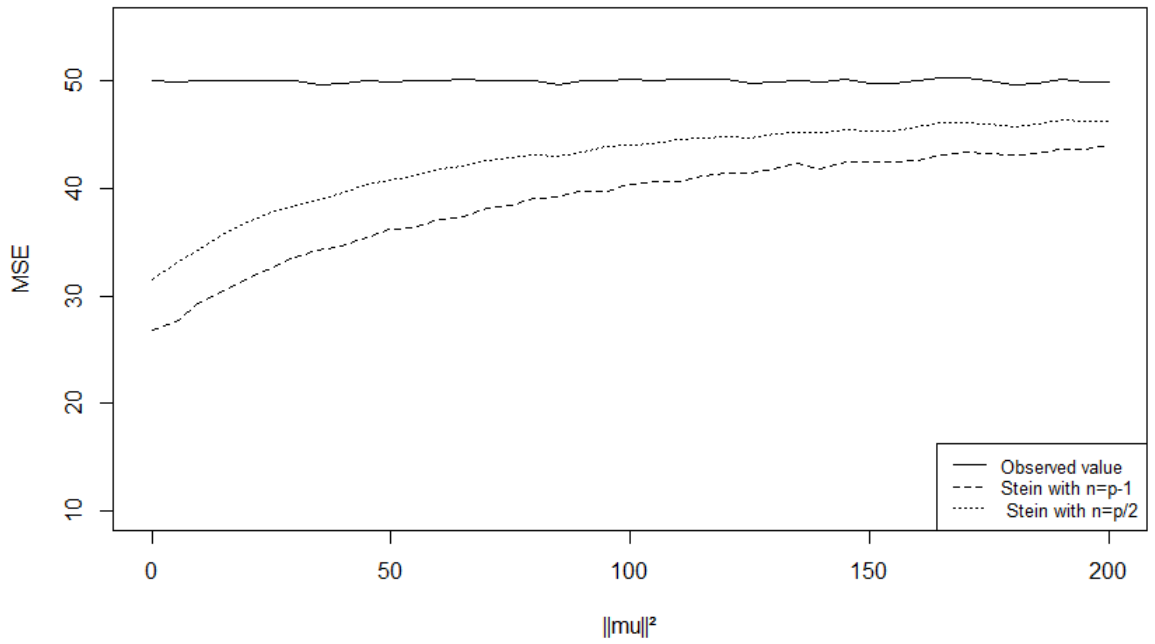

Figure 4.16: MSE when the norm of the mean (a multiple of $(1, ..., 1)'$) increases, with $p = 50$

For both, the evolution is similar to what was observed in Figures 4.3 and 4.5, in the case of identity covariance matrix.

Appendix A

Appendix: Technical results

A.1 Results on Wishart matrices

The two results on Wishart matrices are easily proved using the Bartlett decomposition. We therefore begin by reminding it. Various proofs of this exist: we look at one from [8].

Proposition A.1. *Let S be a Wishart matrix $\mathcal{W}_p(n, I)$. Then S can be written as BB', with*

$$B = \begin{pmatrix} c_1 & 0 & 0 & \dots & 0 \\ n_{21} & c_2 & 0 & \dots & 0 \\ n_{31} & n_{32} & c_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_{p1} & n_{p2} & n_{p3} & \dots & c_p \end{pmatrix}$$

where $c_i^2 \sim \chi^2_{n-i+1}$ and $n_{ij} \sim \mathcal{N}(0,1)$, all independent.

Proof. We know S can be written as $X'X$, where X is a $n \times p$ matrix whose elements are all independent $\mathcal{N}(0,1)$ variables. The columns of X are denoted X_i $(i \in \{1, ..., p\})$. We then build iteratively $Y_1 = X_1$ and

$$Y_i = X_i - b^*_{i1} Y_1 - -b^*_{i2} Y_2 - ... - b^*_{i,i-1} Y_{i-1} \quad (i \in \{2, ..., p\}),$$

where the b^*_{ir} are chosen so that $Y_i' Y_j = 0$ for all $i \neq j$. This means that $b^*_{ir} = \dfrac{Y_r' X_i}{Y_r' Y_r}$. Now defining

$$b_{ir} = (Y_r' Y_r)^{\frac{1}{2}} b^*_{ir} = \frac{Y_r' X_i}{(Y_r' Y_r)^{\frac{1}{2}}}$$

for $i \in \{2, ..., p\}$ and $r \in \{1, ..., i-1\}$, we can verify that

$$X_i' X_i = Y_i' Y_i + \sum_{r=1}^{i-1} b_{ir}^2$$

and

$$X_i'X_j = \sum_{r=1}^{i-1} b_{ir}b_{jr} + b_{ji}(Y_i'Y_i)^{\frac{1}{2}} \text{ for } j > i.$$

As $S_{ij} = X_i'X_j$, this gives us the matrix announced

$$B = \begin{pmatrix} (Y_1'Y_1)^{\frac{1}{2}} & 0 & 0 & \dots & 0 \\ b_{21} & (Y_2'Y_2)^{\frac{1}{2}} & 0 & \dots & 0 \\ b_{31} & b_{32} & (Y_3'Y_3)^{\frac{1}{2}} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{p1} & b_{p2} & b_{p3} & \dots & (Y_p'Y_p)^{\frac{1}{2}} \end{pmatrix}.$$

For all $i \in \{2, ..., p\}$, fix $X_1, ..., X_{i-1}$. Then the b_{ir} ($r \in \{1, ..., i-1\}$) are orthogonal (because $Y_i'Y_j = 0$) combinations of the independent variables $x_{i1}, ..., x_{in}$. We have

$$\begin{aligned} \mathbb{E}\left[b_{ir}\right] &= 0 \\ \text{and} \quad V(b_{ir}) &= \frac{Y_r'Y_r}{Y_r'Y_r} = 1. \end{aligned}$$

The b_{ir} are therefore also independent $\mathcal{N}(0,1)$ variables. From this and Fisher's Lemma, we have that

$$Y_i'Y_i = X_i'X_i - \sum_{r=1}^{i-1} b_{ir}^2$$

follows a $\chi^2_{n-(i-1)}$ law, independent of the b_{ir}. As this is true for all i, this concludes the proof. □

We can now look at the distribution of $Y'S^{-1}Y$ given Y a p-vector, where S follows a Wishart $\mathcal{W}_p(n, I)$ ($n \geq p$). There exists an orthogonal matrix A such that $Y^* = AY$ has its $p-1$ first components equal to 0. Then $Y'S^{-1}Y = (Y^*)'S^{*-1}Y^*$, where $S^* = ASA'$ still follows a Wishart $\mathcal{W}_p(n, I)$. This means that $Y'S^{-1}Y = (Y'Y)(S^{-1})_{pp}$. Denoting by $S_{(p-1)}$ the matrix S deprived of its last row and column, we have that $(S^{-1})_{pp} = \dfrac{\det(S_{(p-1)})}{\det(S)} = Y_p'Y_p$, where Y_p is the one from the proof of the proposition. This means that the distribution of $Y'S^{-1}Y$ given Y is that of $\dfrac{Y'Y}{c_p}$ where c_p is a χ^2_{n-p+1} variable.

From Proposition A.1, we can also get the distribution of the trace of a Wishart matrix. If $S \sim \mathcal{W}_p(n, I)$, then, using the notations of the proposition,

$$\operatorname{tr}(S) = \operatorname{tr}(BB') = \sum_{i=1}^{p} c_p^2 + \sum_{i>j} n_{ij}^2$$

and this follows a χ^2 law with $\sum_{i=1}^{p}(n-i+1) + \sum_{i>j} 1 = np$ degrees of freedom.

A.2 Results on the Moore-Penrose Inverse

We only consider real matrices.

Proposition A.2. *The Moore-Penrose inverse of a matrix A is unique.*

Proof. We suppose A has two Moore-Penrose inverse, B and C. Then, using the definition, we write

$$AB = ACAB = (AC)'(AB)' = C'(ABA)' = C'A' = (AC)' = AC.$$

Similarly, we get that $BA = CA$. Therefore

$$B = BAB = BAC = CAC = C.$$

□

Proposition A.3. *For any matrix A, $A^+ = (A'A)^+A'$.*

Proof. We need two small intermediate results. First, note that if $A'A = 0$, then $A = 0$. This follows easily from

$$0 = \mathrm{tr}(A'A) = \sum_i (A'A)_{ii} = \sum_{i,j} (A')_{ji} A_{ij} = \sum_{i,j} A_{ij}^2$$

which mean that $A_{ij} = 0 \forall i, j$. From this, we get that if $A'AB = 0$, then $AB = 0$, as

$$\begin{aligned} A'AB &= 0 \\ \Rightarrow B'A'AB &= 0 \\ \Rightarrow (AB)'(AB) &= 0 \\ \Rightarrow AB &= 0. \end{aligned}$$

Now, for the actual proof of the proposition, we just have to check that $D = (A'A)^+A'$ verifies the definition of the Moore-Penrose inverse. Start by looking at

$$\begin{aligned} A'A &= A'A(A'A)^+A'A \\ \Leftrightarrow A'A &= A'ADA \\ \Leftrightarrow 0 &= A'A(DA - I) \\ \Leftrightarrow 0 &= A(DA - I) \\ \Leftrightarrow A &= ADA. \end{aligned}$$

To show that $DAD = D$, just look at $(A'A)^+ = (A'A)^+A'A(A'A)^+$ and multiply both side by A' on the right.

Finally we have $(AD)' = (A(A'A)^+A')' = A((A'A)^+)'A' = A(A'A)^+A' = AD$, because the pseudoinverse of a symmetric matrix is also symmetric, and $DA = (A'A)^+A'A$, which is symmetric by definition of the Moore-Penrose pseudoinverse .

Thus we have that $D = A^+$ □

Proposition A.4. *We have, for A invertible and T symmetric,*

$$(ATA)^+ = (T^+TA)^+T^+(AT^+T)^+ \tag{A.1}$$

Proof. These are proved using equations (1.2) and (1.4) in [16], reminded here:

$$\begin{aligned}(AB)^+ &= (A^+AB)^+(ABB^+),\\ (ABC)^+ &= (A^+ABC)^+B(ABCC^+)^+.\end{aligned}$$

Then,

$$\begin{aligned}(ATA)^+ &= (A^+ATA)^+T(ATAA^+)^+\\ &= (TA)^+T(AT)^+\\ &= (T^+TA)^+(TAA^+)^+T(A^+AT)^+(ATT^+)^+\\ &= (T^+TA)^+T^+TT^+(ATT^+)^+\\ &= (T^+TA)^+T^+(ATT^+)^+.\end{aligned}$$

This gives (A.1). □

Appendix B

Appendix R code

The code for the example on singular matrices and Moore-Penrose in Chapter 3.

```
############################################
## Example for the Moore-Penrose inverse ##
############################################

set.seed(12)

# 3 normal vectors of size 4
X <- mvrnorm(n=3,mu = c(0,0,0,0),Sigma = diag(1:4))
round(X,digits = 2)

#
S <- 2*cov(X)

round(S,digits = 2)

det(S)

# singular value decomposition
SVD_S <- svd(S)
D <- SVD_S$d
U <- SVD_S$u
V <-SVD_S$v

round(SVD_S$d,digits = 2)
round(U,digits = 2)
round(V,digits = 2)

round(U%*%diag(D)%*%t(V),digits = 2) #We get S back

round(U%*%t(U),digits = 2) #UU' = I
round(V%*%t(V),digits = 2) #VV' = I

# Moore-Penrose inverse
SP <- ginv(S)
round(SP,digits = 2)

round(SP%*%S,digits = 2) #not the identity, but...
round(S%*%SP%*%S,digits = 2) #We get S back
round(SP%*%S%*%SP,digits = 2) #We get SP back

# SVD of the Moore-Penrose inverse
Dinv <- c(1/D[1],1/D[2],0,0) #We take the inverse of the non-zero values in the SVD of S

round(V%*%diag(Dinv)%*%t(U),digits = 2) #We get SP, using the same U and V as for S
```

The code for the graphs in Chapter 4. It is advised to run each part of the code one by one and to maybe adjust the sample sizes if reasonable computation times are desired.

```
### Required libraries ###
library(MASS)

###  Sample size  ###
N=50000

### Useful functions ###
max0 <- function(x){return(max(0,x))} #returns the max between 0 and x
normsq <- function(x){return(sum(x^2)) } #returns the square norm of x

#returns the square error of Y dependind on the mean m and he covariance A
err_sigma <- function(Y,m,A){
  return(as.numeric((t(Y-m)%*%A)%*%(Y-m)))
}

#creates a block matrix of size p (p must be even)
sigma_block <- function(p){
  SB <- diag(p)

  for(i in 2*(1:(p/2))){
    SB[i,i-1] <- 0.5
    SB[i-1,i] <- 0.5
  }
  return(SB)
}
# returns a "regressive" type matrix of size p
regressive_sigma <- function(p){
  SB <- diag(p)
  rho <- 0.5
  for(i in 1:(p-1)){
    for(j in (i+1):p){
      SB[j,j-i] <- rho^i
      SB[j-i,j] <- rho^i
    }
  }
  return(SB)
}

#######################################
### When Covariance is the identity ###
#######################################

# A function that will compute and plot the MSE for different types of James-Stein
    estimator
# with the mean equal to the parameter "mu" multiplied by the coefficients in "range"
Stein_estimator_Id <- function(mu,range){

  #Dimension, based on the length of the mean
  p <- length(mu)

  #Vectors that will contain the MSE
  E_X_err <- NULL
  E_XS_err <- NULL
  E_XSC_err <- NULL
  E_XSC2_err <- NULL
  #Vector that will contain the norm of the mean, for use in the plot
  Norm_mu <- NULL
  for(i in range){

    X <- mvrnorm(n = N, i*mu ,diag(p)) #N samples of dim d
    XS <- X-(p-2)*(X/rowSums(X^2)) #Stein's estimator

    X_temp <- X-rowMeans(X) #Centered
    XS_temp <- X_temp-(p-3)*(X_temp/rowSums(X_temp^2)) #Centered Stein's estimator
    XS_temp2 <- sapply(1-(p-3)/rowSums(X_temp^2),max0)*X_temp #Centered Stein's estimator
        +

    XSC <- XS_temp + rowMeans(X) # "Un-centering"
    XSC2 <- XS_temp2 + rowMeans(X) # "Un-centering"

    #Square error
    X_err <- colSums((t(X)-i*mu)^2)
    XS_err <- colSums((t(XS)-i*mu)^2)
    XSC_err <- colSums((t(XSC)-i*mu)^2)
    XSC2_err <- colSums((t(XSC2)-i*mu)^2)

    #Mean square error
    E_X_err <- c(E_X_err,mean(X_err))
    E_XS_err <- c(E_XS_err,mean(XS_err))
    E_XSC_err <- c(E_XSC_err,mean(XSC_err))
    E_XSC2_err <- c(E_XSC2_err,mean(XSC2_err))

    #Squared norm of mu (for the x-coordinates in the plot)
    Norm_mu <- c(Norm_mu,normsq(i*mu))
  }

  #Plot the MSE as a function of the norm of the mean
  plot(Norm_mu,E_X_err, lty=1, type = "l", ylim = c(0,p+1),
       main = NULL ,ylab = "MSE",xlab = "||mu||^2")
  lines(Norm_mu,E_XS_err, lty=2)
  lines(Norm_mu,E_XSC_err, lty=3)
  lines(Norm_mu,E_XSC2_err, lty=4)
  legend("bottomright", legend=c("Observed value", "Stein's estimator","Centered Stein's
      estimator","Centered Stein's estimator +"),
         lty = 1:4,cex = 0.8)


}


##################################################
##  Mean = 0: error depending on the dimension  ##
##################################################

E_X_err <- NULL
E_XS_err <- NULL
E_XSP_err <- NULL

dimensions <- 3:20

for(i in dimensions){

  mu <- rep(1,i)

  X <- mvrnorm(n = N, mu ,diag(i)) # N samples of dim d
  XS <- X-(i-2)*(X/rowSums(X^2)) #Stein's estimator
  XSP <- sapply(1-(i-2)/rowSums(X^2),max0)*X

  #Square error
  X_err <- colSums((t(X)-mu)^2)
  XS_err <- colSums((t(XS)-mu)^2)
  XSP_err <- colSums((t(XSP)-mu)^2)


  #Mean square error

  E_X_err <- c(E_X_err,mean(X_err))
  E_XS_err <- c(E_XS_err,mean(XS_err))
  E_XSP_err <- c(E_XSP_err,mean(XSP_err))

}

plot(dimensions,E_X_err, lty=1, type = "l",ylim=c(0,21),
     main = NULL,ylab = "MSE",xlab = "Dimension")
lines(dimensions,E_XS_err, lty=2)
lines(dimensions,E_XSP_err, lty=3)
legend("topleft", legend=c("Observed value", "Stein's estimator","Stein's estimator +"),
       lty = 1:3,cex = 0.8)

###########################################
##  When the norm of the mean increases  ##
###########################################

p=20 #dimension

##################
##  Mean = m*1  ##
##################

mu <- rep(1,p) #mean vector

#while(normsq(m_max*mu)<100){m_max = m_max+0.1}

range <- sqrt(0:100/10) #multiplicator of the mean vector

Stein_estimator_Id(mu,range)

######################
##  Mean = m*(1:p)  ##
######################

mu <- 1:p
range <- sqrt(((0:100)/1400))

Stein_estimator_Id(mu,range)

#############################
##  Mean = m*(0,...,0,1)   ##
#############################

p=5

mu <- rep(0,p-1)
mu <- c(mu,1)
range <- sqrt(0:20)

Stein_estimator_Id(mu,range)

####################################
##  Mean = m*(-1,...,-1,1,...,1)  ##
####################################

p=20

mu <- rep(-1,p/2)
mu <- c(mu,rep(1,p/2))
```

```r
range <- sqrt(0:100/10)

Stein_estimator_Id(mu,range)

###################################
##  Mean = m*(0,...,0,1,...,1)  ##
###################################

mu <- rep(0,p/2)
mu <- c(mu,rep(1,p/2))
range <- sqrt(0:200/10)

Stein_estimator_Id(mu,range)

#######################################
##  When Cov = k*Id with k unknown  ##
#######################################

E_X_err <- NULL
E_XS_err <- NULL
E_XS2_err <- NULL

p = 20
n = 10
N = 50000
mu <- rep(1,p)

variance <- (1:20)/10

for(v in variance){

  #The estimations for k, following a sigma^2 chi-squared
  s <- v*rchisq(n = N, df = n)

  X <- mvrnorm(n = N, mu ,v*diag(p)) #N samples of dim p

  XS <- X-(p-2)*(s/n)*(X/rowSums(X^2)) #Stein's estimator
  XS2 <- X-(p-2)*(s/(n+2))*(X/rowSums(X^2))

  #Square error
  X_err <- colSums((t(X)-mu)^2)
  XS_err <- colSums((t(XS)-mu)^2)
  XS2_err <- colSums((t(XS2)-mu)^2)

  #Mean square error

  E_X_err <- c(E_X_err,mean(X_err))
  E_XS_err <- c(E_XS_err,mean(XS_err))
  E_XS2_err <- c(E_XS2_err,mean(XS2_err))

}

plot(variance,E_X_err, lty=1, type = "l",ylim=c(0,max(E_X_err)),
     main = NULL,ylab = "MSE",xlab = "Variance")
lines(variance,E_XS_err, lty=2)
lines(variance,E_XS2_err, lty=3)
legend("topleft", legend=c("Observed value", "Stein's estimator with c=1/n", "Stein's
    estimator with c=1/(n+2)"),
       lty = 1:3,cex = 0.8)


#######################################
##  Example to show the "shrinkage"  ##
#######################################
```

```
library(ggplot2)

p=20

### Mean = 0 ###

set.seed(6) #example 1: the coefficient is positive

X <- rnorm(p,0,1) #one normal vector of size 20
XS <- X-(p-2)*(X/normsq(X))

height <- c(rep(1,p),rep(0,p)) #this is to put the true obs on top and the stein
    estimation below
pairing <- c(1:p,1:p) #this allows to draw a line between corresponding observations

data <- data.frame( x= c(X,XS), height,pairing) #data frame as it is required by ggplot

ggplot(data, aes(x=x,y=height, group = pairing))+
  ylim(-0.3,1.3)+
  geom_point()+
  geom_line()+ #add the lines between the obs
  annotate("text",x=0,y=c(-0.1,1.1),label = c("Stein's estimator", "Observed value"))+
  theme_bw()+ #set the background white
  theme(axis.title.y=element_blank(),
        axis.title.x=element_blank(), #removes unecessary axis annotations
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank())+
  geom_point(aes(x= 0, y = 1),pch = 8,size = 4)+ #add two stars to locate the mean
  geom_point(aes(x= 0, y = 0),pch = 8,size = 4)

#example 2: the coefficient is negative and the result is bad
set.seed(100)

X <- rnorm(p,0,1) #one normal vector of size 20
XS <- X-(p-2)*(X/normsq(X))

height <- c(rep(1,p),rep(0,p)) #this is to put the true obs on top and the stein
    estimation below
pairing <- c(1:p,1:p) #this allows to draw a line between corresponding observations

data <- data.frame( x= c(X,XS), height,pairing) #data frame as it is required by ggplot

ggplot(data, aes(x=x,y=height, group = pairing))+
  ylim(-0.3,1.3)+
  geom_point()+
  geom_line()+ #add the lines between the obs
  annotate("text",x=0,y=c(-0.1,1.1),label = c("Stein's estimator", "Observed value"))+
  theme_bw()+ #set the background white
  theme(axis.title.y=element_blank(),
        axis.title.x=element_blank(), #removes unecessary axis annotations
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank())+
  geom_point(aes(x= 0, y = 1),pch = 8,size = 4)+ #add two stars to locate the mean
  geom_point(aes(x= 0, y = 0),pch = 8,size = 4)


### Mean = 1 ###
set.seed(123)
X <- rnorm(p,1,1)
XS <- X-(p-2)*(X/normsq(X))

height <- c(rep(1,p),rep(0,p)) #this is to put the true obs on top and the stein
    estimation below
pairing <- c(1:p,1:p) #this allows to draw a line between corresponding observatyion
```

```
data <- data.frame( x= c(X,XS), height,pairing) #data frame as it is required by ggplot

ggplot(data, aes(x=x,y=height, group = pairing))+
  ylim(-0.3,1.3)+
  geom_point()+
  geom_line()+ #add the lines between the obs
  annotate("text",x=1,y=c(-0.1,1.1),label = c("Stein’s estimator", "Observed value"))+
  theme_bw()+ #set the background white
  theme(axis.title.y=element_blank(),
        axis.title.x=element_blank(), #removes unecessary axis annotations
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank())+
  geom_point(aes(x= 1, y = 1),pch = 8,size = 4)+ #add two stars to locate the mean
  geom_point(aes(x= 1, y = 0),pch = 8,size = 4)


################################################
####  Comparison between normal and student  ####
################################################

library(mvtnorm)

p=20 #dimension
Sigma <- diag(p)

###################################
##  Degrees of freedom increase  ##
###################################

mu <- rep(0,p)

#degrees of freedom of the student
df_range <- 3:10

#Vectors that will contain the MSE
E_Xn_err <- NULL
E_XnS_err <- NULL
E_Xt_err <- NULL
E_XtS_err <- NULL

for(df in df_range){

  Xn <- mvrnorm(n=N,mu = mu,Sigma = Sigma) #Multinormal sample of size N to estimate the MSE
  XnS <- Xn-(p-2)*(Xn/rowSums(Xn^2))

  Xt <- rmvt(n=N,sigma = Sigma,df = df) #Multivariate t sample of size N
  XtS <- Xt-(p-2)*(Xt/rowSums(Xt^2))

  Xn_err <- colSums(t(Xn)^2) #Vector of the errors of the observed value for the normal case
  XnS_err <- colSums(t(XnS)^2) #Vector of the errors of Stein’s estimator for the normal case
  Xt_err <- colSums(t(Xt)^2) #Vector of the errors of the observed value for the student case
  XtS_err <- colSums(t(XtS)^2) #Vector of the errors of Stein’s estimator for the student case

  #MSE
  E_Xn_err <- c(E_Xn_err,mean(Xn_err))
  E_XnS_err <- c(E_XnS_err,mean(XnS_err))
  E_Xt_err <- c(E_Xt_err,mean(Xt_err))
  E_XtS_err <- c(E_XtS_err,mean(XtS_err))
}

plot(df_range,E_Xn_err, lty=1, type = "l",ylim = c(0,55),
     main = NULL,ylab = "MSE",xlab = "Degrees of freedom")
lines(df_range,E_XnS_err, lty=2)
lines(df_range,E_Xt_err, lty=3)
lines(df_range,E_XtS_err, lty=4)
legend("topright", legend=c("Normal: Observed value" ,"Normal: Stein's estimator","Student
    : Observed value" ,"Student: Stein's estimator"),
       lty = 1:4,cex = 0.8)

#######################
#### Mean increases ###
#######################

#covariance matrix
Sigma <- diag(p)

#multiplicator of the mean
range <- sqrt(0:50/10)

#Vectors that will contain the MSE
E_Xn_err <- NULL
E_XnS_err <- NULL
E_Xt_err <- NULL
E_XtS_err <- NULL

#Vector that will contain the norm of the mean, for use in the plot
Norm_mu <- NULL

for(m in range){

  mu <- m*rep(1,p) #Mean vector

  Xn <- mvrnorm(n=N,mu = mu,Sigma = Sigma) #Multinormal sample of size N to estimate the
      MSE
  XnS <- Xn-(p-2)*(Xn/rowSums(Xn^2))

  Xt <- rmvt(n=N,delta=mu,sigma = Sigma,df = 6) #Multivariate t sample of size N
  XtS <- Xt-(p-2)*(Xt/rowSums(Xt^2))

  Xn_err <- colSums((t(Xn)-mu)^2) #Vector of the errors of the observed value for the
      normal case
  XnS_err <- colSums((t(XnS)-mu)^2) #Vector of the errors of Stein's estimator for the
      normal case
  Xt_err <- colSums((t(Xt)-mu)^2) #Vector of the errors of the observed value for the
      student case
  XtS_err <- colSums((t(XtS)-mu)^2) #Vector of the errors of Stein's estimator for the
      student case

  #MSE
  E_Xn_err <- c(E_Xn_err,mean(Xn_err))
  E_XnS_err <- c(E_XnS_err,mean(XnS_err))
  E_Xt_err <- c(E_Xt_err,mean(Xt_err))
  E_XtS_err <- c(E_XtS_err,mean(XtS_err))


  #Squared norm of mu (for the x-coordinates in the plot)
  Norm_mu <- c(Norm_mu,normsq(mu))

}

plot(Norm_mu,E_Xn_err, lty=1, type = "l",ylim = c(0,30),
     main = NULL,ylab = "MSE",xlab = "||mu||^2")
```

```
lines(Norm_mu,E_XnS_err, lty=2)
lines(Norm_mu,E_Xt_err, lty=3)
lines(Norm_mu,E_XtS_err, lty=4)
legend("bottomright", legend=c("Normal: Observed value" ,"Normal: Stein's estimator","
    Student: Observed value" ,"Student: Stein's estimator"),
       lty = 1:4,cex = 0.8)



##########################
##  General covariance  ##
##########################

#########################
##  p=50 and n varies  ##
#########################

p=50
N=5000

#Vectors that will contain the MSE
E_X_err <- NULL
E_XS_err <- NULL

#Mean
mu <- rep(1,p)

##  The different types of covariance matrix  ##
Sigma <- diag(c(rep(1,p/2),rep(5,p/2)))
#Sigma<-sigma_block(p)
#Sigma <- regressive_sigma(p)

#inverse of the covariance matrix
Sigma_inv <- solve(Sigma)

### n < p: the Moore-Penrose inverse is used ###

nrange_smaller <- c(seq(5,45,by=5),49)

for(n in nrange_smaller){

  #Sample of size N to estimate the MSE
  X <- mvrnorm(n=N,mu = mu,Sigma = Sigma)

  #Vectors that will contain the errors
  X_err <- NULL
  XS_err <- NULL

  for(j in 1:N){
    #Sample of size n to compute S, so the mean is set to 0
    V <- mvrnorm(n=n+1,mu = rep(0,p),Sigma = Sigma)

    S <- n*cov(V)
    S_MP <- ginv(S) #Moore-Penrose inverse

    # Stein's estimator
    denom <- as.numeric((t(X[j,])%*%S_MP)%*%X[j,])
    coef <- (n-2)/((p-n+3)*denom)
    XSj <- (diag(p)- coef*(S%*%S_MP))%*%X[j,]

    #Square errors
    X_err <- c(X_err,err_sigma(X[j,],mu,Sigma_inv))
    XS_err <- c(XS_err,err_sigma(XSj,mu,Sigma_inv))
  }
  #MSE
  E_X_err <- c(E_X_err,mean(X_err))
  E_XS_err <- c(E_XS_err,mean(XS_err))

}

### p <= n: the real inverse can be used ###

nrange_bigger <- c(50,51,seq(55,100,by=5))

for(n in nrange_bigger){

  #Sample of size N to estimate the MSE
  X <- mvrnorm(n=N,mu = mu,Sigma = Sigma)

  #Vectors that will contain the errors
  X_err <- NULL
  XS_err <- NULL

  for(j in 1:N){

    #Sample of size n to compute S, so the mean is set to 0
    V <- mvrnorm(n=n+1,mu = rep(0,p),Sigma = Sigma)

    S <- n*cov(V)
    S_inv <- solve(S) #Inverse of S

    # Stein's estimator
    denom <- as.numeric((t(X[j,])%*%S_inv)%*%X[j,])
    coef <- (p-2)/((n-p+3)*denom)
    XSj <- (1- coef)*X[j,]

    #Square errors
    X_err <- c(X_err,err_sigma(X[j,],mu,Sigma_inv))
    XS_err <- c(XS_err,err_sigma(XSj,mu,Sigma_inv))
  }

  #MSE
  E_X_err <- c(E_X_err,mean(X_err))
  E_XS_err <- c(E_XS_err,mean(XS_err))

}

plot(c(nrange_smaller,nrange_bigger),E_X_err, lty = 1 , type = "l",ylim = c(0,55),
     main = NULL,ylab = "MSE",xlab = "n")
lines(c(nrange_smaller,nrange_bigger),E_XS_err, lty=2)
abline(v=50)
legend("bottomleft", legend=c("Observed value" , "Stein's estimator"),
       lty = c(1,3),cex = 0.8)

######################
###  When p varies ###
######################

#Takes quite some time...
N=5000

#Dimensions
prange <- seq(10,80,by=10)

#Vectors that will contain the MSE
E_X_err <- NULL
E_XS1_err <- NULL
E_XS2_err <- NULL
```

```
for(p in prange){

  n1 <- p-1
  n2 <- p/2
  mu <- rep(1,p)

  Sigma <- sigma_block(p) #Block covariance

  Sigma_inv <- solve(Sigma)

  X <- mvrnorm(n=N,mu = mu,Sigma = Sigma) #Sample of size N to estimate the MSE

  X_err <- NULL #Vector that will contain the errors of the observed value
  XS1_err <- NULL #Vector that will contain the errors of Stein's estimator using S1 (n=p-
      1)
  XS2_err <- NULL #Vector that will contain the errors of Stein's estimator using S2 (n=p/
      2)

  for(j in 1:N){

    #Sample of size n1 = p-1 to compute S, so the mean is set to 0
    V1 <- mvrnorm(n=n1+1,mu = rep(0,p),Sigma = Sigma)

    S1 <- n1*cov(V1)
    S1_MP <- ginv(S1)

    #Sample of size n2 = p/2 to compute S, so the mean is set to 0
    V2 <- mvrnorm(n=n2+1,mu = rep(0,p),Sigma = Sigma)

    S2 <- n2*cov(V2)
    S2_MP <- ginv(S2)

    # Stein's estimators
    denom1 <- as.numeric((t(X[j,])%*%S1_MP)%*%X[j,])
    coef1 <- (n1-2)/((p-n1+3)*denom1)

    denom2 <- as.numeric((t(X[j,])%*%S2_MP)%*%X[j,])
    coef2 <- (n2-2)/((p-n2+3)*denom2)

    XS1j <- (diag(p)- coef1*(S1%*%S1_MP))%*%X[j,]
    XS2j <- (diag(p)- coef2*(S2%*%S2_MP))%*%X[j,]

    #Square error
    X_err <- c(X_err,err_sigma(X[j,],mu,Sigma_inv))
    XS1_err <- c(XS1_err,err_sigma(XS1j,mu,Sigma_inv))
    XS2_err <- c(XS2_err,err_sigma(XS2j,mu,Sigma_inv))
  }

  #MSE
  E_X_err <- c(E_X_err,mean(X_err))
  E_XS1_err <- c(E_XS1_err,mean(XS1_err))
  E_XS2_err <- c(E_XS2_err,mean(XS2_err))

}

plot(prange,E_X_err, lty=1, type = "l",
     main = NULL,ylab = "MSE",xlab = "Dimension")
lines(prange,E_XS1_err, lty=2)
lines(prange,E_XS2_err, lty=3)
legend("topleft", legend=c("Observed value" ,"Stein with n=p-1", " Stein with n=p/2"),
       lty = 1:3,cex = 0.8)


##########################
##  p=50 and mean vary  ##
```

```
##########################

N=5000

#Dimension and sample sizes for the estimation of S
p=50
n1 <- p-1
n2 <- p/2

#Vectors that will contain the MSE
E_X_err <- NULL
E_XS1_err <- NULL
E_XS2_err <- NULL

Sigma<-sigma_block(p)

Sigma_inv <- solve(Sigma)

range <- sqrt(0:40/10)

Norm_mu <- NULL

for(m in range){

  mu <- m*rep(1,p) #Mean vector

  X <- mvrnorm(n=N,mu = mu,Sigma = Sigma) #Sample of size N to estimate the MSE

  X_err <- NULL #Vector that will contain the errors of the observed value
  XS1_err <- NULL #Vector that will contain the errors of Stein's estimator using S1 (n=p-
      1)
  XS2_err <- NULL #Vector that will contain the errors of Stein's estimator using S2 (n=p/
      2)

  for(j in 1:N){

    #Sample of size n1 = p-1 to compute S, so the mean is set to 0
    V1 <- mvrnorm(n=n1+1,mu = rep(0,p),Sigma = Sigma)

    S1 <- n1*cov(V1)
    S1_MP <- ginv(S1) #Moore-Penrose inverse

    #Sample of size n2 = p/2 to compute S, so the mean is set to 0
    V2 <- mvrnorm(n=n2+1,mu = rep(0,p),Sigma = Sigma)

    S2 <- n2*cov(V2)
    S2_MP <- ginv(S2)

    # Stein's estimators
    denom1 <- as.numeric((t(X[j,])%*%S1_MP)%*%X[j,])
    coef1 <- (n1-2)/((p-n1+3)*denom1)

    denom2 <- as.numeric((t(X[j,])%*%S2_MP)%*%X[j,])
    coef2 <- (n2-2)/((p-n2+3)*denom2)

    XS1j <- (diag(p)- coef1*(S1%*%S1_MP))%*%X[j,]
    XS2j <- (diag(p)- coef2*(S2%*%S2_MP))%*%X[j,]

    #Square error
    X_err <- c(X_err,err_sigma(X[j,],mu,Sigma_inv))
    XS1_err <- c(XS1_err,err_sigma(XS1j,mu,Sigma_inv))
    XS2_err <- c(XS2_err,err_sigma(XS2j,mu,Sigma_inv))
  }

  #MSE
  E_X_err <- c(E_X_err,mean(X_err))
  E_XS1_err <- c(E_XS1_err,mean(XS1_err))
  E_XS2_err <- c(E_XS2_err,mean(XS2_err))

  #Squared norm of mu (for the x-coordinates in the plot)
  Norm_mu <- c(Norm_mu,normsq(mu))

}

plot(Norm_mu,E_X_err, lty=1, type = "l",ylim = c(10,55),
     main = NULL,ylab = "MSE",xlab = "||mu||^2")
lines(Norm_mu,E_XS1_err, lty=2)
lines(Norm_mu,E_XS2_err, lty=3)
legend("bottomright", legend=c("Observed value" ,"Stein with n=p-1", " Stein with n=p/2"),
       lty = 1:3,cex = 0.8)
```

Bibliography

[1] A. Baranchik. Multiple regression and estimation of the mean of a multivariate normal distribution. techreport 51, Stanford University, May 1964.

[2] D. Chételat and M. T. Wells. Improved multivariate normal mean estimation with unknown covariance when p is greater than n. *The Annals of Statistics*, 40(6):3137–3160, 2012.

[3] B. Efron and C. Morris. Stein's estimation rule and its competitors–an empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

[4] B. Efron and C. Morris. Stein's paradox in statistics. *Scientific American - SCI AMER*, 236:119–127, 05 1977.

[5] H. Federer. *Geometric Measure Theory.* Springer Berlin Heidelberg, 1969.

[6] D. Fourdrinier, W. Strawderman, and M. Wells. Robust shrinkage estimation for elliptically symmetric distributions with unknown covariance matrix. *Journal of Multivariate Analysis*, 85:24–39, 02 2003.

[7] G. H. Golub and V. Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis*, 10(2):413–432, 1973.

[8] A. M. Kshirsagar. Bartlett decomposition and wishart distribution. *Ann. Math. Statist.*, 30(1):239–241, 03 1959.

[9] P.-E. Lin and H.-L. Tsai. Generalized bayes minimax estimators of the multivariate normal mean with unknown covariance matrix. *Ann. Statist.*, 1(1):142–145, 01 1973.

[10] J. R. Magnus. On certain moments relating to ratios of quadratic forms in normal variables: Further results. *Sankhyā: The Indian Journal of Statistics*, 52(1):1–13, Apr. 1990.

[11] M. Srivastava. Singular wishart and multivariate beta distributions. *Ann. Statist.*, 31(5):1537–1560, 10 2003.

[12] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In U. of California Press, editor, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1: Contributions to the Theory of Statistics, pages 197–206, Berkeley, Calif., 1956.

[13] C. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, Nov. 1981.

[14] C. Stein and W. James. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1: Contributions to the Theory of Statistics, pages 361–379, Berkeley, Calif., 1961. University of California Press.

[15] W. E. Strawderman. *On minimax estimation of a normal mean vector for general quadratic loss*, volume 42 of *Lecture Notes–Monograph Series*, pages 3–14. Institute of Mathematical Statistics, Beachwood, OH, 2003.

[16] Y. Tian and S. C. †. Some identities for moore–penrose inverses of matrix products. *Linear and Multilinear Algebra*, 52(6):405–420, 2004.

[17] R. A. Wijsman. Random orthogonal transformations and their use in some classical distribution problems in multivariate analysis. *Ann. Math. Statist.*, 28(2):415–423, 06 1957.

[18] Wikipedia contributors. Moore–penrose inverse — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Moore%E2%80%93Penrose_inverse&oldid=893661833`, 2019. [Online; accessed 29-May-2019].

[19] Wikipedia contributors. Proofs involving the moore–penrose inverse — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Proofs_involving_the_Moore%E2%80%93Penrose_inverse&oldid=892842831`, 2019. [Online; accessed 29-May-2019].

[20] Wikipédia. Décomposition en valeurs singulières — wikipédia, l'encyclopédie libre, 2019. [En ligne; Page disponible le 7-avril-2019].