High-dimensionality effects in the Markowitz problem and other quadratic programs with linear equality constraints: risk underestimation

Noureddine El Karoui
Department of Statistics, UC Berkeley*

August 26, 2009

Abstract

We study the properties of solutions of quadratic programs with linear equality constraints whose parameters are estimated from data in the high-dimensional setting where p, the number of variables in the problem, is of the same order of magnitude as n, the number of observations used to estimate the parameters. The Markowitz problem in Finance is a subcase of our study. Assuming normality and independence of the observations we relate the efficient frontier computed empirically to the "true" efficient frontier. Our computations show that there is a separation of the errors induced by estimating the mean of the observations and estimating the covariance matrix. In particular, the price paid for estimating the covariance matrix is an underestimation of the variance by a factor roughly equal to $1 - p/n$. Therefore the risk of the optimal population solution is underestimated when we estimate it by solving a similar quadratic program with estimated parameters.

We also characterize the statistical behavior of linear functionals of the empirical optimal vector and show that they are biased estimators of the corresponding population quantities.

We investigate the robustness of our Gaussian results by extending the study to certain elliptical models and models where our n observations are correlated (in "time"). We show a lack of robustness of the Gaussian results, but are still able to get results concerning first order properties of the quantities of interest, even in the case of relatively heavy-tailed data (we require two moments). Risk underestimation is still present in the elliptical case and more pronounced that in the Gaussian case.

We discuss properties of the non-parametric and parametric bootstrap in this context. We show several results, including the interesting fact that standard applications of the bootstrap generally yields inconsistent estimates of bias.

Finally, we propose some strategies to correct these problems and practically validate them in some simulations. In all the paper, we will assume that p, n and $n - p$ tend to infinity, and $p < n$.

1 Introduction

Many statistical estimation problems are now formulated, implicitly or explicitly, as solutions of certain optimization problems. Naturally, the parameters of these problems tend to be estimated from data and it is therefore important that we understand the relationship between the solutions of two types of optimization problems: those which use the population parameters and those which use the estimated parameters. This question is particularly relevant in high-dimensional inference where one suspects that the differences between the two solutions might be considerable. The aim of this paper is to contribute to this understanding by focusing on quadratic programs with linear equality constraints. An important example of such a program where our questions are very natural is the celebrated Markowitz optimization problem in Finance which will serve as a supporting example throughout the paper.

*I am very grateful to Nizar Touzi and Nicole El Karoui for several very interesting discussions at the beginning of this project and for their interest in it. Support from the France-Berkeley Fund, a Sloan research Fellowship as well as NSF grants DMS-0605169 and DMS-0847647 (CAREER) are gratefully acknowledged. **AMS 2000 SC:** Primary: 62H10. Secondary: 90C20 **Key words and Phrases :** covariance matrices, convex optimization, quadratic programs, multivariate statistical analysis, high-dimensional inference, random matrix theory, concentration of measure, Markowitz problem, Wishart matrices, elliptical distributions. **Contact :** nkaroui@stat.berkeley.edu

The Markowitz problem (Markowitz (1952)) is a classic portfolio optimization problem in Finance, where investors choose to invest according to the following framework: one picks assets in such a way that the portfolio guarantees a certain level of expected returns but minimizes the "risk" associated with them. In the standard framework, this risk is measured the variance of the portfolio.

Markowitz's paper was highly influential and much work has followed. It is now part of the standard textbook literature on these issues (Ruppert (2006), Campbell et al. (1996)). Let us recall the setup of the Markowitz problem.

- We have the opportunity to invest in p assets, $A_1, \ldots, A_p$
- In the ideal situation, the mean returns are known and represented by a p-dimensional vector, μ.
- Also, the covariance between the returns is known; we denote it by Σ
- We want to create a portfolio, with guaranteed mean return μ_P, and minimize its risk, as measured by variance.
- The question is how should items be weighted in portfolio? What are weights w?

We note that Σ is positive semi-definite and hence is in particular symmetric. In the ideal (or population) solution, the covariance and the mean are known. The mathematical formulation is then the following simple quadratic program. We wish to find the weights w that solve the following problem:

$$\begin{cases} \min \frac{1}{2} w' \Sigma w \\ w' \mu = \mu_P \ , \\ w' \mathbf{e} = 1 \end{cases}$$

Here, $\mathbf{e}$ is a p-dimensional vector with 1 in every entry. If Σ is invertible, the solution is known explicitly (see Section 2). If we call w_{optimal} the solution of this problem, the curve $w'_{\text{optimal}} \Sigma w_{\text{optimal}}$, seen as a function of μ_P is called the *efficient frontier*.

Of course, in practice, we do not know μ and Σ and we need to estimate them. An interesting question is therefore to know what happens in the Markowitz problem when we replace population quantities by corresponding estimators.

Naturally, we can ask a similar question for general quadratic programs with linear equality constraints (see below or Boyd and Vandenberghe (2004) for a definition), the Markowitz problem being a particular instance of such a problem. This paper provides an answer to these questions under certain distributional assumptions on the data.

It has been observed by many that there are problems in practice when replacing population quantities by standard estimators (see Lai and Xing (2008), section 3.5), and alternatives have been proposed. A famous one is the Black-Litterman model (Black and Litterman (1990) and e.g Meucci (2008)). Adjustments to the standard estimators have also been proposed: Ledoit and Wolf (2004), partly motivated by portfolio optimization problems, proposed to "shrink" the sample covariance matrix towards another positive definite matrix (often the identity matrix properly scaled), while Michaud (1998) proposed to use the bootstrap and to average bootstrap weights to find better-behaved weights for the portfolio. As noted in Lai and Xing (2008), there is a dearth of theoretical studies regarding, in particular, the behavior of bootstrap estimators.

An aspect of the problem that is of particular interest to us is the study of large-dimensional portfolios (or quadratic programs with linear equality constraints). To make matters clear, we focus on a portfolio with $p = 100$ assets. If we use a year of daily data to estimate Σ, the covariance between the daily returns of the assets, we have $n \simeq 250$ observations at our disposal. In modern statistical parlance, we are therefore in a "large n, large p" setting, and we know from random matrix theory that $\widehat{\Sigma}$, the sample covariance matrix is a poor estimator of Σ, especially when it comes to spectral properties of Σ. There is now a developing statistical literature on properties of sample covariance matrices when n and p are both large - and it is now understood that, though $\widehat{\Sigma}$ is unbiased for Σ, the eigenvalues and eigenvectors of $\widehat{\Sigma}$ behave very differently from those of Σ. We refer the interested reader to Johnstone (2001), El Karoui (2007), El Karoui (2008a), Bickel and Levina (2007a), Rothman et al. (2008), El Karoui (2009) for a partial introduction to these problems. We wish with this study to make clear that the "large n, large p" character of the problem has an important impact of the empirical solution of the problem. By contrast, standard but thorough discussions of these problems (Meucci, 2005) give only a cursory treatment of dimensionality issues (e.g one page out of a whole book).

Another interesting aspect of this problem is that the high-dimensional setting does not allow, by contrast to the classical "small p, large n" setting, a perturbative approach to go through. In the "small p, large n" setting, the paper Jobson and Korkie (1980) is concerned, in the Gaussian case, with issues similar to the ones we will be investigating.

The "large n, large p" setting is the one with which random matrix theory is concerned - and the high-dimensional Markowitz problem has therefore been of interest to random matrix theorists for some time now. We note in particular the paper Laloux et al. (2000), where a random matrix-inspired (shrinkage) approach to improved estimation of the sample covariance matrix is proposed in the context of the Markowitz problem. We also note that other random-matrix based approaches to covariance estimation were later proposed (El Karoui (2008b)), with asymptotic theoretical guarantees on the estimation of the spectral distribution of the covariance matrix.

Let us now remind the reader of some basic facts of random matrix theory that suggests that serious problems may arise if one solves naively the high-dimensional Markowitz problem or other quadratic programs with linear equality constraints. A key result in random matrix theory is the Marčenko-Pastur equation (Marčenko and Pastur (1967)) which characterizes the limiting distribution of the eigenvalues of the sample covariance matrix and relates it to the spectral distribution of the population covariance matrix. We give only in this introduction its simplest form and refer the reader to Marčenko and Pastur (1967), El Karoui (2008b) and El Karoui (2009) for a more thorough introduction and very recent developments, as well as potential geometric and statistical limitations of the models usually considered in random matrix theory.

In the simplest setting, we consider data $\{X_i\}_{i=1}^n$, which are p-dimensional. In a financial context, these vectors would be vectors of (log)-returns of assets, the portfolio consisting of p assets. To simplify the exposition, let us assume that the X_i's are i.i.d with distribution $\mathcal{N}(0, \mathrm{Id}_p)$. We call X the $n \times p$ matrix whose i-th row is the vector X_i. Let us consider the sample covariance matrix

$$\widehat{\Sigma} = \frac{1}{n-1}(X - \bar{X})'(X - \bar{X}) ,$$

where $\bar{X}$ is a matrix whose rows are all equal to the column mean of X. Now let us call F_p the spectral distribution of $\widehat{\Sigma}$, i.e the probability distribution that puts mass $1/p$ at each of the p eigenvalues of $\widehat{\Sigma}$. A graphical representation of this probability distribution is naturally the histogram of eigenvalues of $\widehat{\Sigma}$. A consequence of the main result of the very profound paper Marčenko and Pastur (1967) is that F_p, though a random measure, is asymptotically non-random, and its limit, in the sense of weak convergence of distributions, F has a density (when $p < n$) that can be computed. F depends on $\rho = \lim_{n\to\infty} p/n$ in the following manner: if $p < n$, the density of F is

$$f_\rho(x) = \frac{1}{2\pi\rho}\frac{\sqrt{(y_+ - x)(x - y_-)}}{x} 1_{y_- \leq x \leq y_+} ,$$

where $y_+ = (1 + \sqrt{\rho})^2$ and $y_- = (1 - \sqrt{\rho})^2$. Figure 1 presents a graphical illustration of this result.

What is striking about this result is that it implies that the largest eigenvalue of Σ, λ_1 will be overestimated by l_1 the largest eigenvalue of $\widehat{\Sigma}$. Also, the smallest eigenvalue of Σ, λ_p will be underestimated by the smallest eigenvalue of $\widehat{\Sigma}$, l_p. As a matter of fact, in the model described above, Σ has all its eigenvalues equal to 1, so $\lambda_1(\Sigma) = \lambda_p(\Sigma) = 1$, while l_1 will asymptotically be larger or equal to $(1+\sqrt{\rho})^2$ and l_p smaller or equal to $(1-\sqrt{\rho})^2$ (in the Gaussian case and several others, l_1 and l_p converge to those limits). We note that the result of Marčenko and Pastur (1967) is not limited to the case where Σ is identity, as presented here, but holds for general covariance Σ (F_p has of course a different limit then).

Perhaps more concretely, let us consider a projection of the data along a vector v, with $\|v\|_2 = 1$, where $\|v\|_2$ is the Euclidian norm of v. Here it is clear that, if $X \sim \mathcal{N}(0, \mathrm{Id}_p)$, $\mathrm{var}\,(v'X) = 1$, for all v, since $v'X \sim \mathcal{N}(0, 1)$. However, if we do not know Σ and estimate it by $\widehat{\Sigma}$, a naive (and wrong) reasoning

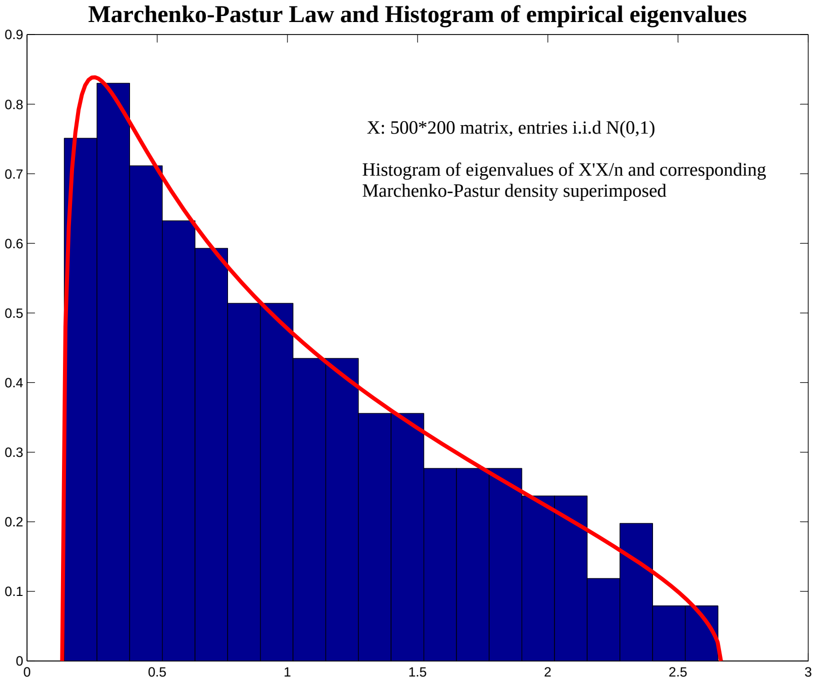


Figure 1: Illustration of Marčenko-Pastur law, n=500, p=200. The red curve is the density of the Marčenko-Pastur -law for $\rho = 2/5$. The simulation was done with i.i.d Gaussian data. The histogram is the histogram of eigenvalues of $X'X/n$

suggests that we can find direction of lower variance than 1, namely those corresponding to eigenvectors of $\widehat{\Sigma}$ associated with eigenvalues that are less than 1. In particular, if v_p is the eigenvector associated with l_p, the smallest eigenvalue of $\widehat{\Sigma}$, by naively estimating, for X independent of $\{X_i\}_{i=1}^n$, the variance in the direction of v_p, var $\left(v_p'X\right)$, by the empirical version $v_p'\widehat{\Sigma}v_p$, one would commit a severe mistake: the variance in any direction is 1, but it would be estimated by something roughly equal to $(1-\sqrt{p/n})^2$ in the direction of v_p.

In a portfolio optimization context, this suggests that by using standard estimators, such as the sample covariance matrix, when solving the high-dimensional Markowitz problem, one might underestimate the variance of certain portfolios (or "optimal" vectors of weights). As a matter of fact, in the previous toy example, thinking (wrongly) that there is low variance in the direction v_p, one might (numerically) "load" this direction more than warranted, given that the true variance is the same in all directions.

This simple argument suggests that severe problems might arise in the high-dimensional Markowitz problem and other quadratic programs with linear equality constraints, and in particular, risk might be underestimated. While this heuristic argument is probably clear to specialists of random matrix theory, the problem has not been investigated at a mathematical level of rigor in that literature. It has received some attention at a physical level of rigor (see e.g Pafka and Kondor (2003), where the authors treat only the Gaussian case, and neglect the effect of the mean, which as we show below creates problems of its own - we also provide exact distributional results in the Gaussian case). In this paper, we propose a theoretical analysis of the problem in a Gaussian and elliptical framework for general quadratic programs with linear equality constraints, one of them involving the parameter μ. Our results and contributions are several-fold. We relate the empirical efficient frontier to the theoretical efficient frontier that is key to the Markowitz theory, in a variety of theoretical settings. We show that the empirical frontier generally yields an underestimation of the risk of the portfolio and that Gaussian analysis gives an over-optimistic view of this problem. We show that the expected returns of the naive "optimal" portfolio are poorly estimated by μ_P. We argue that the bootstrap will not solve the problems we are pointing out here. Beside new formulas, we also provide robust estimators of the various quantities we are interested in.

The paper is divided into four main parts and a conclusion. In Section 2, to make the paper self-contained, we discuss the solution of quadratic problems with linear equality constraints - the focus of this paper. In Section 3, we study the impact of parameter estimation on the solution of these problems when the observed data is i.i.d Gaussian and obtain some exact distributional results for fixed p and n. In Section 4, we obtain results in the case where the data is elliptically distributed. This allows us also to understand the impact of correlation between observations in the Gaussian case and to get information about the behavior of the non-parametric bootstrap. In Section 5, we apply the results of Section 4 to the quadratic programs at hand and compare the elliptical and the Gaussian cases. We show, among other things, that the Gaussian results are not robust in the class of elliptical distribution. In particular, two models may yield the same μ and Σ but can have very different empirical behavior. In Section 5, we also propose various schemes to correct the problems we highlight (see pp. 40, 41 and 51 for pictures). The conclusion summarizes our findings and the Appendix contains various facts and proofs that did not naturally flow in the main text or were better highlighted by being stated separately.

Several times in the paper $\widehat{\Sigma}^{-1}$ and Σ^{-1} will appear. Unless otherwise noted, when taking the inverse of a population matrix, we implicitly assume that it exists. The question of existence of inverse of sample covariance matrices is well-understood in the statistics literature. Because our models will have a component with a continuous distribution, there are essentially no existence problems (unless we explicitly mention and treat them) as proofs similar to standard ones found in textbooks (e.g Anderson (2003)) would show. Hence, we do not belabor this point any further in the rest of the paper as our focus is on another things than rather well-understood technical details and the paper is already a bit long.

2 Quadratic programs with linear equality constraints

We discuss here the properties of the solution of quadratic programs with linear equality constraints as they lay the foundations for our analysis of similar problems involving estimated parameters. We included this section for the convenience of the reader to make the paper as self-contained as possible.

The problem we want to solve is the following:

$$\begin{cases} \min_{w\in\mathbb{R}^p} \frac{1}{2} w'\Sigma w \\ w'v_i = u_i \, , \, 1 \leq j \leq k \ . \end{cases} \tag{QP-eqc}$$

Here Σ is a positive definite matrix of size $p \times p$, $v_i \in \mathbb{R}^p$ and $u_i \in \mathbb{R}$. We have the following theorem:

Theorem 2.1. *Let us call V the $p \times k$ matrix whose i-th column is v_i, U the k dimensional vector whose i-th entry is U_i and M the $k \times k$ matrix*

$$M = V'\Sigma^{-1}V \ .$$

We assume that the v_i's are such that M is invertible. The solution of the quadratic program with linear equality constraints (QP-eqc) *is achieved for*

$$w_{\text{optimal}} = \Sigma^{-1}VM^{-1}U$$

and we have

$$w'_{\text{optimal}}\Sigma w_{\text{optimal}} = U'M^{-1}U \ ,$$

Proof. Let us call λ a k dimensional vector of Lagrange multipliers. The Lagrangian function is, in matrix notation,

$$L(w,\lambda) = \frac{w'\Sigma w}{2} - \lambda'(V'w - U) \ .$$

This is clearly a (strictly) convex function in w, since Σ is positive definite by assumption. We have

$$\frac{\partial L}{\partial w} = \Sigma w - V\lambda \ .$$

So $w_{\text{optimal}} = \Sigma^{-1}V\lambda$. Now we know that $U = V'w_{\text{optimal}}$. So $U = V'\Sigma^{-1}V\lambda = M\lambda$. Therefore,

$$w_{\text{optimal}} = \Sigma^{-1}VM^{-1}U \ .$$

We deduce immediately that

$$w'_{\text{optimal}}\Sigma w_{\text{optimal}} = U'M^{-1}U \; .$$

□

We now turn to another result which will prove useful later. It gives a compact representation of linear combinations of the weights of the optimal solution, and we will rely heavily on it in particular in the case of Gaussian data.

Lemma 2.1. *Let us consider w_{optimal} the solution of the optimization problem* (QP-eqc)*. Let γ be a vector in $\mathbb{R}^p$. Let us call $\mathcal{M}$ the $(k+1)\times(k+1)$ matrix that is written in block form*

$$\mathcal{M} = \begin{pmatrix} V'\Sigma^{-1}V & V'\Sigma^{-1}\gamma \\ \gamma'\Sigma^{-1}V & \gamma'\Sigma^{-1}\gamma \end{pmatrix} \; .$$

Assume that $\mathcal{M}$ is invertible. Then

$$\gamma' w_{\text{optimal}} = -\frac{1}{\mathcal{M}^{-1}_{k+1,k+1}} \left(U'0\right) \mathcal{M}^{-1} \binom{0_k}{1} \tag{1}$$

Proof. The proof is a consequence of the results discussed in the appendix concerning inverses of partitioned matrices (see Subsection A-1 and Equation (A-4) there). Let us write

$$\mathcal{M} = \begin{pmatrix} \mathcal{M}_{11} & \mathcal{M}_{12} \\ \mathcal{M}_{21} & \mathcal{M}_{22} \end{pmatrix} \; ,$$

where $\mathcal{M}_{11}$ is $k\times k$, $\mathcal{M}_{12}$ is naturally $k\times 1$ and $\mathcal{M}_{22}$ is a scalar. With the same block notation, we have

$$\mathcal{M}^{-1} = \begin{pmatrix} \mathcal{M}^{11} & \mathcal{M}^{12} \\ \mathcal{M}^{21} & \mathcal{M}^{22} \end{pmatrix} \; .$$

Then, we know (see Equation (A-4)) that $\mathcal{M}^{12} = -\mathcal{M}_{11}^{-1}\mathcal{M}_{12}\mathcal{M}^{22}$, but since $\mathcal{M}^{22}$ is a scalar, equal to $\mathcal{M}^{-1}(k+1,k+1)$, we have

$$\mathcal{M}_{11}^{-1}\mathcal{M}_{12} = -\mathcal{M}^{12}/\mathcal{M}^{22} \; .$$

Now $\mathcal{M}_{11}^{-1}\mathcal{M}_{12} = (V'\Sigma^{-1}V)^{-1}V'\Sigma^{-1}\gamma$, so $U'\mathcal{M}_{11}^{-1}\mathcal{M}_{12} = w'_{\text{optimal}}\gamma$. Hence,

$$w'_{\text{optimal}}\gamma = -\frac{1}{\mathcal{M}^{22}}(U'\,0)\mathcal{M}^{-1}\binom{0_k}{1} \; .$$

□

We note that here $(\mathcal{M}^{22})^{-1} = \gamma'\Sigma^{-1}\gamma - \gamma'\Sigma^{-1}VM^{-1}V'\Sigma^{-1}\gamma$, as an application of Equation (A-2) clearly shows.

3 QP with equality constraints: impact of parameter estimation in the Gaussian case

From now on, we will assume that we are in the high-dimensional setting where p and n go to infinity. Our study will be divided into two. We will first consider the Gaussian setting (in this Section) and then study an elliptical distribution setting (in Section 4). (We note that for the Markowitz problem, the assumption of Gaussianity would be satisfied if we worked under Black-Scholes diffusion assumptions for our assets and were considering log-returns as our observations.) Interestingly, we will show that the results are not robust against the assumption of Gaussianity, which is not (so) surprising in light of recent random matrix results (see El Karoui (2009)). We will also show that understanding the elliptical setting allows us to understand the impact of correlation between observations and to discuss bootstrap-related ideas. In particular, we will see that various problems arise with the bootstrap in high-dimension and that the results change when one deals with observations that are correlated (in time) or not.

Before we proceed, we need to set up some notations: we call $\mathbf{e}$ the p-dimensional vector whose entries are all equal to 1. We call V, as above, the matrix containing all of our constraint vectors, which we may have to estimate (for instance, if $v_i = \mu$ for a certain i). We call $\widehat{V}$ the matrix of estimated constraint vectors.

The template question for all our investigations will be the following (Markowitz) question: what can be said of the statistical properties of the solution of

$$\begin{cases} \min_{w\in\mathbb{R}^p} w'\widehat{\Sigma}w \\ w'\widehat{\mu} = \mu_P \ , \\ w'\mathbf{e} = 1 \end{cases}$$

compared to the solution of the population version

$$\begin{cases} \min_{w\in\mathbb{R}^p} w'\Sigma w \\ w'\mu = \mu_P \ , \\ w'\mathbf{e} = 1 \end{cases}$$

We will solve the problem at a much greater degree of generality, by considering quadratic programs with linear equality constraints and comparing the solutions of

$$\begin{cases} \min_{w\in\mathbb{R}^p} \frac{1}{2}w'\widehat{\Sigma}w \\ w'v_i = u_i\, , \, 1 \leq j \leq k-1 \ , \\ w'\widehat{\mu} = u_k \end{cases} \qquad \text{(QP-eqc-Emp)}$$

and

$$\begin{cases} \min_{w\in\mathbb{R}^p} \frac{1}{2}w'\Sigma w \\ w'v_i = u_i\, , \, 1 \leq j \leq k-1 \ , \\ w'\mu = u_k \end{cases} \qquad \text{(QP-eqc-Pop)}$$

Here $\widehat{\Sigma}$ and $\widehat{\mu}$ will be estimated from the data. We call w_{emp} the vector that yields a solution of Problem (QP-eqc-Emp) and w_{theo} the vector that yields a solution of Problem (QP-eqc-Pop).

We call $\widehat{V}$ the $p \times k$ matrix containing $\{v_i\}_{i=1}^{k-1}$ and $\widehat{\mu}$, and V its population counterpart, which contains $\{v_i\}_{i=1}^{k-1}$ and μ. We assume that $\{v_i\}_{i=1}^{k-1}$ are deterministic and known (just like the vector $\mathbf{e}$ in the Markowitz problem). In our analysis, k will be held fixed. (The k-th column of $\widehat{V}$ will contain $\widehat{\mu}$ in general or our estimator of μ.)

As should be clear from Theorem 2.1, the properties of the entries of the matrix $\widehat{V}'\widehat{\Sigma}^{-1}\widehat{V}$ as compared to those of the matrix $V'\Sigma^{-1}V$ will be key to our understanding of this question. In what follows, we assume that the vectors $\hat{v}_i$ are either deterministic or equal to $\widehat{\mu}$. The extension to linear combinations of a deterministic vector and $\widehat{\mu}$ is straightforward. We also note that in the Gaussian case, we could just assume that the $\hat{v}_i$ are (deterministic) functions of $\widehat{\mu}$ (because $\widehat{\mu}$ and $\widehat{\Sigma}$ are independent in this case). On the other hand, the vector U is assumed to be deterministic.

Before we proceed, let us mention that after our study was completed, we learned of similar results (restricted to the Markowitz case and not dealing with general quadratic programs with linear equality constraints) by Kan and Smith (2008). We stress the fact that our work was independent of theirs and is more general which is why it is included in the paper.

3.1 Efficient frontier problems

We first study questions concerning the efficient frontier and then turn to information we can get about linear functionals of the empirical weights.

Theorem 3.1. *Let us assume that we observe data* $X_i \overset{iid}{\backsim} \mathcal{N}(\mu, \Sigma)$, *for* $i = 1, \ldots, n$. *Here* Σ *is* $p \times p$ *and* $p < n$. *Suppose we estimate* Σ *with the sample covariance matrix* $\widehat{\Sigma}$, *and* μ *with the sample mean* $\widehat{\mu}$. *Suppose we wish to solve the problem*

$$\begin{cases} \min_{w\in\mathbb{R}^p} w'\Sigma w \\ w'v_j = u_j\, , \, 1 \leq j \leq k \ . \end{cases} , \qquad \text{(QP-eqc-Pop)}$$

where u_j are deterministic, v_j are deterministic and given for $j < k$ and $v_k = \mu$. Assume that we use as a proxy for the previous problem the empirical version with plugged-in parameters. Let us consider the solution of the problem:

$$\begin{cases} \min_{w\in\mathbb{R}^p} w'\widehat{\Sigma}w \\ w'\hat{v}_j = u_j \,,\, 1 \leq j \leq k \ . \end{cases} . \tag{QP-eqc-Emp}$$

Now $\hat{v}_j = v_j$ for $j < k$ and $\hat{v}_k = f(\widehat{\mu})$. Let us call w_{emp} the corresponding "weight" vector. The plug-in estimate of $w'\Sigma w$ is $w'_{\text{emp}}\widehat{\Sigma}w_{\text{emp}}$. Let us call w_{oracle} the optimal solution of the quadratic program obtained under the assumption that Σ is given, but μ is not and is estimated by $f(\widehat{\mu})$. Finally, we assume that $n-1-p+k > 0$.

Then we have

$$\boxed{w'_{\text{emp}}\widehat{\Sigma}w_{\text{emp}} = w'_{\text{oracle}}\Sigma w_{\text{oracle}}\frac{\chi^2_{n-1-p+k}}{n-1}\ ,} \tag{2}$$

where $w'_{\text{oracle}}\Sigma w_{\text{oracle}}$ is random (because $\widehat{\mu}$ is) but is statistically independent of $\chi^2_{n-1-p+k}$. Also,

$$w'_{\text{oracle}}\Sigma w_{\text{oracle}} = U'\left(\widehat{V}'\Sigma^{-1}\widehat{V}\right)^{-1}U\ .$$

The previous theorem means that the cost of not knowing the covariance matrix and estimating it is the apparition of the $\frac{\chi^2_{n-1-p+k}}{n-1}$. In the high-dimensional setting when p and n are of the same order of magnitude and $n-p$ is large, this terms is approximately $1-(p-k)/(n-1)$. Hence, the theorem quantifies the random matrix intuition that having to estimate the high-dimensional covariance matrix at stake here leads to risk **underestimation**, by the factor $1-(p-k)/(n-1)$. In other words, using plug-in procedures leads to over-optimistic conclusions in this situation.

We also note that the previous theorem shows that, in the Gaussian setting under study here, the effect of estimating the mean and the covariance on the solution of the quadratic program are "separable": the effect of the mean estimation is in the oracle term, while the effect of estimating the covariance is in the $\chi^2_{n-p-1+k}/(n-1)$ term. To show risk underestimation, it will therefore be necessary to relate $w'_{\text{oracle}}\Sigma w_{\text{oracle}}$ to $w'_{\text{theo}}\Sigma w_{\text{theo}}$. We do it in Proposition 3.1 but first give a proof of Theorem 3.1.

Proof of Theorem 3.1: The crux of the proof is the following result, which is well-known of statisticians, concerning (essentially) blocks of the inverse of a Wishart matrix: if $S \sim \mathcal{W}_p(\Sigma, m)$, i.e S is a $p\times p$ Wishart matrix with m degree of freedoms and covariance Σ, and A is $p\times k$, deterministic matrix, then, when $m > p$,

$$(A'S^{-1}A)^{-1} \sim \mathcal{W}_k((A'\Sigma^{-1}A)^{-1}, m-p+k)\ .$$

We refer to Eaton (1983), Proposition 8.9 p. 312 for a proof, and to Mardia et al. (1979), pp.70-73 for related results.

Another important remark is the well-known fact that, in the situation we are considering, $\widehat{\mu}$ is $\mathcal{N}(\mu, \Sigma/n)$ and independent of $\widehat{\Sigma}$. Finally, it is also well known that if $S \sim \mathcal{W}_p(\Sigma, m)$ and U is a p-dimensional deterministic vector, then $U'SU = U'\Sigma U\chi^2_m$.

Now $\widehat{\Sigma} \sim \mathcal{W}_p(\Sigma, n-1)/(n-1)$. Therefore, since $\widehat{V}$ is a function of $\widehat{\mu}$, we have, by independence of $\widehat{\mu}$ and $\widehat{\Sigma}$,

$$(\widehat{V}'\widehat{\Sigma}^{-1}\widehat{V})^{-1}\Big|\widehat{\mu} \sim \mathcal{W}_k((\widehat{V}'\Sigma^{-1}\widehat{V})^{-1}, n-1-p+k)/(n-1)\ .$$

Therefore,

$$\left.\frac{U'(\widehat{V}'\widehat{\Sigma}^{-1}\widehat{V})^{-1}U}{U'(\widehat{V}'\Sigma^{-1}\widehat{V})^{-1}U}\right|\widehat{\mu} \sim \frac{\chi^2_{n-p-1+k}}{n-1}\ .$$

Because the right hand side does not depend on $\widehat{\mu}$, we have established the independence of

$$\frac{U'(\widehat{V}'\widehat{\Sigma}^{-1}\widehat{V})^{-1}U}{U'(\widehat{V}'\Sigma^{-1}\widehat{V})^{-1}U} \text{ and } \frac{\chi^2_{n-p-1+k}}{n-1}\ .$$

Hence, we conclude that

$$U'(\widehat{V}'\widehat{\Sigma}^{-1}\widehat{V})^{-1}U = U'(\widehat{V}'\Sigma^{-1}\widehat{V})^{-1}U\frac{\chi^2_{n-p-1+k}}{n-1} ,$$

and the two terms are independent. Now the term $U'(\widehat{V}'\Sigma^{-1}\widehat{V})^{-1}U$ is the estimate we would get for the solution of Problem (QP-eqc-Pop), if Σ were known and μ were estimated by $f(\widehat{\mu})$. In other words, it is the "oracle" solution described above. □

3.1.1 Some remarks on the oracle solution

Theorem 3.1 sheds light on the separate effects of mean and covariance estimation on the problem considered above. To understand further the problem of risk estimation, we need to better understand the role the estimation of the mean might play. This is what we do now.

Proposition 3.1. *Suppose that the last column of $\widehat{V}$ is $\widehat{\mu}$. Let us call V_{-k} the $p\times k-1$ dimensional matrix whose j-th column is v_j, which are known deterministic vectors. Suppose that $M = V'\Sigma^{-1}V = \mathrm{O}(1)$. Suppose further that $\lambda_k(V'\Sigma^{-1}V) \gg n^{-1/2}$, where $\lambda_k(S)$ is the smallest eigenvalue of the $k \times k$ matrix S. Further, call $M = V'\Sigma^{-1}V \in \mathbb{R}^{k\times k}$ and call e_i the canonical basis vectors in $\mathbb{R}^k$. Finally, call $\alpha = \chi^2_p/n$. Then, when $p/n \to \rho \in (0,1)$, asymptotically,*

$$w'_{\text{oracle}}\Sigma w_{\text{oracle}} = w'_{\text{theo}}\Sigma w_{\text{theo}} - \alpha\frac{(U'M^{-1}e_k)^2}{1+\alpha e_k' M^{-1}e_k} + \mathrm{o}_P\left(w'_{\text{theo}}\Sigma w_{\text{theo}}\right) .$$

Let us discuss a little bit this result before we provide a proof. In the asymptotics we have in mind and are considering, $p/n \to \rho \in (0,1)$ and therefore $\alpha \simeq p/n + \mathrm{O}(n^{-1/2})$. So if $\delta_n = (U'M^{-1}e_k)^2/(1 + p/ne_k'M^{-1}e_k)$, when the above analysis applies, the impact of the estimation of μ by $\widehat{\mu}$ will be risk underestimation, just as is the case for the case of the covariance matrix. Here, we can also quantify the impact of this estimation of μ by $\widehat{\mu}$: it leads to risk underestimation by the amount $\alpha\delta_n$.

Proof. Let us write $\widehat{\mu} = \mu + e$, where $e \sim \mathcal{N}(0, \Sigma/n)$. Clearly, $e = n^{-1/2}\Sigma^{1/2}Z$, where Z is $\mathcal{N}(0, \mathrm{Id}_p)$. We have, using block notations,

$$\widehat{V}'\Sigma^{-1}\widehat{V} = V'\Sigma^{-1}V + \begin{pmatrix} 0 & 0 \\ 0 & e'\Sigma^{-1}e \end{pmatrix} + \begin{pmatrix} 0 & V_{-k}'\Sigma^{-1}e \\ e'\Sigma^{-1}V_{-k} & 2\mu'\Sigma^{-1}e \end{pmatrix}$$

Replacing e by its value, we have $\mu'\Sigma^{-1}e \sim \mathcal{N}(0, \mu'\Sigma^{-1}\mu/n)$. By the same token, we can also get that

$$V_{-k}'\Sigma^{-1}e = \frac{1}{\sqrt{n}}V_{-k}'\Sigma^{-1/2}Z \sim \mathcal{N}\left(0, \frac{V_{-k}'\Sigma^{-1}V_{-k}}{n}\right) .$$

Our assumption that $V'\Sigma^{-1}V = \mathrm{O}(1)$ implies that $\mu'\Sigma^{-1}\mu = \mathrm{O}(1)$ and $V_{-k}'\Sigma^{-1}V_{-k} = \mathrm{O}(1)$. Therefore,

$$\begin{pmatrix} 0 & V_{-k}'\Sigma^{-1}e \\ e'\Sigma^{-1}V_{-k} & 2\mu'\Sigma^{-1}e \end{pmatrix} = \mathrm{O}_P\left(\frac{1}{\sqrt{n}}\right) .$$

Hence, since $e'\Sigma^{-1}e = Z'Z/n = \alpha$,

$$\widehat{V}'\Sigma^{-1}\widehat{V} = V'\Sigma^{-1}V + \alpha e_k e_k' + \mathrm{O}_P(n^{-1/2}) .$$

Our assumptions guarantee that $\lambda_k(V'\Sigma^{-1}V) \gg n^{-1/2}$, and therefore $\lambda_k(V'\Sigma^{-1}V + \alpha e_k e_k') \gg n^{-1/2}$. In other respects, let A be a matrix such that $\lambda_p(A) \gg n^{-1/2}$ and E be a matrix such that $E = \mathrm{O}(n^{-1/2})$. Recall that for symmetric matrices, $\lambda_p(A+E) \geq \lambda_p(A)+\lambda_p(E)$ (see e.g Weyl's Theorem, Horn and Johnson (1994), p.185). So in this situation, $(A+E)^{-1} = \mathrm{o}(n^{1/2})$. Let us now consider the implications of this remark on the difference of $(A+E)^{-1}$ and A^{-1}. We claim that $(A+E)^{-1} = A^{-1} + \mathrm{o}(A^{-1})$. By the first resolvent identity, $(A+E)^{-1} = A^{-1} - (A+E)^{-1}EA^{-1}$; our previous remark implies that $\sigma_1[(A+E)^{-1}E] = \mathrm{o}(1)$ and the result follows. Applying the results of this discussion to $A = V'\Sigma^{-1}V + \alpha e_k e_k'$ and $A + E = \widehat{V}'\Sigma^{-1}\widehat{V}$, we have

$$\widehat{V}'\Sigma^{-1}\widehat{V} = (V'\Sigma^{-1}V + \alpha e_k e_k')^{-1} + \mathrm{o}_P((V'\Sigma^{-1}V + \alpha e_k e_k')^{-1}) \ .$$

We can now use well-known results concerning inverses of rank-1 perturbation of matrices, namely

$$(V'\Sigma^{-1}V + \alpha e_k e_k')^{-1} = (M + \alpha e_k e_k')^{-1} = M^{-1} - \alpha \frac{M^{-1}e_k e_k' M^{-1}}{1 + \alpha e_k' M^{-1} e_k} \ .$$

This allows us to conclude that

$$U'(\widehat{V}'\Sigma^{-1}\widehat{V})^{-1}U = U'M^{-1}U - \alpha \frac{(U'M^{-1}e_k)^2}{1 + \alpha e_k' M^{-1} e_k} + \mathrm{o}_P(U'M^{-1}U) \ .$$

This is the result announced in the theorem and the proof is complete. □

We can now combine the results of Theorem 3.1 and Proposition 3.1 to obtain the following corollary.

Corollary 3.1. *We assume that the assumptions of Theorem 3.1 and Proposition 3.1 hold and that p/n has a finite non-zero limit, as $n \to \infty$, and $n - p$ tends to infinity. Then we have,*

$$\boxed{w_{\text{emp}}\widehat{\Sigma}w_{\text{emp}} = \left(1 - \frac{p-k}{n-1}\right)\left(w_{\text{theo}}'\Sigma w_{\text{theo}} - \frac{p}{n}\frac{(U'M^{-1}e_k)^2}{1 + \frac{p}{n}e_k'M^{-1}e_k}\right) + \mathrm{o}_P\left(w_{\text{theo}}'\Sigma w_{\text{theo}} \vee n^{-1/2}\right) \ ,} \qquad (3)$$

where M is the population quantity $M = V'\Sigma^{-1}V$.

The corollary shows that the effects of both covariance and mean estimation are to underestimate the risk, and the empirical frontier is asymptotically deterministic.

3.2 On the optimal weights

Our matrix characterization of the empirical optimal weights (Lemma 2.1) allows us to give a precise characterization of the statistical properties of linear functionals of these weights. We give here some exact results, concerning distributions and expectations of those functionals. A longer discussion, including robustness and more detailed bias issues can be found in Section 5.

Proposition 3.2. *Assume that the assumptions of Theorem 3.1 hold and in particular X_i are i.i.d $\mathcal{N}(\mu, \Sigma_p)$. Let γ be a fixed n-dimensional vector. Let us call $\widehat{V}_\gamma = (\widehat{V}\ \gamma)$ the $p \times (k+1)$ matrix whose first k columns are those of $\widehat{V}$. Let $\widehat{N}_\gamma = (\widehat{V}_\gamma \Sigma^{-1} \widehat{V}_\gamma)^{-1}$ and W_γ be a $(k+1) \times (k+1)$ matrix with distribution $\mathcal{W}_{k+1}(\widehat{N}_\gamma, n - p + k)$ (conditional on $\widehat{\mu}$). Then,*

$$\gamma' w_{\text{emp}} \big| \, \widehat{\mu} \overset{\mathcal{L}}{=} -\frac{\sum_{i=1}^k u_i W_\gamma(i, k+1)}{W_\gamma(k+1, k+1)} \ .$$

In particular,

$$\mathbf{E}\left(\gamma' w_{\text{emp}} \big| \, \widehat{\mu}\right) = -\frac{\sum_{i=1}^k u_i \widehat{N}_\gamma(i, k+1)}{\widehat{N}_\gamma(k+1, k+1)} \ .$$

We note, somewhat heuristically, that when μ is estimated by $\widehat{\mu}$, since $\widehat{\mu} \sim \mathcal{N}(\mu, \Sigma/n)$, $\widehat{\mu}'\Sigma^{-1}\widehat{\mu} \simeq \mu'\Sigma^{-1}\mu + p/n$, when p, n and $n - p$ are all large (we refer again to Section 5 for a more precise statement). Hence $\widehat{N}_\gamma$ is a not a consistent estimator of $N_\gamma = (V_\gamma'\Sigma^{-1}V_\gamma)^{-1}$. As we will see in Subsection 5.2 and can be expected from the previous proposition, this will also implies bias for linear combinations of empirical optimal weights. We will show in particular that returns are overestimated when using $\widehat{\mu}$ as an estimator for μ.

Another interesting aspect of the previous proposition is that it allows us to understand the fluctuation behavior of $\gamma' w_{\text{emp}}$ when $n - p + k$ is large: as a matter of fact, the limiting fluctuation behavior of the entries of a (fixed-dimensional) Wishart matrix with large number of degrees of freedom is well-known (see e.g Anderson (2003), Theorem 3.4.4 p. 87) and the δ-method can be applied to get the information - conditional on $\widehat{\mu}$.

For instance, if we assume that, conditional on $\widehat{\mu}$, the matrix $\widehat{N}_\gamma$ converges to a matrix N_γ^0, which possibly depends on $\widehat{\mu}$, we see that calling ν the last column $\widehat{N}_\gamma$, ν is asymptotically normal (all statements are conditional on $\widehat{\mu}$), if $N = n - p + k$ goes to infinity when p and n go to infinity. Furthermore we know the limiting covariance of ν, using Theorem 3.4.4 in Anderson (2003). Let us call it Γ_0 and let us call ν_0 the limit of ν - which we assume exists.

If we assume that $\nu_0(k+1)$ is not 0, Slutsky's lemma and the δ-method give us through simple computations that

$$\sqrt{n-p+k}\left(\gamma' w_{\text{emp}} + \frac{\sum_{i=1}^k u_i \nu_0(i)}{\nu_0(k+1)}\right)\Bigg|\, \widehat{\mu} \Longrightarrow \frac{1}{\nu_0(k+1)^2}\mathcal{N}(0, C'\Gamma_0 C)\ ,$$

where $C = \nu_0(k+1)\begin{pmatrix} U \\ 0 \end{pmatrix} - \left(\begin{pmatrix} U \\ 0 \end{pmatrix}' \nu_0\right) e_{k+1}$.

We know the distribution of $\widehat{\mu}$, so we could get (limiting) unconditional results for $\gamma' w_{\text{emp}}$. This is not hard but a bit tedious if we want explicit expressions, and because our focus is mostly on first-order properties in this paper, we do not state the result.

Proof of Proposition 3.2: The proof follows from the representation we gave in Lemma 2.1, i.e

$$\gamma' w_{\text{emp}} = -\frac{1}{(\widehat{V}_\gamma' \widehat{\Sigma}^{-1} \widehat{V}_\gamma)^{-1}(k+1, k+1)} (U'\ 0)(\widehat{V}_\gamma' \widehat{\Sigma}^{-1} \widehat{V}_\gamma)^{-1} \begin{pmatrix} 0_k \\ 1 \end{pmatrix}\ ,$$

and the fact that, by the same arguments as before, conditional on $\widehat{\mu}$,

$$(\widehat{V}_\gamma' \widehat{\Sigma}^{-1} \widehat{V}_\gamma)^{-1}\Big|\, \widehat{\mu} \sim \mathcal{W}_{k+1}((\widehat{V}_\gamma' \Sigma^{-1} \widehat{V}_\gamma)^{-1}, n-p+k)/(n-1)\ .$$

We conclude that

$$\gamma' w_{\text{emp}} \big|\, \widehat{\mu} \overset{\mathcal{L}}{=} -\frac{(U'\ 0) W_\gamma \begin{pmatrix} 0_k \\ 1 \end{pmatrix}}{W_\gamma(k+1, k+1)} = -\frac{\sum_{i=1}^k u_i W_\gamma(i, k+1)}{W_\gamma(k+1, k+1)}\ .$$

This shows the fist part of the proposition.

The second part follows from the following observation. Suppose the matrix P is $\mathcal{W}_p(\text{Id}_p, K)$. If α and β are n-dimensional, orthogonal vectors, let us consider

$$\frac{\alpha' P \beta}{\beta' P \beta}\ .$$

We can of course write $P = \sum_{i=1}^K Y_i Y_i'$, where Y_i are i.i.d $\mathcal{N}(0, \text{Id}_p)$. In other respects, $Y_i'\alpha$ and $Y_i'\beta$ are clearly independent normal random variables, since their covariance is $\alpha'\beta = 0$ and they are normal. So

$$\mathbf{E}\left(\frac{\alpha' P \beta}{\beta' P \beta}\Bigg| \{Y_i'\beta\}_{i=1}^K\right) = 0\ ,$$

because the quantity whose expectation we are taking is a linear combination of mean 0 independent normal random variables. Hence, also,

$$\mathbf{E}\left(\frac{\alpha' P \beta}{\beta' P \beta}\right) = 0\ .$$

Now, when α is not orthogonal to β, we write $\alpha = \beta(\alpha'\beta)/\|\beta\|_2^2 + \delta$, where δ is orthogonal to β. We immediately deduce that in general,

$$\mathbf{E}\left(\frac{\alpha' P \beta}{\beta' P \beta}\right) = \frac{\alpha'\beta}{\|\beta\|_2^2} + \mathbf{E}\left(\frac{\delta' P \beta}{\beta' P \beta}\right) = \frac{\alpha'\beta}{\|\beta\|_2^2}\ .$$

Furthermore, when P is $\mathcal{W}_p(\Sigma, K)$, because we can write $P = \Sigma^{1/2} P_0 \Sigma^{1/2}$, where $P_0 \sim \mathcal{W}_p(\mathrm{Id}_p, K)$, we finally have

$$\mathbf{E}\left(\frac{\alpha' P \beta}{\beta' P \beta}\right) = \frac{\alpha' \Sigma \beta}{\beta' \Sigma \beta} \; .$$

In the case of interest to us, we have $\alpha = \begin{pmatrix} U \\ 0 \end{pmatrix}$, $\beta = e_{k+1}$, and $\Sigma = \widehat{N}_\gamma$. Applying the previous formula gives us the second part of the Proposition. □

We now turn to the question of understanding the robustness properties of the Gaussian results we just obtained. We will do so by studying the same problems under more general distributional assumptions, specifically we will now assume that the observations are elliptically distributed.

4 Solutions of quadratic programs when the data is elliptically distributed

In Section 3, we studied the properties of the "plug-in" solution of Problem (QP-eqc-Pop) under the assumption that the data was normally distributed. While this allowed us to shed light on the statistical properties of the solution of Problem (QP-eqc-Emp), it is naturally extremely important to understand how robust the results are to our normality assumptions.

In this Section, we will consider elliptical models, i.e models such that the data can be expressed as:

$$X_i = \mu + \lambda_i \Sigma^{1/2} Y_i \; ,$$

where λ_i is a random variable and Y_i are i.i.d $\mathcal{N}(0, \mathrm{Id}_p)$ entries. λ_i and Y_i are assumed to be independent, and to lift the indeterminacy between Σ and λ, we assume that $\mathbf{E}\left(\lambda_i^2\right) = 1$. Under this assumption, we clearly have $\mathrm{cov}\left(X_i\right) = \Sigma$. We note that this is not the standard definition of elliptical models, which generally replaces Y_i with a vector uniformly distributed on the sphere in $\mathbb{R}^p$, but it captures the essence of the problem. We refer the interested reader to Anderson (2003) and Fang et al. (1990) for extensive discussions of elliptical distributions.

Our motivation for undertaking this study comes also from the fact that for certain types of data, such as financial data, it is sometimes argued that elliptical models are more reasonable that Gaussian ones, for instance because they can capture non-trivial tail dependence (see Frahm and Jaekel (2005), McNeil et al. (2005)). From a theoretical standpoint, considering elliptical models will also help in several other ways: the results will yield alternative proofs to some of the results we obtained in the Gaussian case, they will allow us to deal with some situations where the data X_i are not independent, and they will also allow us to understand the properties of the bootstrap.

We also want to point out that elliptical distributions allow us to not fall into the geometric "trap" of standard random matrix models highlighted in El Karoui (2009): the fact that data vectors drawn from standard random matrix models are essentially assumed to be almost orthogonal to one another and that their norm (after renormalization by $1/\sqrt{p}$) is almost constant. In a sense, studying elliptical models will allow us to understand what is the impact of the implicit geometric assumptions made about the data when assuming normality. (We purposely do so not under minimal assumptions but under assumptions that capture the essence of the problem while allowing us to show in the proofs the key stochastic phenomena at play.) This part of the article can therefore be viewed as a continuation of the investigation we started in El Karoui (2009) where we showed a lack of robustness of random matrix models (contradicting widespread claims of "universality") by thoroughly investigating limiting spectral distribution properties of high-dimensional covariance matrices when the data is drawn according to elliptical models and generalizations. We show here that the theoretical problems we highlighted in El Karoui (2009) have important practical consequences.

We now turn to the problem of understanding the solution of Problem (QP-eqc-Emp) in the setting where the data is elliptically distributed. We will limit ourselves to the case where the matrix $\widehat{V}$ is full of known and deterministic vectors, except possibly for the sample mean. In this section we restrict ourselves to convergence in probability results. It is clear from Section 2 that to tackle the problems we are considering we need to understand at least three types of quantities: $v'\widehat{\Sigma}^{-1}v$ for a deterministic v with unit norm, $\widehat{\mu}'\widehat{\Sigma}^{-1}v$ and $\widehat{\mu}'\widehat{\Sigma}^{-1}\widehat{\mu}$.

Here is a brief overview of our findings. When we consider elliptical models, our results say that roughly speaking, under certain assumptions given precisely later,

1. $\frac{v'\widehat{\Sigma}^{-1}v}{v'\Sigma^{-1}v} \to \mathfrak{s}$, where $\mathfrak{s}$ satisfies, if G is the law of λ_i^2 and $p/n \to \rho \in (0,1)$, $\int \frac{dG(\tau)}{1+\tau\rho\mathfrak{s}} = 1-\rho$.

2. if $\mu = 0$, $\widehat{\mu}'\widehat{\Sigma}^{-1}\widehat{\mu} \to \rho/(1-\rho)$.

3. if $\mu = 0$, $\widehat{\mu}'\widehat{\Sigma}^{-1}v \to 0$.

All these convergence results are to be understood in probability. They naturally allow us - under certain conditions on the population parameters - to conclude about the convergence in probability of the matrix $\widehat{V}'\widehat{\Sigma}^{-1}\widehat{V}$. The results mentioned above are stated in all details in Theorems 4.1 and 4.4.

In the situation where λ_i are i.i.d, the results above hold when λ_i have a second moment and they do not put too much mass near 0. This is interesting in practice because it tells us that our results hold for heavy-tailed data, which are of particular interest in some financial applications.

The bootstrap situation corresponds basically to G being Poisson(1), which we denote by Po(1). Also in the statement above for $\widehat{\mu}'\widehat{\Sigma}^{-1}\widehat{\mu}$, one should replace $\rho/(1-\rho)$ by $\mathfrak{s}-1$ in the bootstrap case. This is explained in Theorem 4.5 and Subsubsection 4.4.4. Finally, in the case of Gaussian data with "temporal" correlation, i.e when the data can be written in matrix form $X = \mathbf{e}_n\mu' + \Lambda Y\Sigma^{1/2}$, where Λ is not diagonal, one should replace G by the limiting spectral distribution of $\Lambda'\Lambda$. The question of convergence of $\widehat{\mu}'\widehat{\Sigma}^{-1}\widehat{\mu}$ is then more involved. We refer to Proposition 4.2 for details about this situation.

Though we are taking a fundamentally random matrix theoretic approach, our presentation purposely avoids borrowing too many techniques from random matrix theory in the hope of making clear(er) the phenomena that yield the results we will obtain. A more general but considerably more technically complicated (for non-specialists of random matrix theory) approach is being developed in our study of a connected problem and will appear in another paper.

This section is divided into four subsections. The first two are devoted to the main technical issues arising in the study of the problem when the data is elliptically distributed. The third discusses the impact of correlation between observations when the data is Gaussian, as it can be recast as a variant of elliptical problems. The last subsection discusses questions related to the (non-parametric) bootstrap.

4.1 On quadratic forms of the type $v'\widehat{\Sigma}^{-1}v$

The focus of this subsection is on understanding statistics of the type $v'\widehat{\Sigma}^{-1}v$, where v is a deterministic vector. We will prove the following important Theorem.

Theorem 4.1. *Suppose we observe n i.i.d observations X_i, where X_i has the form $X_i = \mu + \lambda_i\Sigma^{1/2}Y_i$, with $Y_i \overset{iid}{\backsim} \mathcal{N}(0, \mathrm{Id}_p)$ and $\{\lambda_i\}_{i=1}^n$ is independent of $\{Y_i\}_{i=1}^n$. $\Sigma^{1/2}$ is deterministic and $\mathbf{E}\left(\lambda_i^2\right) = 1$.*

We call $\rho_n = p/n$ and assume that $\rho_n \to \rho \in (0,1)$.

We use the notation $\tau_i = \lambda_i^2$ and assume that the empirical distribution, G_n, of τ_i converges weakly in probability to a deterministic limit G. We also assume that $\tau_i \neq 0$ for all i.

If $\tau_{(i)}$ is the i-th largest τ_k, we assume that we can find a random variable $N \in \mathbb{N}$ and positive real numbers ϵ_0 and C_0 such that

$$\begin{cases} P(p/N < 1-\epsilon_0) \to 1 as\ n \to \infty\ , \\ P(\tau_{(N)} > C_0) \to 1\ , \\ \exists \eta_0 > 0 \text{ such that } P(N/n > \eta_0) \to 1 \text{ as } n \to \infty\ . \end{cases} \qquad \text{(Assumption-BB)}$$

Under these assumptions, if v is a (sequence of) deterministic vector,

$$\frac{v'\widehat{\Sigma}^{-1}v}{v'\Sigma^{-1}v} \to \mathfrak{s} \text{ in probability }\ ,$$

where $\mathfrak{s}$ satisfies,

$$\int \frac{dG(\tau)}{1+\rho\tau\mathfrak{s}} = 1-\rho \; . \tag{4}$$

A few comments are in order before we turn to the proof. First, the assumption that $\lambda_i \neq 0$ for all i could be dispensed of, as long as all assumptions stated above hold when n is understood to denote the number of non-zero λ_i's. Second, (Assumption-BB) concerning N and C will generally hold as soon as G does not put too much mass at 0, the only problem-specific question remaining being how much mass is put at 0 by G compared to ρ, the limit of p/n.

In particular, in the case where the τ_i's are i.i.d, if there exists $C_0 > 0$ and $x_0 > 0$ such that $P_G(X > C_0) = x_0 > 0$, and if G_n is the empirical distribution of the τ_i's, if $G_n \Longrightarrow G$, we see, using e.g Lemma 2.2 in van der Vaart (1998), that

$$\liminf_{n\to\infty} P_{G_n}(X > C_0) = \frac{\text{Card}\,\{\tau_i > C_0\}}{n} \geq P_G(X > C_0) = x_0 \; .$$

So picking $N = (1-\delta)x_0 n$ will guarantee that we have, if $G_n \Longrightarrow G$ in probability, $P(\tau_{(N)} > C_0) \to 1$ and, of course, $P(N/n > \eta) \to 1$. Hence, in checking whether the theorem applies, we just need to see whether p/N stays bounded away from 1.

In the simpler case when all the $|\lambda_i|$ are bounded away from 0, the conditions on N and C apply directly by taking $N = n$. Finally, let us say that (Assumption-BB) is needed in the proof to guarantee that the smallest eigenvalues of $\widehat{\Sigma}$ stay bounded away from 0 with high-probability.

We now briefly compare the Gaussian and elliptical cases. A simple convexity argument (relying on the fact that $1/(1+x)$ is a convex function of x for $x \geq 0$ and Jensen's inequality) shows that, if μ_G is the mean of G,

$$\mathfrak{s} \geq \frac{1}{1-\rho}\frac{1}{\mu_G} \; .$$

In the case of Gaussian data, $G = \delta_1$, i.e it is a point mass at 1 and we have $\mathfrak{s} = 1/(1-\rho)$. In other respects, for X_i to have covariance Σ, we need $\mathbf{E}\left(\lambda_i^2\right) = 1$. When the λ_i's are i.i.d, with λ_i^2 having distribution G, $\mu_G = \mathbf{E}\left(\lambda_i^2\right) = 1$, and we know that $G_n \Longrightarrow G$ in probability. Therefore, in the class of elliptical distributions considered here, risk underestimation, which is essentially measured by $1/\mathfrak{s}$ (see Theorem 2.1 and Section 5) will be least severe in the Gaussian case. In other words, the Gaussian results lead to over-optimistic conclusions (in terms of proximity between sample and population solutions of the quadratic programs we are considering) within the class of elliptical distributions.

We go back to these questions in more detail in Section 5 and now turn to the proof of Theorem 4.1. The proof could be carried out in at least two ways. We take one that is not standard but we feel best explains the phenomenon that is occurring.

Proof of Theorem 4.1: The proof is easier to carry out when we write the problem in matrix form. Because we focus on $\widehat{\Sigma}$, we can assume without loss of generality (wlog) that $\mu = 0$. Let us consider the $n \times p$ data matrix X whose i-th row is X_i. Similarly, we denote by Y the $n \times p$ data matrix whose i-th row is Y_i. Let us call Λ the diagonal matrix with i-th diagonal entry λ_i and $H = \text{Id}_n - \mathbf{e}\mathbf{e}'/n$, where $\mathbf{e}$ is an n-dimensional vector whose entries are all equal to 1. Note that $H'H = H$. With these notations, we have

$$X = \Lambda Y \Sigma^{1/2} \; .$$

Therefore, $X - \bar{X} = HX$, and

$$\widehat{\Sigma} = \frac{1}{n-1}(X-\bar{X})'(X-\bar{X}) = \frac{1}{n-1}\Sigma^{1/2}Y'\Lambda H \Lambda Y \Sigma^{1/2}$$

Let us call L the matrix $L = \Lambda H \Lambda$. Note that $Y'LY$ is a rank p matrix with probability 1, if we assume that $p \leq n-1$ (recall that all the entries of Λ are non-zero). Hence, $Y'LY$ is invertible with probability 1. Therefore,

$$\widehat{\Sigma}^{-1} = \Sigma^{-1/2}\left(\frac{1}{n-1}Y'LY\right)^{-1}\Sigma^{-1/2} \; .$$

Finally, we have

$$\frac{\upsilon'\widehat{\Sigma}^{-1}\upsilon}{\upsilon'\Sigma^{-1}\upsilon} = \nu'\left(\frac{1}{n-1}Y'LY\right)^{-1}\nu \; ,$$

where $\nu = \Sigma^{-1/2}\upsilon/\|\Sigma^{-1/2}\upsilon\|_2$ is a vector of ℓ_2 norm 1.

We now make all of our statements conditional on Λ. Because of the independence of Y and Λ, we can therefore treat the λ_i's as if they were constant and the $Y_{i,j}$'s as i.i.d $\mathcal{N}(0,1)$ random variables. Λ is now assumed to be in the set of matrices $\mathcal{L}_{\epsilon,\delta}$, defined as the end of this proof, for which we have control of the smallest eigenvalue of $\mathcal{S} = Y'LY/(n-1)$. In the steps that follow that are conditional on Λ, we therefore consider that we control the smallest eigenvalue of $\mathcal{S}$ and we will show formally that it is the case at the last step, when we get results unconditionally on Λ. (The arguments are not circular.) We note that if Λ is in $\mathcal{L}_{\epsilon,\delta}$, N is lower bounded. Because N is a function of the λ_i's and hence of Λ, we write all the results conditionally on Λ, but the reader should keep in mind that this conditioning constrains also the possible values of N.

• **Getting results conditionally on $\boldsymbol{\Lambda}$**

If O is an orthogonal matrix, $O'Y'LYO \overset{\mathcal{L}}{=} Y'LY$, because Y is full of i.i.d $\mathcal{N}(0,1)$ random variables and is therefore invariant (in law) by left and right rotation. Therefore the eigenvalues and eigenvectors of $Y'LY$ are independent and its matrix of eigenvectors is uniformly (i.e Haar) distributed on the orthogonal group (see also Chikuse (2003), p. 40, Equation (2.4.4)). Let us write a spectral decomposition of $Y'LY$:

$$\mathcal{S} = \frac{1}{n-1}Y'LY = \sum_{i=1}^{p}\gamma_i\upsilon_i\upsilon_i' \; .$$

We know that a.s $\gamma_i \neq 0$ for all i, so

$$\nu'\mathcal{S}^{-1}\nu = \sum_{i=1}^{p}\frac{1}{\gamma_i}(\nu'\upsilon_i)^2 \; .$$

We claim that

$$\left.\nu'\mathcal{S}^{-1}\nu - \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\gamma_i}\right| (\{\gamma_i\}_{i=1}^{p},\Lambda) \to 0$$

To see this, note that $\mathbf{E}\left((\nu'\upsilon_i)^2\right) = \|\nu\|_2^2/p = 1/p$, because υ_i is uniformly distributed on the unit sphere when Υ (the matrix containing the υ_i) is Haar distributed on the orthogonal group. Hence, given the independence between γ_i and υ_i,

$$\mathbf{E}\left(\nu'\mathcal{S}^{-1}\nu\middle|\, \{\gamma_i\}_{i=1}^{n}, \Lambda\right) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\gamma_i} \; .$$

Now let us call w the vector with $w_i = (\nu'\upsilon_i)^2$, and g the vector with i-th entry $g_i = 1/\gamma_i$. Clearly, since $\nu'\mathcal{S}^{-1}\nu = g'w$, var $\left(\nu'\mathcal{S}^{-1}\nu\,|\{\gamma_i\}\right) = g'\text{cov}\,(w)\,g$. By symmetry it is clear that $\text{cov}\,(w)\,(i,i) = \text{cov}\,(w)\,(1,1)$ and $\text{cov}\,(w)\,(i,j) = \text{cov}\,(w)\,(1,2)$ if $i \neq j$. Further, since the matrix Υ containing the vectors υ_i is Haar distributed on the orthogonal group, we can assume without loss of generality that $\nu = e_1$ for all the computations at stake. As a matter of fact, if O_1 is an orthogonal matrix such that $O_1\nu = e_1$, then $\nu'\upsilon_i = e_1'O_1\upsilon_i = e_1'\tilde{\upsilon}_i$ where the matrix $\widetilde{\Upsilon} = O_1\Upsilon$ is again Haar distributed on the orthogonal group.

So from now on, we assume (without loss of generality) that $\nu = e_1$, and we therefore simply need to understand the correlation between $(\upsilon_1(1))^2$ and $(\upsilon_2(1))^2$. Now, the first row of an orthogonal matrix uniformly distributed on the orthogonal group is a unit vector uniformly distributed on the unit sphere, because if O is Haar distributed, so is O'. We now recall the fact that a vector uniformly distributed on the unit sphere, υ can be generated by drawing at random a $\mathcal{N}(0,\text{Id}_p)$ random vector and normalizing it. In other words, if $Z \sim \mathcal{N}(0,\text{Id}_p)$, $\upsilon = Z/\|Z\|_2$.

So our task has now be considerably simplified, and it consists in understanding the covariance between 2 random variables, r_1 and r_2 such that, if Z_i are i.i.d $\mathcal{N}(0,1)$,

$$r_i = \frac{Z_i^2}{\sum_{i=1}^{p}Z_i^2} \; .$$

Now, by symmetry, $\mathbf{E}\left(r_1 r_2\right) = \mathbf{E}\left(r_i r_j\right)$ for all $i \neq j$ and $p(p-1)\mathbf{E}\left(r_1 r_2\right) = \sum_{i \neq j} \mathbf{E}\left(r_i r_j\right)$. In other words,

$$p(p-1)\mathbf{E}\left(r_1 r_2\right) = \mathbf{E}\left(\sum_{i \neq j} \frac{Z_i^2 Z_j^2}{(\sum_{i=1}^p Z_i^2)^2}\right) = \mathbf{E}\left(\frac{\sum_{i,j} Z_i^2 Z_j^2}{(\sum_{i=1}^p Z_i^2)^2} - \sum_{i=1}^p \frac{Z_i^4}{(\sum_{i=1}^p Z_i^2)^2}\right) .$$

We can therefore conclude that

$$p(p-1)\mathbf{E}\left(r_1 r_2\right) = 1 - p\mathbf{E}\left(\frac{Z_1^4}{(\sum_{i=1}^p Z_i^2)^2}\right) .$$

Hence, $\mathbf{E}\left(r_1 r_2\right) \leq 1/(p(p-1))$. On the other hand,

$$\mathbf{E}\left(\frac{Z_1^4}{(\sum_{i=1}^p Z_i^2)^2}\right) \leq \mathbf{E}\left(\frac{Z_1^4}{(\sum_{i=2}^p Z_i^2)^2}\right) = \frac{3}{(p-3)(p-5)} ,$$

since $\sum_{i=2}^p Z_i^2 \sim \chi_{p-1}^2$, and $\mathbf{E}\left((\chi_{p-1}^2)^r\right) = 2^r\Gamma((p-1)/2+r)/\Gamma((p-1)/2)$, for $r > -(p-1)/2$ (see e.g Mardia et al. (1979), p. 487). Applying these results with $r = -2$ yields the above result as soon as $p > 5$, by using the fact that $\Gamma(x+1) = x\Gamma(x)$. We therefore have

$$1 - \frac{3p}{(p-3)(p-5)} \leq p(p-1)\mathbf{E}\left(r_1 r_2\right) \leq 1 .$$

Since, for instance by symmetry, $\mathbf{E}\left(r_1\right) = 1/p$, and $1/(p(p-1)) - 1/p^2 = (p^2(p-1))^{-1}$, we conclude that

$$\frac{1}{p^2(p-1)} - \frac{3p}{p(p-1)(p-3)(p-5)} \leq \operatorname{cov}\left(r_1, r_2\right) \leq \frac{1}{p^2(p-1)} .$$

We have therefore established the fact that

$$|\operatorname{cov}\left(r_1, r_2\right)| = \mathrm{O}(p^{-3}) .$$

On the other hand, since $\mathbf{E}\left(r_1^2\right) = \mathbf{E}\left(Z_1^4(\sum_{i=1}^p Z_i^2)^{-2}\right)$, we have

$$0 \leq \operatorname{var}\left(r_i\right) \leq \frac{3}{(p-3)(p-5)} - \frac{1}{p^2} .$$

Now using the (standard) fact that, for symmetric matrices M, if $\sigma_1(M)$ is the largest singular value of M,

$$\sigma_1(M) \leq \max_i \sum_j |m_{i,j}| ,$$

(it can easily be proved using for instance, Theorems 5.6.6 and 5.6.9 in Horn and Johnson (1994), or Geršgorin's Theorem (Theorem 6.1.1 in the same reference)) we have

$$\sigma_1(\operatorname{cov}\left(r\right)) \leq \left(\frac{3}{(p-3)(p-5)} - \frac{1}{p^2}\right) + \mathrm{O}(p^{-2}) = \mathrm{O}(p^{-2}) .$$

The first term in the previous bound comes from the contribution of the diagonal and the second term is the sum over the $p-1$ off-diagonal elements on a given row of the upper-bound we had on each such element, i.e Cp^{-3} for some C.

Let us now return to our initial question which was to show that the conditional variance of interest to us was going to zero. Recall that g is a vector whose i-th entry is $1/\gamma_i$. Since

$$\operatorname{var}\left(\nu' \mathcal{S}^{-1} \nu \,|\{\gamma_i\}, \Lambda\right) = g' \operatorname{cov}\left(w\right) g ,$$

and $\operatorname{cov}\left(w\right) = \operatorname{cov}\left(r\right)$, we have, for C a constant, and if $|||A|||_2$ denotes the operator norm (or largest singular value) of the matrix A,

$$\operatorname{var}\left(\nu' \mathcal{S}^{-1} \nu \,|\{\gamma_i\}, \Lambda\right) \leq |||\operatorname{cov}\left(r\right)|||_2 \|g\|_2^2 \leq C\frac{\|g\|_2^2}{p^2} = C\frac{1}{p^2}\sum_{i=1}^p \frac{1}{\gamma_i^2} .$$

Now given the assumptions we made on Λ, according to the arguments given at the end of this proof and Lemma B-1, $\gamma_i^2 \geq \mathfrak{C}_n(1-\sqrt{p/(N-1)})^2/2$, where $\mathfrak{C}_n = C_0(N-1)/(n-1)$, with high $(\{Y_i\}_{i=1}^n)$-probability , so we conclude that all the γ_i's are bounded away (uniformly for Λ in $\mathcal{L}_{\epsilon,\delta}$) from 0, and

$$\text{var}\left(\nu'\mathcal{S}^{-1}\nu\,|\{\gamma_i\},\Lambda\right) \to 0\;.$$

Therefore,

$$\nu'\mathcal{S}^{-1}\nu - \frac{1}{p}\sum_{i=1}^p \frac{1}{\gamma_i}\Bigg|\,\{\gamma_i\}_{i=1}^p, \Lambda \to 0 \quad \text{in probability}\;.$$

Let us now show that this implies convergence in probability to 0 (conditional on Λ only) of $Q_n = \nu'\mathcal{S}^{-1}\nu - \frac{1}{p}\sum_{i=1}^p \frac{1}{\gamma_i}$. Let us call $h_n = C\frac{\|g\|_2^2}{p^2}$. For ζ_n to be determined later, we have

$$P(|Q_n| > \epsilon|\Lambda) \leq P(|Q_n| > \epsilon \,\&\, h_n \leq \zeta_n|\Lambda) + P(h_n > \zeta_n|\Lambda)\;.$$

On the other hand,

$$P(|Q_n| > \epsilon \,\&\, h_n \leq \zeta_n|\Lambda) = \mathbf{E}\left(\mathbf{E}\left(1_{|Q_n|>\epsilon}1_{h_n\leq\zeta_n}\middle|\,\{g_i\},\Lambda\right)|\Lambda\right)\;.$$

Because h_n is a function of the g_i's and $\text{var}\,(Q_n|\{\gamma_i\}_{i=1}^p,\Lambda) \leq h_n$,

$$\mathbf{E}\left(1_{|Q_n|>\epsilon}1_{h_n\leq\zeta_n}\middle|\,\{g_i\},\Lambda\right) = 1_{h_n\leq\zeta_n}\mathbf{E}\left(1_{|Q_n|>\epsilon}\middle|\,\{g_i\},\Lambda\right) \leq 1_{h_n\leq\zeta_n}\frac{h_n}{\epsilon^2} \leq \frac{\zeta_n}{\epsilon^2}\;.$$

But under our assumptions, we have $h_n|\Lambda = \mathrm{O}_P(1/p)$, so taking $\zeta_n = n^{-1/2}$, we have $P(h_n > \zeta_n|\Lambda) \to 0$ and of course, $\zeta_n/\epsilon^2 \to 0$. Hence, for any $\epsilon > 0$,

$$P(|Q_n| > \epsilon|\Lambda) \to 0\;.$$

Let us now turn to the question of identifying the limit.

- **About $\frac{1}{p}\sum_{i=1}^p \frac{1}{\gamma_i}$** The Stieltjes transform of the spectral distribution of $Y'LY/(n-1)$ is

$$s_p(z) = \frac{1}{p}\sum_{i=1}^p \frac{1}{\gamma_i - z}\;.$$

The quantity $\frac{1}{p}\sum_{i=1}^p \frac{1}{\gamma_i}$ is therefore $s_p(0)$ and we are interested in its limit, if it exists, which would correspond to $\mathfrak{s}$.

Recall the Marčenko-Pastur equation, from Marčenko and Pastur (1967), Wachter (1978) and Silverstein (1995): if Y is $n\times p$ has i.i.d entries with mean 0 and variance 1 and L is positive semidefinite, has limiting spectral distribution G and is independent of Y, if $p/n \to \rho > 0$, and if m_p is the Stieltjes transform of the spectral distribution of $Y'LY/p$, then $m_p(z)$ tends (in probability) to $m(z)$ for all z in $\mathbb{C}^+$ and m satisfies

$$-\frac{1}{m(z)} = z - \frac{1}{\rho}\int \frac{\tau dG(\tau)}{1+\tau m(z)}\;. \tag{5}$$

Note that, if $p/n = \rho_n$, we have

$$\rho_n s_p(\rho_n z) = m_p(z)\;.$$

Therefore, according to Marčenko and Pastur (1967), Wachter (1978) and Silverstein (1995), we know that $s_p(z)$ converges for $z \in \mathbb{C}^+$ to a non-random quantity $s(z)$, in probability. Note that s satisfies, in light of Equation (5)

$$-\frac{1}{s(z)} = z - \int \frac{\tau dG(\tau)}{1+\tau\rho s(z)}\;.$$

Here, because we know using our assumptions (see the end of the proof) that γ_i are bounded away from 0 with probability going to 1, we can also conclude that $s_p(0) \to s(0)$ with probability going to 1, because of the weak convergence (in probability) of spectral distributions that pointwise convergence of

Stieltjes transforms implies (as a test function, we can use a function that coincides with $1/x$ except in a interval near 0 where we are guaranteed that there are no eigenvalues asymptotically). We also know that s is continuous (and actually analytic) at 0 in this situation since the s is the Stieltjes transform of a measure who has support bounded away from 0. So the previous equation holds for $z = 0$ and we have

$$-\frac{1}{s(0)} = -\int \frac{\tau dG(\tau)}{1+\tau\rho s(0)} \; .$$

Multiplying both sides by $-\rho s(0)$, we get, after we recall that G is a probability measure,

$$\rho = \int \frac{\rho\tau s(0) dG(\tau)}{1+\tau\rho s(0)} = \int \left(1 - \frac{1}{1+\tau\rho s(0)}\right) dG(\tau) = 1 - \int \frac{1}{1+\tau\rho s(0)} dG(\tau) \; .$$

Calling $s(0) = \mathfrak{s}$, we have the result we announced - conditionally on Λ. Now, here G is the limiting spectral distribution of $\Lambda H \Lambda$, but because this matrix is a rank one perturbation of Λ^2, these two matrices have the same limiting spectral distribution. This concludes this part of the proof.

• **Getting results unconditionally on $\boldsymbol{\Lambda}$** All the statements above were made conditional on Λ. If we can show that our probability bounds and our characterization of the limit hold uniformly in Λ, we will have an unconditional statement, as we seek.

The fact that the limit does not depend on Λ is essentially obvious from its description: all that matters is the limiting spectral distribution, which is the same for all Λ. Let us consider the question of uniform probability bounds. All we need to do is show that we control $P(h_n > \zeta_n|\Lambda)$ uniformly in Λ. At this point, it is helpful to recall that N can be viewed as a function of Λ. Recall also the results and the proof of Lemma B-1: in particular, when Λ is such that $p/N < 1-\epsilon$, if $\mathfrak{C}_n = C_0 \frac{N-1}{n-1}$ and γ_p is the smallest eigenvalue of $Y'LY/n - 1$, we have, if P_Λ denotes probability conditional on Λ,

$$P_\Lambda\left(\sqrt{\gamma_p} \leq \sqrt{\mathfrak{C}_n}\left[(1-\sqrt{1-\epsilon}) - t\right]\right) \leq \exp\left(-(N-1)t^2\right) \; .$$

Let us call $\mathcal{L}_{\epsilon,\delta}$ the set of matrices Λ such that $p/N < 1-\epsilon$ and $C_0(N-1)/(n-1) > \delta$. Under (Assumption-BB), for a δ bounded away from 0 (e.g $\delta = C_0\eta_0/2$, since we need a bound on $\liminf C_0 N/n$ that holds with probability going to 1), $P(\Lambda \in \mathcal{L}_{\epsilon,\delta}) \to 1$. In other respects, if $\Lambda \in \mathcal{L}_{\epsilon,\delta}$,

$$P_\Lambda\left(\sqrt{\gamma_p} \leq \sqrt{\delta}\left[(1-\sqrt{1-\epsilon}) - t\right]\right) \leq \exp\left(-(n-1)\delta t^2/C\right) \; .$$

Hence, when $\Lambda \in \mathcal{L}_{\epsilon,\delta}$, if $\zeta_n = n^{-1/2}$, $P(h_n > \zeta_n|\Lambda) \leq f_n(C,\epsilon,\delta)$, where $f_n(C,\epsilon,\delta)$ tends to 0 as n tends to infinity. In other words, we have now established that if $\Lambda \in \mathcal{L}_{\epsilon,\delta}$, and $Q_n = \nu'\mathcal{S}^{-1}\nu - \frac{1}{p}\sum_{i=1}^p \frac{1}{\gamma_i}$, for any $t > 0$,

$$P(|Q_n| > t|\Lambda) \leq \frac{\zeta_n}{t^2} + f_n(C,\epsilon,\delta) \; .$$

Using the fact that $P(|Q_n| > t) \leq P\left(|Q_n| > t\&\{\Lambda \in \mathcal{L}_{\epsilon,\delta}\}\right) + P\left(\Lambda \notin \mathcal{L}_{\epsilon,\delta}\right)$, we conclude that $P(|Q_n| > t) \to 0$ as n tends to infinity for any $t > 0$ and the proof is complete. □

As a consequence of Theorem 4.1, we have the following practically useful result.

Lemma 4.1. *We assume that the assumptions of Theorem 4.1 hold and that G is such that $\mathfrak{s}$ is not ∞.*

Suppose that v_1 and v_2 are deterministic vectors such that

$$\frac{v_1'\Sigma^{-1}v_2}{(v_1+v_2)'\Sigma^{-1}(v_1+v_2)} \text{ and } \frac{v_1'\Sigma^{-1}v_2}{(v_1-v_2)'\Sigma^{-1}(v_1-v_2)}$$

are bounded away from 0. Then under the assumptions of Theorem 4.1,

$$\frac{v_1'\widehat{\Sigma}^{-1}v_2}{v_1'\Sigma^{-1}v_2} \to \mathfrak{s} \text{ in probability.}$$

In other respects, suppose that $v_1'\Sigma^{-1}v_2 \to 0$, while $v_1'\Sigma^{-1}v_1$ and $v_2'\Sigma^{-1}v_2$ stay bounded away from ∞. Then, under the assumptions of Theorem 4.1,

$$v_1'\widehat{\Sigma}^{-1}v_2 \to 0 \text{ in probability.}$$

Proof. The proof of the first part of the Lemma is an immediate consequence of Theorem 4.1, after writing

$$\begin{aligned} 2\frac{v_1'\widehat{\Sigma}^{-1}v_2}{v_1'\Sigma^{-1}v_2} &= \frac{(v_1+v_2)'\widehat{\Sigma}^{-1}(v_1+v_2)}{(v_1+v_2)'\Sigma^{-1}(v_1+v_2)}\frac{(v_1+v_2)'\Sigma^{-1}(v_1+v_2)}{v_1'\Sigma^{-1}v_2} \\ &- \frac{(v_1-v_2)'\widehat{\Sigma}^{-1}(v_1-v_2)}{(v_1-v_2)'\Sigma^{-1}(v_1-v_2)}\frac{(v_1-v_2)'\Sigma^{-1}(v_1-v_2)}{v_1'\Sigma^{-1}v_2} \end{aligned}$$

For the proof of the second part, we note that Theorem 4.1 implies that

$$v'\widehat{\Sigma}^{-1}v = \mathfrak{s}v'\Sigma^{-1}v + \mathrm{o}_P(v'\Sigma{-}1v)\ .$$

Note that since for $i = 1, 2$, $v_i'\Sigma^{-1}v_i$ is assumed to stay bounded, the same is true of $(v_1+\epsilon v_2)'\Sigma^{-1}(v_1+\epsilon v_2)$, where $\epsilon = \pm 1$. Now we write

$$2\,v_1'\widehat{\Sigma}^{-1}v_2 = (v_1+v_2)'\widehat{\Sigma}^{-1}(v_1+v_2) - (v_1-v_2)'\widehat{\Sigma}^{-1}(v_1-v_2)\ .$$

Our previous remark and the assumption of boundedness of $v_i'\Sigma^{-1}v_i$ implies that

$$\begin{aligned} 2\,v_1'\widehat{\Sigma}^{-1}v_2 &= \mathfrak{s}((v_1+v_2)'\Sigma^{-1}(v_1+v_2) - (v_1-v_2)'\Sigma^{-1}(v_1-v_2)) + \mathrm{o}_P(1)\ , \\ &= \mathfrak{s}\,2v_1'\Sigma^{-1}v_2 + + \mathrm{o}_P(1) = \mathrm{o}_P(1)\ . \end{aligned}$$

□

4.2 On quadratic forms involving $\widehat{\mu}$ and $\widehat{\Sigma}^{-1}$

As is clear from the solutions of Problems (QP-eqc) and (QP-eqc-Emp), when $\widehat{\mu}$ appears in the matrix $\widehat{V}$, its influence on the solution of our quadratic program will manifest itself in the form of quantities of the type $\widehat{\mu}'\widehat{\Sigma}^{-1}\widehat{\mu}$ and $v_i'\widehat{\Sigma}^{-1}\widehat{\mu}$. It is therefore important that we get a good understanding of those quantities.

Compared to the Gaussian case, in the elliptical case, $\widehat{\mu}$ is not independent of $\widehat{\Sigma}$ anymore, which generates some complications. They are fully addressed in Theorem 4.4, but as a stepping stone to that result (the main of this subsection), we need the following theorem, which essentially takes care of the problem of understanding $\widehat{\mu}'\widehat{\Sigma}^{-1}\widehat{\mu}$ for the class of elliptical distributions we consider when the population mean is 0.

Theorem 4.2. *Suppose Y is an $n \times p$ matrix whose rows are the vectors Y_i, which are i.i.d $\mathcal{N}(0, \mathrm{Id}_p)$.*

Suppose Λ is a diagonal matrix whose i-th entry is λ_i, which is possibly random and is independent of Y. Call $\tau_i = \lambda_i^2$. We assume that $\tau_i \neq 0$ for all i and

$$\frac{1}{n^2}\sum_{i=1}^{n}\lambda_i^4 = \frac{1}{n^2}\sum_{i=1}^{n}\tau_i^2 \to 0 \textit{ in probability}\ . \qquad \text{(Assumption-BLa)}$$

If $\tau_{(i)}$ is the i-th largest τ_k, we assume that we can find a random variable $N \in \mathbb{N}$ and positive real numbers ϵ_0 and C_0 such that

$$\begin{cases} P(p/N < 1-\epsilon_0) \to 1 \textit{as } n \to \infty\ , \\ P(\tau_{(N)} > C_0) \to 1\ , \\ \exists \eta_0 > 0 \textit{ such that } P(N/n > \eta_0) \to 1 \textit{ as } n \to \infty\ . \end{cases} \qquad \text{(Assumption-BB)}$$

Let us call $\rho_n = p/n$ and $\rho = \lim_{n\to\infty}\rho_n$. We assume that $\rho \in (0, 1)$. We call

$$Z_{n,p} = \frac{1}{n^2}\mathbf{e}'\Lambda Y(Y'\Lambda^2 Y/n)^{-1}Y'\Lambda\mathbf{e}\ .$$

Then we have

$$Z_{n,p} \to \rho\ , \textit{ in probability.}$$

If the $n \times p$ data matrix $\widetilde{X}$ is written $\widetilde{X} = \Lambda Y\Sigma^{1/2}$, and if $\widehat{m} = \Sigma^{1/2}Y'\Lambda\mathbf{e}/n$ is the vector of column means of $\widetilde{X}$, and if $\widehat{\Sigma}$ is the sample covariance matrix computed from $\widetilde{X}$, we have

$$\widehat{m}'\widehat{\Sigma}^{-1}\widehat{m} \to \kappa = \frac{\rho}{1-\rho} \textit{ in probability}\ .$$

Some comments on this theorem are in order. First, $Z_{n,p}$ is unchanged if we rescale all the λ_i's by the same constant. So it appears we could assume that they are all less than 1 for instance and dispense entirely with (Assumption-BLa). However, that would potentially violate the conditions of (Assumption-BB) which appear to guarantee that $Z_{n,p}$ has variance going to zero. We also note that because the Y_i's have a continuous distribution and we know that all the λ_i's are different from 0, the existence of $Z_{n,p}$ is guaranteed with probability 1.

Some practical clarifications are also in order concerning the condition

$$\frac{1}{n^2}\sum_{i=1}^{n}\lambda_i^4 = \frac{1}{n^2}\sum_{i=1}^{n}\tau_i^2 \to 0 \text{ in probability} .$$

When the λ_i's are i.i.d, this condition is satisfied (almost surely and hence in probability) if for instance the λ_i's have finite second moment according to the Marcinkiewicz-Zygmund law of large numbers (see Chow and Teicher (1997), p. 125). This is very interesting from a practical standpoint as it basically means that we only require our random variables X_i to have a second moment for the theorem to hold. We note that if there were no variance, the premises of the problem would be essentially flawed (after all the quadratic form we are optimizing involves a proxy for the population covariance and in the absence of a second moment for the λ_i's, the population covariance would not exist), and hence we require minimal conditions from the point of view of the practical problem at stake.

Finally, and remarkably, the limit of $Z_{n,p}$ does not depend on the empirical distribution of the λ_i's. In particular, in the class of elliptical distributions (satisfying the assumptions of Theorem 4.2), the limit of $\widehat{m}'\widehat{\Sigma}^{-1}\widehat{m}$ is always the same: $\kappa = \rho/(1-\rho)$.

We now turn to proving Theorem 4.2. The proof will be facilitated by the following lemma, which essentially gives us $\mathbf{E}\left(Z_{n,p}\right)$.

Lemma 4.2. *Let Y be an $n \times p$ random matrix, with $n \geq p$ with for instance independent rows, Y_i. Assume that Y_i have symmetric distributions, i.e $Y_i \overset{\mathcal{L}}{=} -Y_i$. Let Λ be an $n \times n$ diagonal matrix with possibly random entries. Let $P = \Lambda Y(Y'\Lambda^2 Y)^{-1}Y'\Lambda$ be a random projection matrix. Y is assumed to be independent of Λ and Y and Λ are assumed to be such that P exists with probability 1. Then,*

$$\mathbf{E}\left(\mathbf{e}'P\mathbf{e}|\Lambda\right) = \mathbf{E}\left(\mathbf{e}'P\mathbf{e}\right) = p .$$

In particular, the result applies when Y_i are normally distributed and Λ is such that (Assumption-BB) holds and P is defined with probability one.

Proof of the lemma: Let us note that $P = f_\Lambda(Y_1, \ldots, Y_n)$. Now, conditional on Λ, $P \overset{\mathcal{L}}{=} f_\Lambda(-Y_1, Y_2, \ldots, Y_n) = \tilde{P}$. However $\tilde{P}(1,j) = -P(1,j)$, if $j \neq 1$. As a matter of fact,

$$P(1,j) = \lambda_1\lambda_j Y_1'(\sum_{i=1}^{n}\lambda_i^2 Y_i Y_i')^{-1}Y_j .$$

Hence, conditional on Λ, $P(1,j) \overset{\mathcal{L}}{=} -P(1,j)$. Now P is an orthogonal projection matrix, $P = P'$, so all its entries are less than 1, the operator norm of P. In particular, all the entries have an expectation. Since, if $j \neq 1$, $P(1,j)$ has a symmetric distribution (conditional on Λ), we conclude that

$$\mathbf{E}\left(P(1,j)|\Lambda\right) = 0 , \text{ if } j \neq 1 .$$

Note that the same arguments would apply if 1 were replaced by i, so we really have

$$\mathbf{E}\left(P(i,j)|\Lambda\right) = 0 , \text{ if } j \neq i .$$

Therefore,

$$\mathbf{E}\left(\mathbf{e}'P\mathbf{e}|\Lambda\right) = \mathbf{E}\left(\text{trace}\left(P\right)|\Lambda\right) = p ,$$

since P has rank p and is a projection matrix.

The same results hold when we take expectations over Λ by similar arguments. □

To prove Theorem 4.2, all we have to do (in light of Lemma 4.2) is to show that we control the variance of

$$Z_{n,p} = \frac{1}{n}\mathbf{e}'P\mathbf{e} \; .$$

We are going to do this now by using rank 1 perturbation arguments, in connection with the Efron-Stein inequality.

Proof of Theorem 4.2: As before, we first work conditionally on Λ. We assume until further notice that $\Lambda \in \mathcal{L}_{\epsilon_0,\delta_0}$, a set of matrices who is defined at the end of the proof, will have measure going to 1 asymptotically, and is such that all the technical issues appearing in the proof can be taken care of. (The arguments are not circular.)

We will use the notation

$$\mathcal{S} = \frac{1}{n}\sum_{k=1}^{n} \lambda_k^2 Y_k Y_k' \; , \text{ and } \mathcal{S}_i = \mathcal{S} - \frac{1}{n}\lambda_i^2 Y_i Y_i' \; .$$

Note that $\mathcal{S}_i$ is symmetric and positive semi-definite. Naturally, in matrix form we can write $\mathcal{S} = (Y'\Lambda^2 Y)/n$ and $\mathcal{S}_i = (Y'\Lambda_i^2 Y)/n$, where Λ_i^2 is the same matrix as Λ, except that $\Lambda_i(i,i) = 0$. Our aim is to approximate

$$Z_{n,p} = \frac{\mathbf{e}'\Lambda Y}{n}\left(\frac{Y'\Lambda^2 Y}{n}\right)^{-1}\frac{Y'\Lambda\mathbf{e}}{n} = f(X_1,\ldots,X_n) \; ,$$

by a random variable involving only $(Y_1,\ldots,Y_{i-1},Y_{i+1},\ldots,Y_n)$, i.e not involving Y_i. Using classic matrix perturbation results (see Horn and Johnson (1990), p. 19), we have

$$\mathcal{S}^{-1} = \left(\mathcal{S}_i + \frac{\lambda_i^2}{n}Y_i Y_i'\right)^{-1} = \mathcal{S}_i^{-1} - \frac{\lambda_i^2}{n}\frac{\mathcal{S}_i^{-1}Y_i Y_i'\mathcal{S}_i^{-1}}{1+\lambda_i^2(Y_i'\mathcal{S}_i^{-1}Y_i/n)} \; .$$

Of course, if e_i is the i-th canonical basis vector in $\mathbb{R}^n$,

$$W \triangleq \Lambda Y = \sum_{i=1}^{n}\lambda_i e_i Y_i' \triangleq W_i + \lambda_i e_i Y_i' \; .$$

Let us now call $q_i = Y_i'\mathcal{S}_i^{-1}Y_i/n$ and $r_i = W_i\mathcal{S}_i^{-1}Y_i$. We have

$$\Lambda Y\mathcal{S}^{-1} = W_i\mathcal{S}_i^{-1} - \frac{\lambda_i^2}{n}\frac{r_i Y_i'\mathcal{S}_i^{-1}}{1+\lambda_i^2 q_i} + \lambda_i e_i Y_i'\mathcal{S}_i^{-1} - \lambda_i^3 q_i\frac{e_i Y_i'\mathcal{S}_i^{-1}}{1+\lambda_i^2 q_i} \; . \tag{6}$$

Similarly,

$$\begin{aligned}\Lambda Y\mathcal{S}^{-1}Y'\Lambda = W_i\mathcal{S}_i^{-1}W_i' &- \frac{\lambda_i^2}{n}\frac{r_i r_i'}{1+\lambda_i^2 q_i} + \lambda_i e_i r_i' - \lambda_i^3 q_i\frac{e_i r_i'}{1+\lambda_i^2 q_i}\\ &+ \lambda_i r_i e_i' - \lambda_i^3 q_i\frac{r_i e_i'}{1+\lambda_i^2 q_i} + \lambda_i^2 n q_i e_i e_i' - \lambda_i^4 n q_i^2\frac{e_i e_i'}{1+\lambda_i^2 q_i}\end{aligned} \tag{7}$$

This is, in some sense, the key expansion in this proof. Now let us call $\widehat{\mu}_i' = \mathbf{e}'W_i/n$ and $w_i = \mathbf{e}'r_i/n = \widehat{\mu}_i'\mathcal{S}_i^{-1}Y_i$. We have

$$Z_{n,p} = \widehat{\mu}_i'\mathcal{S}_i^{-1}\widehat{\mu}_i - \frac{\lambda_i^2}{n}\frac{w_i^2}{1+\lambda_i^2 q_i} + 2\frac{\lambda_i}{n}w_i - \frac{2}{n}\lambda_i^3\frac{q_i w_i}{1+\lambda_i^2 q_i} + \frac{\lambda_i^2}{n}q_i - \frac{\lambda_i^4}{n}\frac{q_i^2}{1+\lambda_i^2 q_i}$$

Now let us call $Z_i = \widehat{\mu}_i'\mathcal{S}_i^{-1}\widehat{\mu}_i$. Clearly, Z_i does not depend on Y_i. Now, it is easily verified that

$$\left(2\lambda_i w_i + \lambda_i^2 q_i - \frac{\lambda_i^4 q_i^2}{1+\lambda_i^2 q_i} - \frac{\lambda_i^2 w_i^2}{1+\lambda_i^2 q_i} - 2\frac{\lambda_i^3 q_i w_i}{1+\lambda_i^2 q_i}\right) = 1 - \frac{(1-\lambda_i w_i)^2}{1+\lambda_i^2 q_i} \; .$$

We finally conclude that

$$Z_{n,p} = Z_i + \frac{1}{n}\left(1 - \frac{(1-\lambda_i w_i)^2}{1+\lambda_i^2 q_i}\right) \ . \tag{8}$$

We now recall the Efron-Stein inequality, as formulated in Theorem 9 of Lugosi (2006): if $\alpha = f(X_1, \ldots, X_n)$, where the X_i's are independent, and α_i is a measurable function of $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$, then

$$\operatorname{var}(\alpha) \leq \sum_{i=1}^{n} \mathbf{E}\left((\alpha - \alpha_i)^2\right) \ .$$

In particular, for us, it means that

$$\operatorname{var}(Z_{n,p}|\Lambda) \leq \sum_{i=1}^{n} \mathbf{E}\left((Z_{n,p} - Z_i - \frac{1}{n})^2|\Lambda\right) \ .$$

If we now use Equation (8) and the fact that $q_i \geq 0$, we have

$$n\left|Z_{n,p} - Z_i - \frac{1}{n}\right| = \frac{(1-\lambda_i w_i)^2}{1+\lambda_i^2 q_i} \leq 2(1+\lambda_i^2 w_i^2) \ .$$

Moreover, conditional on $Y_{(-i)} = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)$ (and Λ since all our arguments at this point are made conditional on Λ), w_i is $\mathcal{N}(0, \widehat{\mu}_i' \mathcal{S}_i^{-2} \widehat{\mu}_i)$ when the Y's are $\mathcal{N}(0, \mathrm{Id}_p)$, because $w_i = \widehat{\mu}_i' \mathcal{S}_i^{-1} Y_i$. Therefore,

$$\mathbf{E}\left(w_i^4|\Lambda\right) = 3\mathbf{E}\left((\widehat{\mu}_i' \mathcal{S}_i^{-2} \widehat{\mu}_i)^2|\Lambda\right) \ .$$

Almost by definition, we have $\widehat{\mu}_i' \mathcal{S}_i^{-1} \widehat{\mu}_i \leq 1$, since the vector $\mathbf{e}/\sqrt{n}$ has norm 1 and $W_i(W_i'W_i)^{-1}W_i'$ is a projection matrix (recall that $\mathcal{S}_i = W_i'W_i/n$ and $\widehat{\mu}_i' = \mathbf{e}'W_i/n$). So we would be done if we had uniform control on $|||\mathcal{S}_i^{-1}|||_2$. Let us now go around this difficulty.

- **Regularization interlude** Let us consider, for $t > 0$, $Z(t) = \widehat{\mu}'(\mathcal{S} + t\mathrm{Id}_p)^{-1}\widehat{\mu}$, where $\widehat{\mu}' = \mathbf{e}'W/n$. Clearly, $0 \leq Z(t) \leq Z_{n,p} = Z(0)$, because $\mathcal{S} + t\mathrm{Id}_p \succeq \mathcal{S} \succeq 0$ in the positive-semidefinite ordering. In other respects, the decomposition in Equation (8) is still valid if we replace Z_i by $Z_i(t)$ and $\mathcal{S}_i$ by $\mathcal{S}_i(t)$ everywhere. However, $|||(\mathcal{S}_i(t))^{-1}|||_2 \leq 1/t$. We therefore have

$$\widehat{\mu}_i' \mathcal{S}_i(t)^{-2} \widehat{\mu}_i \leq |||\mathcal{S}_i^{-1}(t)|||_2 \|\mathcal{S}_i^{-1/2} \widehat{\mu}_i\|_2^2 \leq \frac{\widehat{\mu}_i' \mathcal{S}_i^{-1}(t) \widehat{\mu}_i}{t} \leq \frac{\widehat{\mu}_i' \mathcal{S}_i^{-1} \widehat{\mu}_i}{t} \leq \frac{1}{t}$$

So applying the previous analysis and using the fact that $\widehat{\mu}_i'(\mathcal{S}_i(t))^{-2}\widehat{\mu}_i \leq 1/t$, we conclude that

$$\operatorname{var}(Z(t)|\Lambda) \leq \frac{8}{n^2} \sum_{i=1}^{n} (1 + 3\frac{\lambda_i^4}{t^2}) \ .$$

So under our assumptions, $Z(t)$ can be approximated, in probability, at least conditionally on Λ, by $\mathbf{E}(Z(t)|\Lambda)$. If we write the singular value decomposition of $W/\sqrt{n} = \sum_{i=1}^{p} \sigma_i u_i v_i'$, where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p$, we have $W\mathcal{S}^{-1}W'/n = \sum_{i=1}^{p} u_i u_i'$, $W(\mathcal{S}(t))^{-1}W'/n = \sum_{i=1}^{p} \sigma_i^2/(\sigma_i^2 + t) u_i u_i'$, and therefore

$$0 \leq Z_{n,p} - Z(t) = \frac{t}{n} \sum_{i=1}^{p} \frac{1}{\sigma_i^2 + t}(u_i'\mathbf{e})^2 \leq \frac{t}{\sigma_p^2 + t} \frac{1}{n} \sum_{i=1}^{p} (u_i'\mathbf{e})^2 \leq \frac{t}{\sigma_p^2 + t} \frac{\|\mathbf{e}\|_2^2}{n} = \frac{t}{\sigma_p^2 + t} \ .$$

To get the inequality above, we used the fact that the $\{u_i\}_{i=1}^p$ are orthonormal in $\mathbb{R}^n$, and can therefore be completed to form an orthonormal basis of this vector space. The quantities $u_i'\mathbf{e}$ are naturally the coefficients of $\mathbf{e}$ in this basis, and we know that their sum of squares should be the squared norm of $\mathbf{e}$, which is n.

Let us now call $\mathcal{L}_{\epsilon_0,\delta}$ the set of matrices Λ such that $p/N < 1 - \epsilon_0$ and $C_0(N-1)/(n-1) > \delta$. Under our assumptions, for a δ_0 bounded away from 0 (e.g $\delta_0 = 1/2C_0\eta_0$), $P(\Lambda \in \mathcal{L}_{\epsilon_0,\delta_0}) \to 1$. Let us pick such a δ_0. If $\Lambda \in \mathcal{L}_{\epsilon_0,\delta_0}$, according to Lemma B-1 and the proof of Theorem 4.1, if P_Λ denotes probability conditional on Λ,

$$P_\Lambda\left(\sigma_p \leq \sqrt{\delta_0}\left[(1 - \sqrt{1-\epsilon_0}) - t\right]\right) \leq \exp\left(-(n-1)\delta_0 t^2/C_0\right) \ .$$

Hence, when $\Lambda \in \mathcal{L}_{\epsilon_0,\delta_0}$, we can find, for any $u > 0$, an $\eta(u) > 0$,

$$P(|Z_{n,p} - Z(\eta(u))| > u) \leq f_n(\epsilon_0, \delta_0, \eta(u), u) = f_n(u) \ ,$$

where, $f_n(u) = f_n(\epsilon_0, \delta_0, \eta(u), u) \to 0$ as $n \to \infty$, for fixed u.

On the other hand, our conditional variance computations have established that, for any $\eta > 0$, $Z(\eta) - \mathbf{E}\left(Z(\eta)|\Lambda\right)$ converges in probability (conditional on Λ) to 0 if $\eta^{-2}\sum \lambda_i^4/n^2$ tends to 0. We note that $0 \leq Z_{n,p} \leq 1$ and that the same is true for $\gamma_n(u) = \mathbf{E}\left(Z(\eta(u))|\Lambda\right)$. Therefore, $|Z_{n,p} - \gamma_n(u)| \leq 1$ and $\mathbf{E}\left((Z_{n,p} - \gamma_n(u))^2|\Lambda\right)$ goes to zero, since

$$\begin{aligned}\mathbf{E}\left((Z - \gamma_n(u))^2|\Lambda\right) &\leq u^2 P(|Z_{n,p} - \gamma_n(u)| \leq u|\Lambda) + P(|Z_{n,p} - \gamma_n(u)| > u|\Lambda) \\ &\leq u^2 + P(|Z_{n,p} - Z(\eta(u))| > u/2|\Lambda) + \frac{4}{u^2}\mathrm{var}\left(Z(\eta(u))|\Lambda\right) \ .\end{aligned}$$

In other words, we also have, if $\Lambda \in \mathcal{L}_{\epsilon_0,\delta_0}$, for any $u > 0$,

$$\mathrm{var}\left(Z_{n,p}|\Lambda\right) \leq u^2 + f_n(u/2) + \frac{32}{u^2}\frac{1}{n^2}\sum_{i=1}^n \left(1 + 3\frac{\lambda_i^4}{\eta(u)^2}\right) \ .$$

Hence, if $\Lambda \in \mathcal{L}_{\epsilon_0,\delta_0}$ and $\sum_{i=1}^n \lambda_i^4/n^2 \to 0$, $\mathrm{var}\left(Z|\Lambda\right)$ goes to zero as n goes to infinity, and we conclude that, since $\mathbf{E}\left(Z|\Lambda\right) = p/n$,

$$Z - \frac{p}{n} \to 0 \text{ in probability, conditional on } \Lambda.$$

• **Deconditioning on Λ**

Let us call $\mathcal{L}^2_{\epsilon_0,\delta_0,t}$ the set of matrices such that $\mathcal{L}^2_{\epsilon_0,\delta_0,t} = \mathcal{L}_{\epsilon_0,\delta_0} \bigcap \{(\frac{1}{n^2}\sum_{i=1}^n \lambda_i^4) \leq t\}$. Our previous computations clearly show that we can find a function $g_n(u)$, with $g_n(u) \to 0$ as $n \to \infty$, such that, for any $u > 0$, when $\Lambda \in \mathcal{L}^2_{\epsilon_0,\delta_0,u^4\eta(u)^2} \triangleq \mathcal{L}^2(u)$, $\mathrm{var}\left(Z_{n,p}|\Lambda\right) \leq 97u^2 + g_n(u)$, and hence we have the "uniform bound", if $\Lambda \in \mathcal{L}^2(u)$,

$$P(\left|Z_{n,p} - \frac{p}{n}\right| > x|\Lambda) \leq \frac{97u^2 + g_n(u)}{x^2} \ .$$

Now under our assumptions, $P(\Lambda \in \mathcal{L}^2(u))$ goes to 1 for any given u, so we conclude, using the fact that

$$P(|Z_{n,p} - p/n| > x) \leq P[|Z_{n,p} - p/n| > x \,\&\Lambda \in \mathcal{L}^2(u)] + P[\Lambda \in \mathcal{L}^2(u)] \ ,$$

that

$$Z_{n,p} - \frac{p}{n} \to 0 \text{ in probability.}$$

This last statement is now understood of course unconditionally on Λ and this proves the first part of the theorem.

• **Proof of the second part of the theorem**

We now focus on the $\widehat{m}\widehat{\Sigma}^{-1}\widehat{m}$ part of the theorem. Let us call $\mathfrak{S} = \widetilde{X}'\widetilde{X}/n$. Then, $\frac{n-1}{n}\widehat{\Sigma} = \mathfrak{S} - \widehat{m}\widehat{m}'$. Therefore,

$$\frac{n}{n-1}\widehat{\Sigma}^{-1} = \mathfrak{S}^{-1} + \frac{\mathfrak{S}^{-1}\widehat{m}\widehat{m}'\mathfrak{S}^{-1}}{1 - \widehat{m}'\mathfrak{S}^{-1}\widehat{m}} \ .$$

Hence,

$$\frac{n}{n-1}\widehat{m}'\widehat{\Sigma}^{-1}\widehat{m} = \frac{\widehat{m}'\mathfrak{S}^{-1}\widehat{m}}{1 - \widehat{m}'\mathfrak{S}^{-1}\widehat{m}} = \frac{Z_{n,p}}{1 - Z_{n,p}} \ .$$

Since $Z_{n,p} \to \rho$ in probability, we have the result announced in the theorem. □

Now that we have proved Theorem 4.2, we need to turn to results that will allow us to handle the case of non-zero population mean, as well as questions such as the convergence of $\widehat{\mu}'\widehat{\Sigma}^{-1}v$, for deterministic v.

4.2.1 On quantities of the type $(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}$

Recall that the key quantity in the solution of Problem (QP-eqc-Emp), the problem of main interest in this paper, is of the form $\widehat{V}'\widehat{\Sigma}^{-1}\widehat{V}$. Therefore, it is important for us to understand quantities of the type

$$\zeta = \widehat{\mu}'\widehat{\Sigma}^{-1}v \;,$$

for a fixed vector v. At this point, we focus on the particular case where $\mu = \mathbf{E}\left(X_i\right) = 0$. To do so, we will need to study, if $\mathcal{S} = Y'\Lambda'\Lambda Y/n$,

$$\zeta = \frac{1}{n}\mathbf{e}'\Lambda Y \mathcal{S}^{-1}v \;,$$

for a fixed vector v. As it turns out, this random variable goes to zero in probability when for instance $\|v\|_2 = 1$.

Theorem 4.3. *Suppose v is a deterministic vector, with $\|v\|_2 = 1$. Suppose the assumptions stated in Theorem 4.2 hold and also that*

$$\frac{1}{n}\sum_{i=1}^{n}\lambda_i^2 \textit{ remains bounded with probability going to 1.} \qquad \text{(Assumption-BLb)}$$

Consider

$$\zeta = \frac{1}{n}\mathbf{e}'\Lambda Y \mathcal{S}^{-1}v \;,$$

where $\mathcal{S} = \frac{1}{n}Y'\Lambda^2 Y$. Then

$$\zeta \to 0 \textit{ in probability} \;.$$

Before giving the proof, we note that if the λ_i's are i.i.d and have a second moment, the "extra" condition on $\sum_{i=1}^n \lambda_i^2/n$ introduced in this Theorem (as compared to Theorem 4.2) is clearly satisfied by the law of large numbers.

Proof. The proof is quite similar to the proof of Theorem 4.2 above. We start by conditioning on Λ.

Let us call $\zeta(t)$ the quantity obtained when we replace $\mathcal{S}$ by $\mathcal{S}(t) = \mathcal{S} + t\text{Id}$ in the definition of ζ. Note that since Y is symmetric, $\zeta(t) \overset{\mathcal{L}}{=} -\zeta(t)$, conditionally on Λ, by arguments similar to those given in the proof of Lemma 4.2. Now $\zeta(t)$ clearly has an expectation (conditional on Λ), because $|||S^{-1}(t)|||_2 \leq 1/t$, for $t > 0$, so $\mathbf{E}\left(\zeta(t)|\Lambda\right) = 0$. Now recall Equation (6): with the notations used there,

$$\Lambda Y \mathcal{S}^{-1} = W_i\mathcal{S}_i^{-1} - \frac{\lambda_i^2}{n}\frac{r_i Y_i'\mathcal{S}_i^{-1}}{1+\lambda_i^2 q_i} + \lambda_i e_i Y_i'\mathcal{S}_i^{-1} - \lambda_i^3 q_i \frac{e_i Y_i'\mathcal{S}_i^{-1}}{1+\lambda_i^2 q_i} \;.$$

Let us now call $q_i(t) = Y_i'\mathcal{S}_i(t)^{-1}Y_i/n$, $w_i(t) = \mathbf{e}'W_i\mathcal{S}_i(t)^{-1}Y_i/n = \widehat{\mu}_i\mathcal{S}_i(t)^{-1}Y_i$ and $\theta_i(t) = Y_i'\mathcal{S}_i(t)^{-1}v$. Clearly, if $\zeta_i(t)$ is the random variable obtained by excluding Y_i from the computation of $\zeta(t)$ (e.g by replacing λ_i by 0), we have

$$\begin{aligned}\zeta(t) &= \zeta_i(t) - \frac{\lambda_i^2}{n}\frac{w_i(t)\theta_i(t)}{1+\lambda_i^2 q_i(t)} + \frac{\lambda_i\theta_i(t)}{n} - \frac{\theta_i(t)}{n}\frac{\lambda_i^3 q_i(t)}{1+\lambda_i^2 q_i(t)} \\ &= \zeta_i(t) + \frac{1}{n}\left(\frac{\lambda_i\theta_i(t)(1-\lambda_i w_i(t))}{1+\lambda_i^2 q_i(t)}\right) .\end{aligned}$$

We remark that $\theta_i(t)|(Y_{(-i)},\Lambda) \sim \mathcal{N}(0, v'\mathcal{S}_i^{-2}(t)v)$ and recall that $w_i|(Y_{(-i)},\Lambda) \sim \mathcal{N}(0, \widehat{\mu}_i'\mathcal{S}_i^{-2}\widehat{\mu}_i)$. Using the fact that $\|v\|_2 = 1$, $|||\mathcal{S}_i^{-2}(t)|||_2 \leq t^{-2}$, and the remarks we made in the proof of Theorem 4.2, we get that $\mathbf{E}\left([\theta_i(t)]^{2k}\middle|\,(Y_{(-i)},\Lambda)\right) \leq C_k t^{-2k}$, $\mathbf{E}\left([w_i(t)]^{2k}\middle|\,(Y_{(-i)},\Lambda)\right) \leq C_k t^{-k}$, where $C_1 = 1$ and $C_2 = 3$. We also have

$$\left[\lambda_i\theta_i(t)(1-\lambda_i w_i(t))\right]^2 \leq 2\left[\lambda_i^2\theta_i^2(t) + \lambda_i^4\theta_i^2(t)w_i^2(t)\right] .$$

Hence, simply using the fact that $2(ab)^2 \leq (a^4 + b^4)$, we get

$$\mathbf{E}\left(\left(\frac{\lambda_i\theta_i(t)(1-\lambda_i w_i(t))}{1+\lambda_i^2 q_i(t)}\right)^2\middle|\,\Lambda\right) \leq 2\frac{2\lambda_i^2}{t^2} + 3\lambda_i^4\left(\frac{1}{t^2}+\frac{1}{t^4}\right) \;.$$

We conclude by the Efron-Stein inequality that, when Λ is such that $\sum_{i=1}^n \lambda_i^4/n^2 \to 0$, for any $t > 0$,

$$\zeta(t) \to 0 \text{ in probability, conditionally on } \Lambda.$$

As before, let us call $\mathcal{L}_{\epsilon_0,\delta}$ the set of matrices Λ such that $p/N < 1 - \epsilon_0$ and $C_0(N-1)/(n-1) > \delta$. Recall that under our assumptions, for δ_0 bounded away from 0 (e.g $\delta_0 = C_0\eta_0/2$), $P(\Lambda \in \mathcal{L}_{\epsilon_0,\delta_0}) \to 1$.

As we saw before, when $\Lambda \in \mathcal{L}_{\epsilon_0,\delta_0}$, $|||\mathcal{S}^{-1}|||_2$ is bounded with high-probability (conditional on Λ), so we conclude that, for any $\eta > 0$, we can find a t such that

$$|||\mathcal{S}^{-1} - \mathcal{S}^{-1}(t)|||_2 < \eta \text{ with probability (conditional on } \Lambda\text{) going to 1 .}$$

We also notice that conditionally on Λ, $\widehat{\mu} \sim \mathcal{N}\left(0, \frac{\sum \lambda_i^2}{n^2}\mathrm{Id}_p\right)$ and hence, $\|\widehat{\mu}\|_2^2 \sim \chi_p^2/n(\sum \lambda_i^2)/n$. We recall that $\|v\|_2 = 1$, and since

$$|\zeta - \zeta(t)| \leq \|\widehat{\mu}\|_2 |||\mathcal{S}^{-1} - \mathcal{S}^{-1}(t)|||_2 \|v\|_2 ,$$

we conclude that with high-probability (conditional on Λ), for any $\eta > 0$, $|\zeta - \zeta(t)| \leq \eta$ and finally,

$$\zeta \to 0 \text{ in probability, conditionally on } \Lambda .$$

Now along the same lines as what was done in the proof of Theorem 4.2, we can make all these probability bounds uniform in Λ when Λ is in a set of matrices such as $\mathcal{L}_{\epsilon_0,\delta_0}$ and when we also have bounds on $\sum_{i=1}^n \lambda_i^4/n^2$ and $\sum_{i=1}^n \lambda_i^2/n$. Under our assumptions, the set of Λ for which these conditions hold has measure going to 1, so we can finally conclude - along the same lines (omitted here) as in the proof of Theorem 4.2 - that, unconditionally on Λ,

$$\zeta \to 0 \text{ in probability .}$$

□

After these preliminaries, we can finally state the theorem of main interest. Recall that under the assumptions of Theorem 4.1, if v is deterministic,

$$\frac{v'\widehat{\Sigma}^{-1}v}{v'\Sigma^{-1}v} \to \mathfrak{s} \text{ in probability } ,$$

where $\mathfrak{s}$ is defined in Equation (4).

Theorem 4.4. *Suppose that $X_i = \mu + \lambda_i \Sigma^{1/2} Y_i$, where Y_i are i.i.d $\mathcal{N}(0, \mathrm{Id}_p)$ and $\{\lambda_i\}_{i=1}^n$ are random variables, independent of $\{Y_i\}_{i=1}^n$. Let v be a deterministic vector. Suppose that $\rho_n = p/n$ has a finite non-zero limit, ρ and that $\rho \in (0,1)$.*

We call $\tau_i = \lambda_i^2$. We assume that $\tau_i \neq 0$ for all i as well as

$$\frac{1}{n^2}\sum_{i=1}^n \lambda_i^4 \to 0 \textit{ in probability, and } \frac{1}{n}\sum_{i=1}^n \lambda_i^2 \textit{ remains bounded in probability.} \qquad \text{(Assumption-BL)}$$

If $\tau_{(i)}$ is the i-th largest τ_k, we assume that we can find a random variable $N \in \mathbb{N}$ and positive real numbers ϵ_0 and C_0 such that

$$\begin{cases} P(p/N < 1 - \epsilon_0) \to 1 \textit{as } n \to \infty , \\ P(\tau_{(N)} > C_0) \to 1 , \\ \exists \eta_0 > 0 \textit{ such that } P(N/n > \eta_0) \to 1 \textit{ as } n \to \infty . \end{cases} \qquad \text{(Assumption-BB)}$$

We also assume that the empirical distribution of τ_i's converges weakly in probability to a deterministic limit G.

We call Λ the $n \times n$ diagonal matrix with $\Lambda(i,i) = \lambda_i$, Y the $n \times p$ matrix whose i-th row is Y_i, $W = \Lambda Y$ and $\mathcal{S} = W'W/n = \sum_{k=1}^n \lambda_k^2 Y_k Y_k'/n$. Finally, we use the notation $\widehat{\omega} = W'\mathbf{e}/n$, $\widetilde{\mu} = \Sigma^{-1/2}\mu$.

Then, we have, for $\mathfrak{s}$ defined as in Equation (4),

$$\frac{\widehat{\mu}'\widehat{\Sigma}^{-1}v}{\sqrt{v'\Sigma^{-1}v}} = \frac{\mu'\widehat{\Sigma}^{-1}v}{\sqrt{v'\Sigma^{-1}v}} + \mathrm{o}_P(1) = \mathfrak{s}\frac{\mu'\Sigma^{-1}v}{\sqrt{v'\Sigma^{-1}v}} + \mathrm{o}_P\left(1 \vee \frac{\mu'\Sigma^{-1}v}{\sqrt{v'\Sigma^{-1}v}}\right) , \tag{9}$$

the second statement holding if for instance μ and v are such that the first set of conditions in Lemma 4.1 are met.
Also,

$$\widehat{\mu}'\widehat{\Sigma}^{-1}\widehat{\mu} = \mu'\widehat{\Sigma}^{-1}\mu + \frac{\rho_n}{1-\rho_n} + 2\frac{n-1}{n}\frac{\widehat{\omega}'\mathcal{S}^{-1}\widetilde{\mu}}{1-\widehat{\omega}'\mathcal{S}^{-1}\widehat{\omega}} + \mathrm{o}_P(1) \tag{10}$$

and we recall that $\widehat{\omega}'\mathcal{S}^{-1}\widetilde{\mu}/\|\widetilde{\mu}\| = \mathrm{o}_P(1)$ and $\widehat{\omega}'\mathcal{S}^{-1}\widehat{\omega} = p/n + \mathrm{o}_P(1)$.

To be able to exploit Equation (10) in practice, we make the following remarks. We can consider three cases, having to do with the size of $\mu'\Sigma^{-1}\mu = \|\widetilde{\mu}\|_2^2$.

1. If $\mu'\Sigma^{-1}\mu \to 0$, then, $\widehat{\mu}'\widehat{\Sigma}^{-1}\widehat{\mu} = \frac{\rho_n}{1-\rho_n} + \mathrm{o}_P(1)$.

2. If $\mu'\Sigma^{-1}\mu \to \infty$, then $\widehat{\mu}'\widehat{\Sigma}^{-1}\widehat{\mu} \sim \mathfrak{s}\mu'\Sigma^{-1}\mu$.

3. Finally, if $\mu'\Sigma^{-1}\mu$ stays bounded away from 0 and infinity,

$$\widehat{\mu}'\widehat{\Sigma}^{-1}\widehat{\mu} \simeq \mathfrak{s}\mu'\Sigma^{-1}\mu + \frac{\rho_n}{1-\rho_n} + \mathrm{o}_P(1) .$$

A noticeable feature of these results is that the "extra bias" $\kappa_n = \rho_n/(1-\rho_n)$, which comes essentially from mis-estimation of μ, is constant within the class of elliptical distributions considered here. This should be contrasted with the "scaling", $\mathfrak{s}$, which strongly depends on the empirical distribution of the λ_i^2's.

We now give a brief proof of Theorem 4.4.

Proof of Theorem 4.4. We first note that $\Sigma^{1/2}\widehat{\omega} = \widehat{m}$ in the notation of Theorem 4.2. Also, $\widehat{\mu} = \mu + \Sigma^{1/2}\widehat{\omega} = \mu + \widehat{m}$. Finally,

$$\frac{n-1}{n}\widehat{\Sigma} = \Sigma^{1/2}\mathcal{S}\Sigma^{1/2} - \widehat{m}\widehat{m}' = \Sigma^{1/2}\left(\mathcal{S} - \widehat{\omega}\widehat{\omega}'\right)\Sigma^{1/2} .$$

Proof of Equation (10). By writing $\widehat{\mu} = \mu + \widehat{m}$, we clearly have

$$\widehat{\mu}'\widehat{\Sigma}^{-1}\widehat{\mu} = \mu'\widehat{\Sigma}^{-1}\mu + 2\widehat{m}'\widehat{\Sigma}^{-1}\mu + \widehat{m}'\widehat{\Sigma}^{-1}\widehat{m} .$$

We have already seen in Theorem 4.2 that the third term tends to $\kappa = \rho/(1-\rho)$. On the other hand, half of the middle term is equal to

$$\frac{n}{n-1}\widehat{\omega}'\left(\mathcal{S} - \widehat{\omega}\widehat{\omega}'\right)^{-1}\widetilde{\mu} .$$

Since $(\mathcal{S} - \widehat{\omega}\widehat{\omega}')^{-1} = \mathcal{S}^{-1} + \mathcal{S}^{-1}\widehat{\omega}\widehat{\omega}'\mathcal{S}^{-1}/(1-\widehat{\omega}'\mathcal{S}^{-1}\widehat{\omega})$, we have

$$\frac{n}{n-1}\widehat{m}'\widehat{\Sigma}^{-1} = \widehat{\omega}'\mathcal{S}^{-1}\left(1 + \frac{\widehat{\omega}'\mathcal{S}^{-1}\widehat{\omega}}{1-\widehat{\omega}'\mathcal{S}^{-1}\widehat{\omega}}\right)\Sigma^{-1/2} = \frac{1}{1-\widehat{\omega}'\mathcal{S}^{-1}\widehat{\omega}}\widehat{\omega}'\mathcal{S}^{-1}\Sigma^{-1/2} ,$$

and we deduce the result of Equation (10). We now remark that $\widehat{\omega}'\mathcal{S}^{-1}\widehat{\omega}$ is equal to the quantity $Z_{n,p}$ in Theorem 4.2. The fact that $\widehat{\omega}'\mathcal{S}^{-1}\widetilde{\mu}/\|\widetilde{\mu}\| = \mathrm{o}_P(1)$ follows from applying Theorem 4.3 with $v = \widetilde{\mu}/\|\widetilde{\mu}\|_2$.
Proof of Equation (9) The proof of this result follows from a decomposition similar to the one we just made. Clearly the only question is whether $\widehat{m}'\widehat{\Sigma}^{-1}v/v'\Sigma^{-1}v$ goes to 0. As we just saw,

$$\frac{n}{n-1}\widehat{m}'\widehat{\Sigma}^{-1}v = \frac{1}{1-\widehat{\omega}'\mathcal{S}^{-1}\widehat{\omega}}\widehat{\omega}'\mathcal{S}^{-1}\Sigma^{-1/2}v .$$

The results of Theorem 4.3 guarantee that

$$\frac{\widehat{\omega}'\mathcal{S}^{-1}\Sigma^{-1/2}v}{\|\Sigma^{-1/2}v\|} \to 0 \text{ in probability} .$$

Since $\widehat{\omega}'\mathcal{S}^{-1}\widehat{\omega}$ tends to $\rho < 1$ and $\|\Sigma^{-1/2}v\|^2 = v'\Sigma^{-1/2}v$, we have shown the result stated in Equation (9). □

4.3 On the effect of correlation between observations

It is clear that in financial practice and other applied settings, the assumption that the returns (or observed data vectors) are independent is often questionable. So for quadratic programs with linear equality constraints (including the Markowitz problem but also going beyond it), it is natural to ask what is the impact of correlation in our observations on the empirical solution of the problem. In our notation, this means that the vectors X_i and X_j are correlated - we refer to this situation as the correlated case or as the case of temporal correlation.

Our work on the elliptical case comes in handy here and allows us to also draw conclusions concerning the correlated case. We consider a particular model, namely we assume that the $n \times p$ data matrix X is given by

$$X = \mathbf{e}_n \mu' + \Lambda Y \Sigma^{1/2} \text{ , where } \Lambda \text{ is a deterministic but } \textbf{not} \text{ necessarily diagonal matrix,}$$

and Y is a matrix with i.i.d $\mathcal{N}(0,1)$ entries. We assume throughout that Λ is full rank. The model we consider now is more general than the one we looked at before, since if $\Lambda = \text{Id}_n$, we get the i.i.d Gaussian case, and if Λ is diagonal we are back in an "elliptical" case (where the ellipticity parameters are assumed to be deterministic, which amounts to doing computations conditional on Λ). But when Λ is not diagonal, X_i and X_j might be correlated. (In all the situations where Λ is deterministic, the marginal distribution of X_i is $\mathcal{N}(\mu, s_i^2 \Sigma)$, where s_i is the norm of the i-th row of Λ.)

Because we want to focus here on robustness questions arising when going from independent Gaussian random variables to correlated ones, we will assume throughout that Λ is deterministic. (Allowing Λ to be random simply requires some minor technical modifications but would make the exposition a bit less clear.) Our main results in this Subsection can be interpreted as saying that that the Gaussian analysis of Section 3, carried out in the setting of independent observations, is not robust against this independence assumptions. The results change quite significantly when the vectors of observations are correlated.

In general, we write the singular value decomposition of the $n \times n$ matrix Λ as $\Lambda = ADB'$ (see Horn and Johnson (1990), p.414), where A and B are orthogonal and D is diagonal. Therefore, $AA' = \text{Id}_n$, and

$$\frac{1}{n}(X - \mathbf{e}_n \mu')'(X - \mathbf{e}_n \mu') = \frac{1}{n} \Sigma^{1/2} Y' B D^2 B' Y \Sigma^{1/2} \overset{\mathcal{L}}{=} \frac{1}{n} \Sigma^{1/2} Y' D^2 Y \Sigma^{1/2} \; .$$

So we are almost back in the elliptical case. The key difference now is that what will matter in our analysis are not the diagonal entries of $\Lambda'\Lambda$, but rather its eigenvalues (see Proposition 4.1). Also, we will see (in Proposition 4.2) that the results change quite significantly when we look at quantities like $\widehat{\mu}'\widehat{\Sigma}^{-1}\widehat{\mu}$.

4.3.1 On quadratic forms involving $\widehat{\boldsymbol{\Sigma}}^{-1}$

As a counterpart to Theorem 4.1, we have the following Proposition.

Proposition 4.1. *Suppose the $n \times p$ data matrix X (whose i-th row is the i-th vector of observations can be written as*

$$X = \mathbf{e}_n \mu' + \Lambda Y \Sigma^{1/2} \text{ , } \textit{where } \Lambda \textit{ is a deterministic but } \textbf{\textit{not}} \textit{ necessarily diagonal matrix,}$$

Suppose that the eigenvalues of $\Lambda'\Lambda$ satisfy (Assumption-BB) *with a deterministic N and that the spectral distribution of $\Lambda'\Lambda$ converges weakly to a probability distribution G. Suppose also that $p/n \to \rho \in (0,1)$. Call $\widehat{\Sigma}$ the classical sample covariance matrix, i.e*

$$\widehat{\Sigma} = \frac{1}{n}(X - \bar{X})'(X - \bar{X}) \; .$$

Then, if v is a deterministic vector, we have

$$\frac{v'\widehat{\Sigma}^{-1}v}{v'\Sigma^{-1}v} \to \mathfrak{s} \textit{ in probability} \quad ,$$

where $\mathfrak{s}$ satisfies, if G is the limiting spectral distribution $\Lambda'\Lambda$

$$\int \frac{dG(\tau)}{1 + \rho\tau\mathfrak{s}} = 1 - \rho \; .$$

The proposition shows that Theorem 4.1 essentially applies again, however now what matters - unsurprisingly - are the singular values of Λ and not its diagonal entries. The proof of Proposition 4.1, or rather the adjustments needed to make the proof of Theorem 4.1 go through, are given in the Appendix, Subsection C-1.

4.3.2 On quadratic forms involving $\widehat{\mu}$ and $\widehat{\Sigma}^{-1}$

This is the situation where the results are most different from that of the uncorrelated case. Once again, here we will be content to just state the results - a detailed justification of our claims is in the Appendix, Subsection C-2.

As before, the most complicated aspect of the problem is to understand quantities of the type $\widehat{\mu}'\widehat{\Sigma}^{-1}\widehat{\mu}$, in the situation where $\mu = 0$. In this setting, we have the following result.

Proposition 4.2. *Suppose the $n \times p$ data matrix $\widetilde{X}$ is such that, for Y an $n \times p$ matrix with i.i.d $\mathcal{N}(0,1)$ entries, and Λ a deterministic matrix,*

$$\widetilde{X} = \Lambda Y \Sigma^{1/2} .$$

We assume that (Assumption-BB) *holds for the eigenvalues of $\Lambda'\Lambda$, for a deterministic sequence $N(n)$. We write the singular value decomposition of Λ as $\Lambda = ADB'$.*

We call $\mathcal{S} = \widetilde{X}'\widetilde{X}/n$ and $\widehat{m} = \Sigma^{1/2}Y'\Lambda'\mathbf{e}/n$, i.e the sample mean of the columns of $\widetilde{X}$. We denote by d_i the diagonal elements of D, and $\widetilde{Y} = B'Y \overset{\mathcal{L}}{=} Y$. We also call

$$F = \frac{1}{n}\sum_{i=1}^{n} d_i^2 \widetilde{Y}_i\widetilde{Y}_i' \ , \ F_i = F - \frac{1}{n} d_i^2 \widetilde{Y}_i\widetilde{Y}_i' \ , P = D\widetilde{Y}\left(\widetilde{Y}'D^2\widetilde{Y}\right)^{-1}\widetilde{Y}'D \ .$$

If we call $\omega = A'\mathbf{e}$, and $q_i = Y_i'F_i^{-1}Y_i/n$, we have, if $\|\omega\|_4^4/n^2$ and $\|d\|_4^4/n^2 \to 0$,

$$\widehat{m}'\mathcal{S}^{-1}\widehat{m} - \kappa(n,p) \to 0 \quad \textit{in probability}$$

where

$$\kappa(n,p) = \frac{1}{n}\sum_{i=1}^{n}\omega_i^2 \mathbf{E}\left(P(i,i)\right) \ \textit{and } P(i,i) = 1 - \frac{1}{1+q_i d_i^2} \ .$$

Further,

$$\widehat{m}\widehat{\Sigma}^{-1}\widehat{m} - \frac{\kappa(n,p)}{1-\kappa(n,p)} \to 0 \ \textit{in probability} \ .$$

Furthermore, under the above assumptions, if the spectral distribution of $\Lambda'\Lambda$ converges to G and $(\sum_{i=1}^{n}\omega_i^2 d_i^2)/n$ remains bounded, a result similar to Theorem 4.4 holds, with $\mathfrak{s}$ being computed by solving Equation (4) *with the corresponding G and $\kappa(n,p)$ playing the role of $\rho_n/(1-\rho_n)$.*

Essentially the previous proposition tells us that when dealing with correlated variables, the new $\kappa(n,p)$ replaces the old $\kappa = \rho/(1-\rho)$. We note that there are no inconsistencies with our previous results as $\sum_i P(i,i) = \operatorname{trace}(P) = p$ and in the elliptical case, $\omega_i^2 = 1$, so the previous proposition is consistent with the results we have obtained in the elliptical case. We also remark that $\|\omega\| = \sqrt{n}$, since A is orthogonal.

Finally, in the case where the d_i's have a limiting spectral distribution and satisfy (Assumption-BB), further computations show that $q_i - \rho_n\mathfrak{s} \to 0$. However, this does not help (in general) in getting a simpler expression for $\kappa(n,p)$.

4.4 On the bootstrap

An interesting aspect of the analysis of elliptical models is that it also shed lights on the properties of the bootstrap in this context. As a matter of fact, the non-parametric bootstrap yields covariance matrices that have a structure similar to those computed from elliptical distributions: if we call D the diagonal matrix whose i-th diagonal entry is the number of times observation X_i appears in our bootstrap sample, we have, if $\widehat{\Sigma}^*$ is the bootstrapped covariance matrix,

$$\widehat{\Sigma}^* = \frac{1}{n-1}X'DX - \frac{n}{n-1}\widehat{\mu}^*(\widehat{\mu}^*)' ,$$

where X is our original data matrix, and $\widehat{\mu}^*$ is the sample mean of our bootstrap sample, which can also be written $\widehat{\mu}^* = X'D\mathbf{e}/n$. Unless otherwise noted, we assume in the discussion that follows that the population mean μ is 0. Since the covariance matrix is shift-invariant, we can make this assumption without loss of generality. We call

$$\mathfrak{S}^* = \frac{1}{n} X'DX \text{ , and } \mathcal{S}^* = \Sigma^{-1/2}\mathfrak{S}^*\Sigma^{-1/2} \; .$$

As we will see shortly, understanding the properties of $\widehat{\Sigma}^*$ boils down to understanding those of $\mathcal{S}^*$ so we will focus on this slightly more convenient object in this short discussion.

We note that if X is Gaussian, $\mathfrak{S}^*$ can be thought of as a "covariance matrix" computed from the elliptical data $\widetilde{X}_i = d_i^{1/2} X_i$. The same remark applies when X is elliptical - i.e, for us, $X_i = \lambda_i \mathcal{N}(0, \Sigma)$: all we need to do is change the "ellipticity parameter" λ_i to $\sqrt{d_i}\lambda_i$. The same remark is also applicable to the case of correlated observations, i.e $X = \Lambda Y \Sigma^{1/2}$, where Λ is not diagonal anymore. Studying the bootstrap properties of such a model is the same as studying that of the model where we replace Λ by $\sqrt{D}\Lambda$. We therefore would like to apply directly all the results we have obtained above in our study of elliptical models to better understand the bootstrap. For quantities of the form $v'(\widehat{\Sigma}^*)^{-1}v$, we will see that we can essentially do it, but differences will appear when dealing with $(\widehat{\mu}^*)'(\widehat{\Sigma}^*)^{-1}\widehat{\mu}^*$, which yields statistics that are not exactly analogous to corresponding statistics appearing in the elliptical case.

Our focus will be on bias properties of bootstrapped replications - so we will aim for convergence in probability results and not fluctuation behavior. Our overall strategy here is to show convergence in probability of the quantities we are interested in as functions of both the d_i's and X_i's. We will derive the convergence properties of our bootstrapped statistics by then conditioning on the data and arguing that with high probability (over the X_i's), this does not change the results much. We first give some needed background on the bootstrap in subsubsections 4.4.1 and 4.4.2, then turn to properties of quantities like $v'(\widehat{\Sigma}^*)^{-1}v$ (in 4.4.3) and finally study $(\widehat{\mu}^*)'(\widehat{\Sigma}^*)^{-1}\widehat{\mu}^*$ (in 4.4.4), where we will see (in Proposition 4.5) some key differences with the elliptical case. We conclude this subsection with a brief discussion of the parametric bootstrap and the conclusions that can be reached about it through our results.

4.4.1 A remark on needed convergence properties

Making statements about bootstrapped statistics requires us to make statements that are conditional on the observed data. This is not a trivial matter for the statistics we deal with since they cannot be easily described in terms of simple formulas involving the original observations. However, we can take a roundabout way: by showing joint convergence in probability (joint here refers to the "new" data being the vectors of bootstrapped weights and observations), we can obtain interesting conclusions conditional on the data. Though this is not difficult to show, we give full arguments here for the sake of completeness.

We will look at our statistics as functions of the number of times an observation appears in the sample and also, of course, of our observations. In other words, the original statistic, T_n can be written

$$T_n = f(1, \ldots, 1, X_1, \ldots, X_n)$$

and, the bootstrapped version T_n^* is, if observation X_i appears w_i^* times in the bootstrap sample,

$$T_n^* = f(w_1^*, \ldots, w_n^*, X_1, \ldots, X_n) \; .$$

The following simple proposition is used repeatedly in our bootstrap work.

Proposition 4.3. *Let us consider a statistic $T_n = f(w_1, \ldots, w_n, X_1, \ldots, X_n)$, where w_i is the number of times X_i appears in our sample. Suppose that the vector of weights, w is independent of the data matrix X. Denote by $\mathcal{Q}_n$ the joint probability distribution of the w_i's, $\mathcal{P}_n$ the joint probability distribution of the X_i's and $\mathcal{R}_n = \mathcal{Q}_n \times \mathcal{P}_n$ the probability distribution of $(w_1, \ldots, w_n, X_1, \ldots, X_n)$.*

Suppose we have established that T_n tends in $\mathcal{R}_n$-probability to c, a deterministic object, as $n \to \infty$.

Then we have: with $\mathcal{P}_n$-probability going to 1 as $n \to \infty$,

$$T_n | \{X_i\}_{i=1}^n \to c \text{ in } \mathcal{Q}_n\text{-probability.}$$

In other words, calling $\mathcal{X}_n = \{X_i\}_{i=1}^n$, for all $\epsilon, \eta > 0$, if $Q_n(\epsilon) = \mathcal{Q}_n(|T_n - c| > \epsilon | \mathcal{X}_n)$, $\mathcal{P}_n(Q_n(\epsilon) > \eta) \to 0$ as n tends to infinity.

In the case where the weights w_i are obtained by standard bootstrapping, Q_n is Multinomial$(1/n, \ldots, 1/n, n)$. Then, $T_n | X_n$ has the distribution of the usual bootstrap quantity T_n^*. We will focus on this case more specifically later.

Proof. The proof and the statement are almost obvious but we include them for the sake of completeness. Let us call $\tau_n = |T_n - c|$ and $\mathcal{X}_n = \{X_1, \ldots, X_n\}$. By assumption, $\tau_n \to 0$ in $\mathcal{R}_n$ probability. Hence,

$$\mathbf{E}_{\mathcal{R}_n}\left(1_{\tau_n > \epsilon}\right) = \mathbf{E}_{\mathcal{P}_n}\left(\mathbf{E}_{\mathcal{Q}_n}\left[1_{\tau_n > \epsilon} | \mathcal{X}_n\right]\right) \to 0 \ .$$

Let us call $Q_n(\epsilon) = \mathcal{Q}_n(|T_n - c| > \epsilon | \mathcal{X}_n)$. Clearly, $0 \leq Q_n(\epsilon) \leq 1$ and $\mathbf{E}_{\mathcal{P}_n}\left(Q_n(\epsilon)\right) \to 0$, so for any $\eta > 0$,

$$\mathcal{P}_n(Q_n(\epsilon) > \eta) \to 0 \ .$$

□

We now investigate the case of the classical bootstrap, i.e the situation in which $\mathcal{Q}_n$ is Multinomial$(\frac{1}{n}, \ldots, \frac{1}{n}, n)$.

4.4.2 Empirical distribution of bootstrap weights

As we saw in Theorem 4.1, the empirical distribution of the ellipticity parameters affect crucially statistics of the type $v'\widehat{\Sigma}^{-1}v$, so to understand the effect of bootstrapping, we need to understand the empirical distribution of the bootstrap weights. This question has surely been investigated but we did not find a good reference so we provide the result and a simple proof for the convenience of the reader.

Proposition 4.4. *Let the vector w be distributed according to a Multinomial$(\frac{1}{n}, \ldots, \frac{1}{n}, n)$ distribution. Call F_n the empirical distribution of the vector w. Then*

$$F_n \Longrightarrow \text{Po}(1) \textit{ in probability } ,$$

where Po(1) *is the Poisson distribution with parameter 1.*

Proof of the proposition: Let us first start by an elementary remark: suppose $\pi_1, \ldots, \pi_n$ are i.i.d with distribution Po(1). Call $\Pi_n = \sum_{i=1}^n \pi_i$. Then

$$(\pi_1, \ldots, \pi_n) \,|\, \{\Pi_n = n\} \sim \text{Multinomial}(\frac{1}{n}, \ldots, \frac{1}{n}, n) \ .$$

This result is a simple application of Bayes' rule and the fact that $\Pi_n \sim \text{Po}(n)$.

Let us now show that is f is bounded and continuous, and if $W \sim \text{Po}(1)$,

$$\mathbf{E}_{F_n}(f) = \frac{1}{n}\sum_{i=1}^n f(w_i) \to \mathbf{E}\left(f(W)\right) \text{ in probability } .$$

To do so, we note that $w_i \sim \text{Binomial}(n, 1/n)$ and therefore its marginal distribution is asymptotically Po(1). Therefore,

$$\mathbf{E}\left(\mathbf{E}_{F_n}(f)\right) \to \mathbf{E}\left(f(W)\right) \ .$$

Now all we need to do is therefore to show that $\text{var}\left(\mathbf{E}_{F_n}(f)\right)$ goes to zero. Clearly, by independence of the π_i's,

$$\text{var}\left(\frac{1}{n}\sum_{i=1}^n f(\pi_i)\right) = \frac{1}{n}\text{var}\left(f(W)\right) = \text{O}\left(\frac{1}{n}\right) \ ,$$

because f is bounded. But our first remark implies that

$$\text{var}\left(\mathbf{E}_{F_n}(f)\right) = \text{var}\left(\frac{1}{n}\sum_{i=1}^n f(\pi_i) \middle| \Pi_n = n\right)$$

Now,

$$\operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} f(\pi_i)\right) = \mathbf{E}\left(\operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} f(\pi_i)\middle|\Pi_n\right)\right) + \operatorname{var}\left(\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} f(\pi_i)\middle|\Pi_n\right)\right)$$
$$\geq \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} f(\pi_i)\middle|\Pi_n = n\right) P(\Pi_n = n) \ .$$

Since Π_n has Po(n) distribution, $P(\Pi_n = n) \sim 1/\sqrt{2\pi n}$. Hence,

$$\operatorname{var}\left(\mathbf{E}_{F_n}(f)\right) = \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} f(\pi_i)|\Pi_n = n\right) = \mathrm{O}(n^{-1/2}) \to 0 \ ,$$

and the result is established. □

We will also need later to use on the following (coarse) fact:

Fact 4.1. *Let the vector w be distributed according to a Multinomial$(\frac{1}{n},\ldots,\frac{1}{n},n)$ distribution. Then*

$$P(\max_{i=1,\ldots,n} w_i > (\log n)) = \mathrm{O}\left(\frac{n^{3/2}}{(\log n)!}\right) \ .$$

In particular, this probability goes to 0 faster than any n^{-a}, $a > 0$.

The proof of the fact is elementary, and relies on the representation used above for the vector w, a simple union bound, the fact that $P(\text{Po}(n) = n) \sim n^{-1/2}$ and the fact that $P(\text{Po}(1) \geq M) \leq (M!)^{-1}M/(M-1)$ which is easy to see by writing explicitly the probability we are trying to compute.

With these preliminaries behind us, we are now ready to tackle the question of understanding the (first-order) bootstrap properties of the statistics appearing in the study of quadratic programs with linear equality constraints.

4.4.3 On inverse covariance matrices computed from bootstrapped data

Our aim in this subsubsection and the next is to find analogs to Theorems 4.1 and Theorems 4.4. Our first result along these lines is an analog of Theorem 4.1.

We present the result in the case of Gaussian data, where we can get a somewhat explicit expression for the quantity we care about, and discuss possible extensions below.

Theorem 4.5. *Suppose we observe n i.i.d observations X_i, where X_i are i.i.d in $\mathbb{R}^p$ with distribution $\mathcal{N}(\mu, \Sigma_p)$. Call $\rho_n = p/n$ and assume that $\rho_n \to \rho \in (0, 1 - \mathrm{e}^{-1})$. Call $\widehat{\Sigma}^*$ the covariance matrix computed after bootstrapping the X_i's. Call $\mathcal{P}_n$ the joint distribution of the X_i's.*

If v is a (sequence of) deterministic vector, then conditional on $\{X_i\}_{i=1}^n$, with high $\mathcal{P}_n$ probability,

$$\frac{v'(\widehat{\Sigma}^*)^{-1}v}{v'\Sigma^{-1}v} \to \mathfrak{s} \textit{ in probability} \ ,$$

where $\mathfrak{s}$ satisfies, if G is a Po(1) *distribution*

$$\int \frac{dG(\tau)}{1 + \rho\tau\mathfrak{s}} = 1 - \rho \ . \tag{11}$$

Proof. As before, we call $\mathcal{Q}_n$ the law of the bootstrap weights (i.e Multinomial$(\frac{1}{n},\ldots,\frac{1}{n},n)$) and $\mathcal{R}_n = \mathcal{Q}_n \times \mathcal{P}_n$. Without loss of generality, we can assume that $\mu = 0$. Let us call D the diagonal matrix containing the bootstrap weights. We have $\widehat{\mu}^* = X'D\mathbf{e}/n$. Also, it is true that

$$\widehat{\Sigma}^* = \frac{1}{n-1}(X - \frac{\mathbf{e}(\widehat{\mu}^*)'}{n})'D(X - \frac{\mathbf{e}(\widehat{\mu}^*)'}{n}) \ .$$

Since $\mathbf{e}'D\mathbf{e} = n$, we also have

$$(n-1)\widehat{\Sigma}^* = X'D(\text{Id} - \frac{\mathbf{e}\mathbf{e}'D}{n})X = X'D^{1/2}\left(\text{Id} - \frac{1}{n}D^{1/2}\mathbf{e}\mathbf{e}'D^{1/2}\right)D^{1/2}X \; .$$

Because X is of the form $X = Y\Sigma^{1/2}$ under our assumptions, we see that

$$\begin{aligned}\widehat{\Sigma}^* &= \Sigma^{1/2}\mathcal{S}^*\Sigma^{1/2} \; , \text{ where} \\ \mathcal{S}^* &= \frac{1}{n-1}Y'D^{1/2}\left(\text{Id} - \frac{1}{n}D^{1/2}\mathbf{e}\mathbf{e}'D^{1/2}\right)D^{1/2}Y \; .\end{aligned}$$

If we call $\delta = D^{1/2}\mathbf{e}$, we have $\|\delta\|_2^2 = n$, because the sum of the bootstrap weights is n. Therefore, $H_\delta = \text{Id}_n - \delta\delta'/n \succeq 0$. Also, H_δ (like H) is a projection matrix and a rank 1 perturbation of Id_n.

The situation is therefore very similar to the question we studied in Theorem 4.1, except that $H = \text{Id} - \mathbf{e}\mathbf{e}'/n$ is replaced by $H_\delta = \text{Id}_n - \delta\delta'/n$. All the arguments given there hold provided we can show that (Assumption-BB) is satisfied for the bootstrap weights in the situation we have here.

Now let us call N the number of non-zero bootstrap weights. In the notation of Theorem 4.1, $\lambda_i = \sqrt{d_i}$ and $\tau_i = \delta_i$. So clearly, $\tau_{(N)} \geq 1$. So $C_0 = 1$. Also, $N/n \to 1 - 1/e$ in probability, so p/N has a limit in probability and this limit is bounded away from 1 because of our assumption that $\rho_n \to \rho \in (0, 1-1/e)$. Finally, we can pick $\eta_0 = 1 - 1/e$.

So the proof of Theorem 4.1 applies (it is easy to see here that the assumption that $\tau_i \neq 0$ can be dispensed of, because we know that the non-zero τ_i's are large enough for our arguments to go through, and there are enough of them that we do not have problems with $\widehat{\Sigma}^{-1}$ not being defined) and we have the announced result. □

The previous theorems settled the question of understanding the impact of the non-parametric bootstrap on statistics of the form $v'\widehat{\Sigma}^{-1}v$ in the situation where the original data were Gaussian. A similar analysis could be carried out in the case of elliptical data, when we assume that the "ellipticity" parameters, λ_i are such Assumption-BB is satisfied for the "new weights" $\tau_i = \lambda_i^2 w_i$. The result would then depend on the limiting distribution of $\lambda_i^2 w_i$ (if it exists), where w_i is the bootstrap weight given to observation i.

4.4.4 Bootstrap analogs of Theorems 4.3 and 4.4

An important piece of our analysis of quadratic programs with linear equality constraints when the data are elliptically distributed was the study of quadratic forms of the type $\widehat{\mu}'\widehat{\Sigma}^{-1}\widehat{\mu}$. It is natural to ask what happens to them when we bootstrap the data. In the elliptical case, we saw that the key statistic was of the form, when $\mu = 0$ and $\mathfrak{S} = \Sigma^{1/2}Y'\Lambda^2 Y\Sigma^{1/2}/n$,

$$\widehat{\mu}'\mathfrak{S}^{-1}\widehat{\mu} = \frac{1}{n}\mathbf{e}'\Lambda Y(Y'\Lambda^2 Y)^{-1}Y'\Lambda\mathbf{e} \; .$$

However, in the bootstrap case, if Λ is the diagonal matrix containing the bootstrap weights, we have $\mathfrak{S}^* = \Sigma^{1/2}Y'\Lambda Y\Sigma^{1/2}/n$, but $\widehat{\mu}^* = Y'\Lambda\mathbf{e}/n$, so the key statistic is going to be of the form

$$(\widehat{\mu}^*)'(\mathfrak{S}^*)^{-1}(\widehat{\mu}^*) = \frac{1}{n}\mathbf{e}'\Lambda Y(Y'\Lambda Y)^{-1}Y'\Lambda\mathbf{e} \; .$$

This creates complications because the matrix $\Lambda Y(Y'\Lambda Y)^{-1}Y'\Lambda$ is not a projection matrix, and hence some of our previous analysis cannot be applied directly. However, this statistic can be rewritten, if we denote $w = \Lambda^{1/2}\mathbf{e}$, as

$$\frac{1}{n}w'\Lambda^{1/2}Y(Y'\Lambda Y)^{-1}Y'\Lambda^{1/2}w = \frac{1}{n}w'P_{\Lambda^{1/2}}w \; ,$$

where $P_{\Lambda^{1/2}}$ is now a projection matrix. As before its off-diagonal elements have mean 0 (conditional on Λ), but now we also need to understand $\sum_{i=1}^n w_i P_{i,i}/n$ and not only $\sum_{i=1}^n P_{i,i}/n$. A detailed analysis of the former quantity is done in Appendix C-3.

We naturally now assumes that p/n has a finite limit, ρ in $(0,1-1/e)$. As explained in Appendix C-3, $\sum_{i=1}^n w_i P_{i,i}/n \to (\mathfrak{s}-1)/\mathfrak{s}$ in probability, with $\mathcal{P}_n$ probability going to 1, where $\mathfrak{s}$ is computed by solving Equation (11) (i.e using Po(1) for G in that equation).

Similarly, it is explained there, that with $\mathcal{P}_n$ probability going to 1, when X_i have mean 0,

$$(\widehat{\mu}^*)'(\widehat{\Sigma}^*)^{-1}\widehat{\mu}^* \to \mathfrak{s}-1 \geq \frac{\rho}{1-\rho}\ , \text{ in } \mathcal{Q}_n \text{ probability.}$$

Finally, an analog of Theorem 4.3 holds, so we have an analog of Theorem 4.4, where $\mathfrak{s}$ is as defined above and $\rho_n/1-\rho_n$ needs to be replaced by $\mathfrak{s}-1$.

In summary, we have the following proposition.

Proposition 4.5. *Call $\mathfrak{s}$ the quantity defined by Equation* (11).

Suppose the data $X_1,\ldots,X_n$ is i.i.d $\mathcal{N}(\mu,\Sigma)$, and call $\mathcal{P}_n$ the corresponding probability distribution. Suppose v is a given deterministic sequence of vectors. We have, when bootstrapping the data, with $\mathcal{P}_n$ probability going to 1:

$$\frac{v'(\widehat{\Sigma}^*)^{-1}v}{v'\Sigma^{-1}v} \to \mathfrak{s} \textit{ in } \mathcal{Q}_n - \textit{probability.}$$

$$\frac{(\widehat{\mu}^*)'(\widehat{\Sigma}^*)^{-1}v}{\sqrt{v'\Sigma^{-1}v}} \to 0 \textit{ in } \mathcal{Q}_n - \textit{probability, when } \mu = 0\ ,$$

$$(\widehat{\mu}^*)'(\widehat{\Sigma}^*)^{-1}\widehat{\mu}^* \simeq \mathfrak{s}\mu'\Sigma^{-1}\mu + (\mathfrak{s}-1) + \mathrm{o}_{\mathcal{Q}_n}\left(\sqrt{\mu'\Sigma^{-1}\mu}, 1\right)$$

We note that our techniques could yield generalizations of the previous fact for the case where the data is elliptically distributed. However, in the case where X_i have mean 0, the quantity $(\widehat{\mu}^*)'(\widehat{\Sigma}^*)^{-1}\widehat{\mu}^*$ does not seem to have a limiting value that is writable in compact form, so we do not dwell on this question further.

Naturally, the motivation behind the previous proposition is practical and the results are interesting from that standpoint. They show that the bootstrap yields inconsistent estimators of the population quantities, something that is not completely unexpected when we understand the random matrix aspects of these questions. Perhaps even more interesting is that bootstrap estimates of bias are themselves inconsistent: as a matter of fact, the key quantity that measures bias in the Gaussian case is $1/(1-p/n)$; when bootstrapping it is replaced by $\mathfrak{s}$, as defined in Equation (11). These results therefore cast some doubts on the practical relevance of the bootstrap for the high-dimensional problems we are considering, at least when it is used in "classical" ways.

4.4.5 On the parametric bootstrap

In the settings considered here, it is also natural to ask how the parametric bootstrap would behave. For instance, if we assumed Gaussianity of the data, we could just estimate Σ and μ (by e.g, naively, $\widehat{\Sigma}$ and $\widehat{\mu}$) and use a parametric bootstrap to get at the quantities we are interested in.

Naturally, the analysis of such a scheme is similar to the analysis of the Gaussian case carried out in Section 3, where the population parameters Σ and μ need to be replaced by the estimators we use in our parametric bootstrap. The same would be true if we were to do a parametric bootstrap for elliptical data, but we would have to use the results of Section 4 instead.

Our computations show that the parametric bootstrap could be used in the problems under study to estimate the bias of various plug-in estimators: we would for instance recover the correct $\mathfrak{s}$ by considering $v'(\Sigma^*_{\text{parametric}})^{-1}v/v'\widehat{\Sigma}^{-1}v$. We note, however, that our analyses, and the estimation work we carry out in Section 5 could do this too, at a cheaper numerical cost.

Finally and very interestingly, we see that a naive use the parametric bootstrap to estimate the bias in the empirical efficient frontier - a reasonable idea at first glance - would yield inconsistent estimates of bias.

5 Robustness, bias, and improved estimation

We now go back to our original problem, which was to understand the relationship between the solution of Problem (QP-eqc-Emp) and the solution of Problem (QP-eqc-Pop) (see page 7 for definitions).

It is naturally important to understand the effect of making the assumption that the data is normally distributed as compared to, say, an assumption of elliptical distribution for the data. The following discussion fleshes out some our theoretical results and what their significance is when solving quadratic programs with linear equality constraints. The discussion is an application of the work done in Sections 2 to 4. It might appear to be mainly heuristic, but precise statements can be easily deduced from the precise statements of the theorems given in the corresponding technical sections.

We discuss here only the case of i.i.d data. As we have shown above, the bootstrap case and the case of correlated observations are more complicated to handle, and the formulas are not as explicit in those cases as they are in the case of i.i.d data. But for certain cases, one could plug-in our earlier results for those situations to obtain explicit results about efficient frontiers and weight vectors in those cases too.

As a matter of notation, all of our approximation statements hold with high-probability asymptotically, unless otherwise noted. We will carry out our work under the model put forward in Theorem 4.1, assuming that the λ_i's are i.i.d and the following assumptions:

1. Assumption A1: for all $i \in \{1, \ldots, k\}$, $v_i'\Sigma^{-1}v_i$ stays bounded away from 0. v_k is assumed to be equal to μ.
2. Assumption A2: the smallest eigenvalue of $M = V'\Sigma^{-1}V$ stays bounded away from 0 and the condition number of M remains bounded.
3. Assumption A3: if $\epsilon = \pm 1$, $(v_i + \epsilon v_j)'\Sigma^{-1}(v_i + \epsilon v_j)$ stays bounded away from infinity.
4. Assumption A4: (Assumption-BB) and (Assumption-BL) hold. (See Theorem 4.4 for definitions.)
5. Assumption A5: The operator norm of Σ, $|||\Sigma|||_2$, remains bounded.

These assumptions guarantee that the noise terms involving $\widehat{\mu}$ do not overwhelm the signal terms involving μ, and also that we can safely take inverses of our approximations to get approximations of their inverses. Also, all the key results we obtained in Sections 3 and 4 are applicable, and our conclusions will of course heavily rely on them.

We will use the notation $\rho_n = p/n$. We recall that in the Gaussian case, the quantity $\mathfrak{s}$ appearing below is approximately equal to $1/(1-\rho_n)$ and in the elliptical case, it is always greater than $1/(1-\rho_n)$, as we explained after the proof of Theorem 4.1.

5.1 Relative positions of efficient frontiers: Gaussian vs. elliptical case

When assumptions (A1-A4) hold, it is clear that

$$\widehat{M} = \widehat{V}'\widehat{\Sigma}^{-1}\widehat{V} \simeq \mathfrak{s}V'\Sigma^{-1}V + \frac{\rho_n}{1-\rho_n}e_k e_k' \; . \tag{12}$$

Now recall that in the elliptical case, $\mathfrak{s} \geq 1/(1-p/n) = \mathfrak{s}^G$, i.e the "$\mathfrak{s}$" corresponding to the Gaussian case. Calling $\widehat{M_E}$ the empirical estimator of M we get in the elliptical case and $\widehat{M_G}$ its analog in the Gaussian case, we have, when A1-A4 are satisfied, with high-probability,

$$\widehat{M_E} \succeq \widehat{M_G} \; ,$$

at least asymptotically.

We now call $f_{\text{emp}}^{(E)}$ and $f_{\text{emp}}^{(G)}$ the "efficient frontiers" obtained by solving Problem (QP-eqc-Emp) when the data is respectively elliptical and Gaussian. Recall that under our assumptions, μ and Σ are the same for the two problems, so the population version corresponding to the two problems is the same. We call the population solution, i.e the efficient frontier computed with the population parameters, f_{theo}. Naturally, this is the quantity we are fundamentally interested in estimating.

Using the fact that $f_{\text{emp}} = U'\widehat{M}^{-1}U$, the following important results.

Theorem 5.1. *When Assumptions A1-A4 are satisfied, we have with high-probability and asymptotically,*

$$f_{\text{emp}}^{(E)} \leq f_{\text{emp}}^{(G)} \leq f_{\text{theo}} \; .$$

In other words, risk underestimation in the empirical quadratic program with linear equality constraints is least severe (within the class of elliptical models) in the Gaussian case.

In other respects, we have, asymptotically, with high-probability, if $\kappa = \rho_n/(1-\rho_n)$,

$$f_{\text{emp}}^{(E)} \simeq \frac{1}{\mathfrak{s}}\left(f_{\text{theo}} - \frac{\kappa}{\mathfrak{s}}\frac{\left(e_k' M^{-1}U\right)^2}{1+\frac{\kappa}{\mathfrak{s}}e_k'M^{-1}e_k}\right) \; . \tag{13}$$

Another way of phrasing this result is the fact that the Gaussian analysis gives the most optimistic view of risk underestimation within the class of elliptical models considered here.

Practically, it means that users of Markowitz-type optimization should be wary of the empirical solution they get, and even of the correction that Gaussian results suggest. If the data is elliptical, they will underestimate the risk of their portfolio even more than the Gaussian results suggest.

Let us now give a proof of Theorem 5.1.

Proof. Under the assumptions of the Theorem, we can use the approximation in Equation (12). The first part of the theorem has been argued before, so we do not need to do anything else to obtain it.

The second part follows directly from a rank one perturbation argument. We have

$$f_{\text{emp}}^{(E)} \simeq U'\left(\mathfrak{s}V'\Sigma^{-1}V + \frac{\rho_n}{1-\rho_n}e_k e_k'\right)^{-1} U = \frac{1}{\mathfrak{s}}U'\left(M + \frac{\kappa}{\mathfrak{s}}e_k e_k'\right)^{-1} U \; .$$

Using the classic result $(M+\nu\nu')^{-1} = M^{-1} - M^{-1}\nu\nu' M^{-1}/(1+\nu' M^{-1}\nu)$, we conclude that

$$U'\left(M + \frac{\kappa}{\mathfrak{s}}e_k e_k'\right)^{-1} U = U'M^{-1}U - \frac{\kappa}{\mathfrak{s}}\frac{(U'M^{-1}e_k)^2}{1+\frac{\kappa}{\mathfrak{s}}e_k'M^{-1}e_k} \; .$$

We now recall from Section 2 that $f_{\text{theo}} = U'M^{-1}U$, and we have the announced result. □

Equation (13) naturally suggests better ways of estimating f_{theo} than using f_{emp}. We postpone a discussion of this issue to Subsection 5.4 , because it requires somewhat lengthy preliminaries.

5.2 Issues concerning the weights of the portfolio

Beside problems in the location of the efficient frontiers, our analysis reveals another very interesting phenomenon: problems with estimating w_{theo}, the optimal vector of weights. In particular, one can show that the mean return of the portfolio is poorly estimated and the weight given to each asset is biased.

Theorem 5.2 (Bias in weights)**.** *Suppose assumptions A1-A4 hold. We have, asymptotically and with high-probability,*

$$w_{\text{emp}} \simeq w_{\text{theo}} - \zeta(\mathfrak{s})\frac{\kappa}{\mathfrak{s}}w_b \; , \tag{14}$$

where

$$\zeta(\mathfrak{s}) = \frac{e_k'M^{-1}U}{1+\frac{\kappa}{\mathfrak{s}}e_k'M^{-1}e_k} \; , \; w_b = \Sigma^{-1}VM^{-1}e_k \; .$$

This approximation is valid when looking at linear combinations of the vector of weights: if $\gamma \in \mathbb{R}^n$ *is deterministic and assumption A3 extended to include this vector holds,*

$$\gamma' w_{\text{emp}} \sim \gamma'\left(w_{\text{theo}} - \zeta(\mathfrak{s})\frac{\kappa}{\mathfrak{s}}w_b\right) \; .$$

We note that the last assertion of the theorem does not necessarily immediately follow from equation (14) in high-dimension, but it is true in the setting we consider. A particularly interesting corollary is the following statement concerning inconsistent estimation of the returns.

Corollary 5.1 (Poor estimation of returns). *Recall that with our notations, $w'_{\text{theo}}\mu = u_k = \mu_P$. In practical terms, μ_P corresponds to the desired expected returns we wish to have for our "portfolio". Under the same assumptions as that of Theorem 5.2, we have*

$$\mu' w_{\text{emp}} \simeq \mu_P \frac{1}{1+\frac{\kappa}{\mathfrak{s}} e'_k M^{-1} e_k} - \frac{\kappa}{\mathfrak{s}} \frac{\sum_{i<k} u_i e'_k M^{-1} e_i}{1+\frac{\kappa}{\mathfrak{s}} e'_k M^{-1} e_k} \; .$$

The previous corollary is a statement about poor estimation of returns for the following reason: $\widehat{\mu}' w_{\text{emp}} = \mu_P$ by construction, so one might naively hope that, for a new observation X_{n+1}, independent of $X_1, \ldots, X_n$ and with the same distribution as them, $\mathbf{E}\left(w'_{\text{emp}} X_{n+1} | X_1, \ldots, X_n\right) = w'_{\text{emp}}\mu \simeq \mu_P$. However, as the previous corollary shows, this is not satisfied. We note that the factor affecting μ_P is a shrinkage factor, always smaller than 1 because M is positive semi-definite. The other term could have either sign, so its effect on return estimation is less interpretable. For large μ_P, it is nonetheless clear that the previous corollary shows that the returns are overestimated: the realized returns are (asymptotically and with high-probability) less than μ_P.

We now prove these two results. The proof of the corollary is at the end of the proof of the theorem.

Proof of Theorem 5.2: Under the assumptions of the theorem we have

$$\widehat{M} \simeq \mathfrak{s} M + \kappa e_k e'_k \; ,$$

and our assumptions guarantee that we can take inverses and still have valid approximations. Hence, using the classic formula for inversion of a rank one perturbation of a matrix (see Horn and Johnson (1990), p. 19), we have

$$\widehat{M}^{-1} \simeq \frac{1}{\mathfrak{s}} \left(M^{-1} - \frac{\kappa}{\mathfrak{s}} \frac{M^{-1} e_k e'_k M^{-1}}{1+\frac{\kappa}{\mathfrak{s}} e'_k M^{-1} e_k} \right) \; .$$

Now recall that $w_{\text{emp}} = \widehat{\Sigma}^{-1} \widehat{V} \widehat{M}^{-1} U$ and $w_{\text{theo}} = \Sigma^{-1} V M^{-1} U$. For a deterministic γ, our work in Section 4 indicates that $\gamma' \widehat{\Sigma}^{-1} \widehat{V} \sim \mathfrak{s} \gamma' \Sigma^{-1} V$. So we conclude that

$$\gamma' w_{\text{emp}} \sim \mathfrak{s} \gamma' \Sigma^{-1} V \frac{1}{\mathfrak{s}} \left(M^{-1} - \frac{\kappa}{\mathfrak{s}} \frac{M^{-1} e_k e'_k M^{-1}}{1+\frac{\kappa}{\mathfrak{s}} e'_k M^{-1} e_k} \right) U \; .$$

In other words, we have

$$\gamma' w_{\text{emp}} \sim \gamma' \Sigma^{-1} V M^{-1} U - \frac{\kappa}{\mathfrak{s}} \frac{\gamma' \Sigma^{-1} V M^{-1} e_k e'_k M^{-1} U}{1+\frac{\kappa}{\mathfrak{s}} e'_k M^{-1} e_k} \; ,$$

or, as announced,

$$\gamma' w_{\text{emp}} \sim \gamma' w_{\text{theo}} - \frac{\kappa}{\mathfrak{s}} \gamma' w_b \, \zeta(\mathfrak{s}) \; .$$

It seems difficult to say more, because w_b and ζ are population parameters and their properties and values may vary from problem to problem.

- **Proof of the corollary** We now assume that $\gamma = \mu$. We remark that $\mu = V e_k$, by construction of V. Therefore,

$$\mu' w_b = e'_k V' \Sigma^{-1} V M^{-1} e_k = e'_k M M^{-1} e_k = 1 \; .$$

Further,

$$e'_k M^{-1} U = \sum_{i=1}^{k} u_i e'_k M^{-1} e_i = \sum_{i<k} u_i e'_k M^{-1} e_i + \mu_P e'_k M^{-1} e_k \; .$$

These two remarks and the result of Theorem 5.2 give the conclusion of the corollary. □

5.3 Bias correction for the weights

An important question now that we have identified possible problems with the empirical weights is to try and correct them. We propose such a scheme, suggested by our computations.

Our investigations will rely on the following asymptotic result, discussed in Theorem 5.2: in the notations of this theorem,

$$\gamma' w_{\text{emp}} \sim \gamma' w_{\text{theo}} - \frac{\kappa}{\mathfrak{s}} \gamma' w_b \, \zeta(\mathfrak{s}) \ .$$

Our efforts will focus on trying to estimate $w_b/\mathfrak{s}$ and $\zeta(\mathfrak{s})$, as $\kappa = \rho_n/(1-\rho_n)$ is known and computable from the data.

Recall that we assumed that $v_k = \mu$ and let us call

$$\widetilde{M} = \widehat{M} - \kappa e_k e_k' \ .$$

Under the assumptions underlying the previous computations, we have

$$\widetilde{M} \sim \mathfrak{s} M \ .$$

In practice, we wish $\widetilde{M}$ to be a positive semi-definite matrix - something that is guaranteed asymptotically, but might require checking and potentially corrections in practice.

We propose to use

1. As an estimator of w_b,
$$\widehat{w_b} = \widehat{\Sigma}^{-1} \widehat{V} \widetilde{M}^{-1} e_k$$

2. As an estimator of $\zeta(\mathfrak{s})/\mathfrak{s}$,
$$\widehat{z} = \frac{e_k' \widetilde{M}^{-1} U}{1 + \kappa e_k' \widetilde{M}^{-1} e_k} \ .$$

For any deterministic γ (such that the assumptions of Theorem 5.2 hold), $\gamma' \widehat{w_b} \sim \gamma' w$, because $\gamma' \widehat{\Sigma}^{-1} \widehat{V} \sim \mathfrak{s} \gamma' \Sigma^{-1} V$ and $\widetilde{M}^{-1} U \sim M^{-1} U/\mathfrak{s}$. Also, $e_k' \widetilde{M}^{-1} U \sim \mathfrak{s}^{-1} e_k' M^{-1} U$, and $e_k' \widetilde{M}^{-1} e_k \sim \mathfrak{s}^{-1} e_k' M^{-1} e_k$, so $\widehat{z} \sim \zeta(\mathfrak{s})/\mathfrak{s}$. Hence,

$$\gamma' \widehat{w_b} \widehat{z} \sim \gamma' w \frac{\zeta(\mathfrak{s})}{\mathfrak{s}} \ .$$

In other words, we have found an asymptotically consistent way of estimating the quantities of interest. Hence, the estimator we propose to use is

$$\boxed{\widehat{w_{\text{theo}}} = (w_{\text{emp}} + \kappa \widehat{z} \widehat{w_b}) = \widehat{\Sigma}^{-1} \widehat{V} \widetilde{M}^{-1} U} \ . \tag{15}$$

Interestingly, this proposal does not require us to estimate $\mathfrak{s}$. Furthermore, because we have consistency of the estimator in the whole class of elliptical distributions, this estimator is fairly robust to distributional assumptions about the data. Finally, the estimator is consistent in the sense that all (deterministic) linear combinations of $\widehat{w_{\text{theo}}}$ are consistent for the corresponding linear combinations of w_{theo} (provided these linear combinations are such that the assumptions of Theorem 5.2 apply to them).

The estimator satisfies the constraints It is nonetheless natural to raise the following question: does the proposed estimator satisfy the constraints of the problem? If not, our proposal would be problematic, but it is indeed the case that our estimator satisfies the constraints $\widehat{w_{\text{theo}}}' v_i = u_i$ for all $i \in \{1, \ldots, k-1\}$. Naturally, the last constraint (i.e $\widehat{w_{\text{theo}}}' \mu = u_k = \mu_P$) is difficult to satisfy exactly because μ is unknown, so it is also less of a concern.

Let us now briefly justify our claim concerning the satisfaction of the equality constraints. By construction, w_{emp} satisfies the constraints $w_{\text{emp}}' v_i = u_i$, $1 \leq i \leq k-1$, so all we have to show is that the $k \times 1$ vector $\widehat{V}' \widehat{w_b}$ is proportional to e_k. We recall that $\widetilde{M} = \widehat{M} - \kappa e_k e_k'$, so

$$\widehat{w_b} = \widehat{\Sigma}^{-1} \widehat{V} \left(\widehat{M} - \kappa e_k e_k' \right)^{-1} e_k \ .$$

Using the standard formula for the inverse of a rank-1 perturbation of a matrix, we therefore get

$$\begin{aligned}\widehat{w}_b &= \widehat{\Sigma}^{-1}\widehat{V}\left(\widehat{M}^{-1}+\kappa\frac{\widehat{M}^{-1}e_k e_k'\widehat{M}^{-1}}{1-\kappa e_k'\widehat{M}^{-1}e_k}\right)e_k\\ &= \widehat{\Sigma}^{-1}\widehat{V}\widehat{M}^{-1}e_k+\kappa\widehat{\Sigma}^{-1}\widehat{V}\widehat{M}^{-1}e_k\frac{e_k'\widehat{M}^{-1}e_k}{1-\kappa e_k'\widehat{M}^{-1}e_k}\\ &= \frac{1}{1-\kappa e_k'\widehat{M}^{-1}e_k}\widehat{\Sigma}^{-1}\widehat{V}\widehat{M}^{-1}e_k\end{aligned}$$

Once we recall that $\widehat{M}=\widehat{V}'\widehat{\Sigma}^{-1}\widehat{V}$, we immediately get the equality

$$\widehat{V}'\widehat{w}_b=\frac{1}{1-\kappa e_k'\widehat{M}^{-1}e_k}e_k \ ,$$

which shows that $v_i'\widehat{w}_b=0$ for $1\leq i\leq k-1$, as announced.

Finally, from a practical point of view, one might be worried that the estimator proposed in Equation (15) "puts too much weight on the theory and not enough on the data", and that better practical performance might be achieved by tuning more finely our corrections to the data. For instance, one might propose, we think reasonably, to use, instead of $\widetilde{M}$ the matrix $\widetilde{M}(\lambda_1)=\widehat{M}-\lambda_1\kappa e_k e_k'$, where λ_1 would be picked by some form of cross-validation based on the new estimator $\widehat{w_{\text{theo}}}(\lambda_1)=w_{\text{emp}}+\kappa\widehat{z}(\lambda_1)\widehat{w}_b(\lambda_1)$. We do not discuss this issue any further in this paper as we plan to address it in another, more applied, article. We do however show the performance of our estimator in simulations in Subsection 5.5.

5.4 Improved estimation of the frontier

We now discuss the question of improved estimation of the efficient frontier. This is naturally an important quantity in the problem, and, as we hope to have shown, a difficult one to estimate by naive methods. One aspect of its importance is that it gives us a benchmark of performance for optimal portfolios. We therefore think that in a financial context, it might be of great interest in particular to regulators.

5.4.1 Estimation of $\mathfrak{s}$

Though we have seen that we could devise a scheme to improve the estimation of the weights without having to estimate $\mathfrak{s}$, this latter quantity is still an important one to estimate if we want to better understand the pitfalls we might be facing.

In the elliptical case, where $X_i=\mu+\lambda_i\Sigma^{1/2}Y_i$, we wish to estimate λ_i^2, as we have seen that $\mathfrak{s}$ is "driven" by this quantity. To do so, we recall the concentration of measure results put forward in El Karoui (2009), which say that with very high probability, if the largest eigenvalue of Σ stays bounded,

$$\frac{\|\Sigma^{1/2}Y_i\|_2^2}{p}\simeq\frac{\operatorname{trace}(\Sigma)}{p} \ .$$

Now, note that $\|\mu-\widehat{\mu}\|_2^2\simeq\frac{\operatorname{trace}(\Sigma)}{n}$, because under our assumptions A1-A4 and the assumption of independence of the λ_i's, $\sum_{i=1}^n\lambda_i^2/n\to 1$ and A5 implies that the previous approximation holds. Hence,

$$\frac{\|X_i-\widehat{\mu}\|_2^2}{p}\simeq\lambda_i^2\frac{\operatorname{trace}(\Sigma)}{p} \ .$$

We now propose the following estimator for λ_i^2 :

$$\widehat{\lambda_i^2}=\frac{\|X_i-\widehat{\mu}\|_2^2}{\sum_{i=1}^n\|X_i-\widehat{\mu}\|_2^2/n} \ .$$

If we denote $\rho_n = p/n$, we then propose to estimate $\mathfrak{s}$ using the positive solution of

$$g(x) = 1 - \rho_n \,,$$
$$\text{where } g(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 + x\widehat{\lambda_i^2}\rho_n} \,.$$

We note that this is just the discretized version of the equation characterizing $\mathfrak{s}$. (g is clearly a continuous convex decreasing function of x on $[0,\infty)$, so the existence and uniqueness of a solution to $g(x) = 1 - \rho_n$ is clear.)

5.4.2 Estimation of the efficient frontier

We recall an important result from Theorem 5.1: under the assumptions made in this section,

$$f_{\text{emp}} \simeq \frac{1}{\mathfrak{s}} f_{\text{theo}} - \kappa \frac{\left(e_k' M^{-1} U/\mathfrak{s}\right)^2}{1 + \frac{\kappa}{\mathfrak{s}} e_k' M^{-1} e_k} \,.$$

Now recall that we have a consistent estimator of $e_k' M^{-1}/\mathfrak{s}$, that is $e_k'\widetilde{M}^{-1}$, and we just discussed how to estimate $\mathfrak{s}$.

As an estimator of the efficient frontier we therefore propose

$$\boxed{\widehat{f_{\text{theo}}} = \widehat{\mathfrak{s}}\left(f_{\text{emp}} + \kappa \frac{(e_k'\widetilde{M}^{-1}U)^2}{1 + \kappa e_k'\widetilde{M}^{-1} e_k}\right)} \,.$$

We also note that $\widetilde{M}$ could be replaced by $\widetilde{M}(\lambda_1)$ described above with a similar cross-validation scheme.

5.5 Numerical results and practical considerations

This subsection gives some numerical results to assess the quality of the proposed estimators for both weights and "efficient frontier".

Our aim was to investigate among other things the improvement in the quality of our approximations as n and p grew to infinity. Hence, we present the results of two simulation setups: one where $n = 250, p = 100$ and one where $n = 2500, p = 1000$. We chose to work with simulations where we picked both Σ and μ so that we could guarantee - for instance - that the efficient frontier was basically the same for both simulations.

More specifically, we chose Σ to be a $p \times p$ Toeplitz matrix, with $\Sigma(i,j) = \alpha^{|i-j|}$, where $\alpha = .4$. In the smaller dimensional simulation, i.e $p = 100$, we picked v_1 to be the eigenvector associated with the 90th smallest eigenvalue of Σ. Calling β_2 the eigenvector associated with the 15th smallest eigenvalue of Σ, we picked $v_2 = \mu$ to be $\sqrt{.3}v_1 + \sqrt{.7}\beta_2$. In the larger dimensional simulation, we used for v_1 the eigenvector associated with the 900th smallest eigenvalue of Σ, while β_2 was now associated with the 150th smallest eigenvalue of Σ. $\mu = v_2$ was computed in the same fashion in both simulations.

We did simulations both in the Gaussian case and in the case of an elliptical distribution as described above, i.e $X_i = \mu + \lambda_i \Sigma^{1/2} Z_i$, where λ_i was proportional to a t-distributed random variables with 6 degrees of freedom and scaled to have variance 1. We picked 6 degrees of freedom to have simulations with relatively heavy tails and capture visually the corresponding effects. It was also naturally a way to investigate the practical robustness of our estimators and compare with the Gaussian case. We call below the set of simulations involving the t-distribution the "t_6" case because of its similarity with multivariate t-distributions.

We repeated 1000 times the simulations in all the cases considered. We chose $u_1 = 1$ and u_2 (the "target returns" in a financial context) ranging from .1 to 5.

We note that our estimators require taking inverses of matrices - which naturally raises the question of how well-conditioned those matrices are. This is in particular the case when we deal with M and $\widetilde{M}$:

if M is poorly conditioned, even though $\widetilde{M}$ is a good estimator of $\mathfrak{s}M$, it can turn out that $\widetilde{M}^{-1}$ is a relatively poor estimator of $M^{-1}\mathfrak{s}^{-1}$. In our simulations, both M and Σ were well-conditioned but in practice, one should be aware of potential difficulties that may arise if, for instance, $\widetilde{M}$ indicates that M may be ill-conditioned.

5.5.1 Estimation of portfolio weights

As we have seen earlier, the “naive” weights obtained by plugging-in the sample mean and the sample covariance matrix in our quadratic program with linear equality constraints are biased, in the sense that their projection in any given direction will generally be biased.

Here we show the performance of our estimator as measured by its projection on $v_k = \mu$. It is a natural direction to consider since, for instance in a financial context and under our modeling assumptions, it gives us the expected returns of our portfolio (conditional on $X_1, \ldots, X_n$).

As our simulations indicate, our estimator appears to be practically unbiased (even in the “lower-dimensional” case), which means in a financial context that the corresponding investment strategy will yield the returns that the investor expected. (We note that from a mean-variance point of view, we do not claim that our estimator is optimal. Work is under way to find better performing portfolios - but it requires a new set of theoretical investigations whose results are postponed to another paper. In limited simulations, it appeared that our “debiased” portfolio performed similarly to the naive one from a mean-variance point of view, its main advantage being that it delivers the returns that the investor expects.)

We present two pictures on (pp. 40 and 41) to give a sense to the reader of the impact of the size of n and p on the estimators we proposed (the “larger-dimensional” case gives quite significantly better results, with narrower confidence bands, though (empirical) near-unbiasedness is present in both cases).

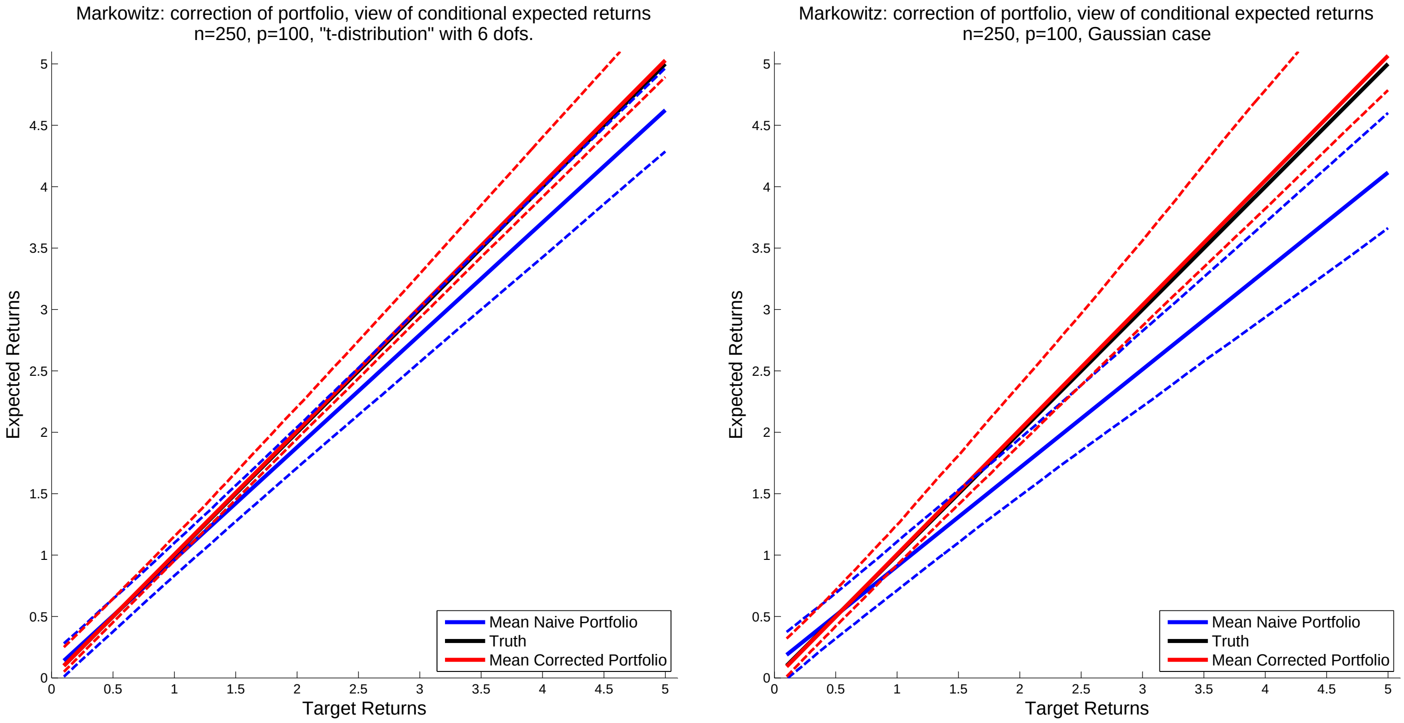


Figure 2: Performance of naive and corrected portfolios, for scaled “t_6” (left picture) and Gaussian returns. Here $n = 250, p = 100$ and the number of simulations is 1000. The dashed lines represent 95% confidence bands. The x-axis represents the returns an investor expects. The y-axis represents what s/he would actually get on average (i.e $\mu'\widehat{w}$). The plots show both the bias in the naive solution (blue solid lines) and the fact that our estimator is nearly unbiased (red solid lines). They also illustrate the robustness of our corrections. The black line is very close to the red line, showing a very good correction (on average).

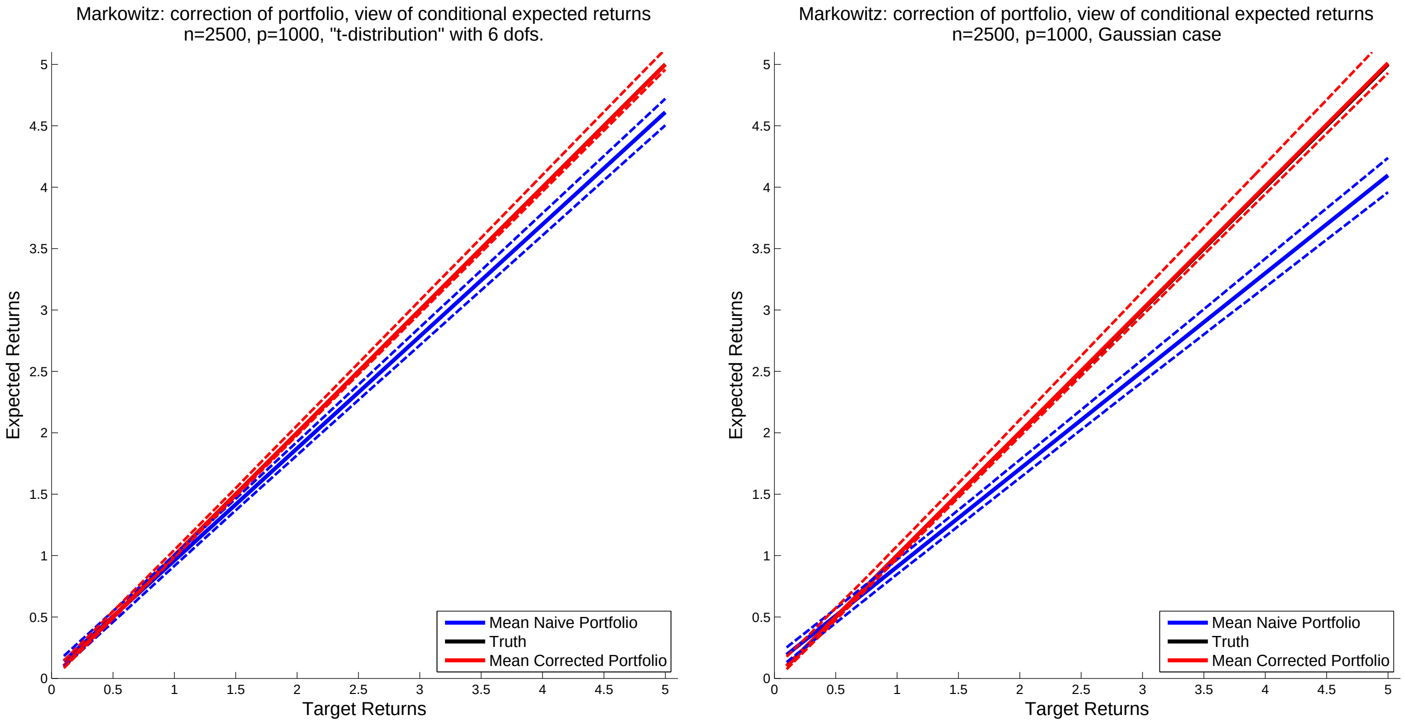


Figure 3: Performance of naive and corrected portfolios, for scaled "t_6" (left picture) and Gaussian returns. Here $n = 2500, p = 1000$ and the number of simulations is 1000. The dashed lines represent 95% confidence bands. The x-axis represents the returns an investor expects. The y-axis represents what s/he would actually get on average (i.e $\mu'\widehat{w}$). The plots show both the bias in the naive solution (blue solid lines) and the fact that our estimator is nearly unbiased (red solid lines). They also illustrate the robustness of our corrections. Note the narrower confidence bands as compared to Figure 2. The black line is essentially hidden under the red line, showing a near perfect correction (on average).

5.5.2 Correction to the frontier

We now turn to the issue of estimating the "efficient frontier", i.e the curve that represents the minima of our convex optimization problem (QP-eqc), on p. 5. The pictures we present (see p. 51) were obtained from the simulations we described above. We chose to plot the variance (i.e $\min w'\Sigma w$) on the x-axis and the target returns (i.e the u_k's in the notation of Equation (QP-eqc)) on the y-axis as this is the convention in financial applications.

As the reader can see, our estimator turns out to be essentially unbiased, even in the "lower-dimensional" case. We note too that the variance can be quite large but that the confidence bands obtained from our corrections were always to the right of the confidence bands obtained from the naive estimator - meaning, if one is concerned with risk estimation that in (essentially) the worst case for our estimator, we still obtained a better performing estimator than in (essentially) the best case for the naive estimator.

Finally, for graphical purposes and to help comparisons, we chose to put all the graphs on the same scale. Some of the information on our original graphs (for the "lower-dimensional" case) was therefore left out but can be inferred by "naturally" extrapolating the curves shown on our graphs which are essentially parabolas.

6 Conclusion

This study of quadratic programs with linear equality constraints whose parameters are estimated from data has highlighted the difficulties created by the high-dimensionality of the data. In particular, we have shown that the fact that n (the number of observations used to estimate the parameters) and p both grew to infinity lead to a systematic underestimation of the minimal "risk" one exposed itself to when approaching the optimization problem (QP-eqc-Pop) by solving its proxy (QP-eqc-Emp).

Our study produced exact distributional results in the Gaussian case (Section 3) and convergence results in probability in the elliptical case (Section 4), which also allowed us to reach conclusions for the bootstrap and the case of non-independent data (in particular, it covers the case of Gaussian data correlated in time). As explained in Section 5, the study of the Gaussian case gives an over-optimistic assessment of risk underestimation in the context we study: in the class of elliptical distributions we consider, risk is minimally underestimated in the Gaussian case, and the situation is more dire for other elliptical distributions. Our study also highlights the fact that standard bootstrap estimates of bias will be inconsistent. It also suggests that in the case of correlated Gaussian observations, risk underestimation is likely to be more severe than in the i.i.d case.

Another benefit of our analysis is that it sheds light on what is creating those difficulties and allows us to propose robust corrections to these problems. As shown in the theoretical part of the paper and illustrated in our simulation work, they are robust in the class of elliptical distributions we consider. They also appear to work quite well in practice - as our (somewhat limited) simulation work seems to indicate.

Perhaps surprisingly, we did not need to make very strong assumptions about the covariance matrix at stake or its mean, whereas recent statistical work focused on estimation of covariance matrices (see El Karoui (2008a) or Bickel and Levina (2007b)) tends to do so. This is in part because our theoretical analysis clearly showed what functionals of these two parameters one needed to estimate, and hence we were able to bypass stronger requirements by focusing on those particular functionals.

Beside the interesting statistical and mathematical questions this study raised, we hope that it might also be helpful to, for instance, financial regulators by perhaps providing them with more realistic benchmarks for the performance of optimal portfolios and that it sheds light on how the high-dimensionality of the data affects the proper assessment of risk of large portfolios obtained by solving high-dimensional optimization problems.

APPENDIX

A Classical results of linear algebra

A-1 On inverses of partitioned matrices

In our study of the Gaussian case, and in particular in connection with properties of Wishart matrices, we relied several times on properties of the inverse of a partitioned matrix. Here is a detailed statement of what we needed.

Let A be a generic matrix, and let us decompose it by blocks:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

Let us call A^{-1} the inverse of A. We assume that all inverses we take are well-defined. Let us write

$$A^{-1} = \begin{pmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{pmatrix}$$

Then, it is well known that (see e.g Mardia et al. (1979), pp. 458-459, or Boyd and Vandenberghe (2004), p. 650)

$$A^{11} = \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1} \ , \tag{A-1}$$

$$A^{22} = \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1} \ , \tag{A-2}$$

$$A^{12} = -A_{11}^{-1}A_{12}A^{22} \ , \tag{A-3}$$

$$A^{21} = -A^{22}A_{21}A_{11}^{-1} \ . \tag{A-4}$$

B Random matrix results

B-1 Lower bounds on smallest eigenvalue

In many proofs in the course of the paper we needed to have quantitative bounds on the behavior of the smallest eigenvalue of a number of matrices and made repeated use of the following lemma.

Lemma B-1. *Suppose Y is a $n \times p$ matrix, with i.i.d $\mathcal{N}(0,1)$ entries, with $p/n \to \rho$, and $0 < \rho < 1$.*

Suppose Λ is an $n \times n$ diagonal and deterministic matrix and that we can find $N(n)$, $C > 0$ and $\epsilon > 0$ such that, if τ_i is the i-th largest eigenvalue of $\Lambda'\Lambda$, $\tau_{N(n)} > C$, for some fixed $C > 0$, while N is such that, for p and n large, $p/N < 1 - \epsilon$ and N/n stays bounded away from 0. Finally, we assume that all the diagonal entries of Λ are different from 0.

Call $H = \mathrm{Id} - \delta\delta'/n$, where $\|\delta\|_2^2 = n$. Then λ_p, the smallest eigenvalue of $Y'\Lambda' H \Lambda Y/n-1$, is bounded away from 0 with high-probability.

In particular, when $p/N < 1 - \epsilon$, if $\mathfrak{C}_n = C\frac{N-1}{n-1}$,

$$P\left(\sqrt{\lambda_p} \leq \sqrt{\mathfrak{C}_n}\left[(1-\sqrt{1-\epsilon}) - t\right]\right) \leq \exp\left(-(N-1)t^2\right) \ .$$

The following proof makes clear that the result holds also when some of the diagonal entries of Λ are equal to zero if we make the following modification: n should now denote the number of non-zero entries on the diagonal of Λ and the corresponding assumptions about p and N should then hold. We also point out that under our assumptions H is an orthogonal projection matrix.

Proof. Before we start the proof per se, we need some notations: we call λ_k the k-th largest eigenvalue of a symmetric matrix. In other words, the eigenvalues are decreasingly ordered and $\lambda_1 \geq \lambda_2 \geq \ldots$

The result is known if $\Lambda = \mathrm{Id}_n$, since

$$\frac{1}{n-1} Y' H Y \overset{\mathcal{L}}{=} \frac{1}{n-1} \mathcal{W}_p(\mathrm{Id}_p, n-1) \ ,$$

and therefore (see e.g Bai (1999))

$$\frac{1}{n-1} Y' H Y \to (1-\sqrt{\rho})^2 \quad \text{a.s} \quad ,$$

which is bounded away from 0 - this is a weak form of our statement. Using Davidson and Szarek (2001), Theorem II.13, we have the following stronger and more quantitative result: the smallest eigenvalue of a matrix with distribution $\mathcal{W}(\mathrm{Id}_p, n_0)/n_0$ is strongly concentrated around $(1-\sqrt{p/n_0})^2$ when $p < n_0$, and

$$P\left(\sqrt{\lambda_p} < (1-\sqrt{p/n_0}) - t\right) \leq \exp(-n_0 t^2) \ .$$

This gives our result in the case where $\Lambda = \mathrm{Id}_n$. Let us now investigate what happens when Λ is not Id_n.

The matrix $M = \Lambda' H \Lambda$ is a rank-1 perturbation of $\Lambda'\Lambda$ and is positive semi-definite, because H is. Therefore, for any $k \geq 2$, $\lambda_{k-1}(\Lambda' H \Lambda) = \lambda_{k-1}(M) \geq \lambda_k(\Lambda'\Lambda)$, by the interlacing Theorem 4.3.4 in Horn and Johnson (1990). M has rank $n-1$ matrix since, $M\Lambda^{-1}\delta = 0$ and $\mathrm{rank}\,(M) \geq \mathrm{rank}\,(\Lambda'\Lambda) - 1 = n-1$.

We can diagonalize $M = ODO'$, where D has $(n-1)$ non-zero coefficients, and because $O'Y \overset{\mathcal{L}}{=} Y$, we have

$$Y'MY = Y'\Lambda' H \Lambda Y \overset{\mathcal{L}}{=} Y'DY = \sum_{i=1}^{n-1} d_i Y_i Y_i' \ ,$$

where d_i are the non-zero diagonal entries of D. Because M is positive semi-definite, we have $d_i \geq 0$ for all i. In other respects, because for all $k \leq n-1$, $d_k \geq \lambda_{k+1}(\Lambda'\Lambda) = \tau_{k+1}$ by our remark on interlacing inequalities, we have, if $\succeq$ denotes positive-semidefinite ordering,

$$\sum_{i=1}^{n-1} d_i Y_i Y_i' \succeq \sum_{i=1}^{N-1} d_i Y_i Y_i' \succeq \tau_N \sum_{i=1}^{N-1} Y_i Y_i' = \tau_N \mathcal{W}_p(\mathrm{Id}_p, N-1) \ .$$

Therefore,

$$\frac{1}{n-1}Y'\Lambda' H\Lambda Y \succeq C\frac{N-1}{n-1}\,\frac{1}{N-1}\mathcal{W}_p(\mathrm{Id}_p, N-1)$$

As we recalled above, the smallest eigenvalue of $\mathcal{W}_p(\mathrm{Id}_p, N-1)/(N-1)$ remains bounded away from 0 with high-probability in our setting, because p/N remains bounded away from 1 by assumption. We also assumed that N/n and C were bounded away from 0. If we call $\mathfrak{C} = \liminf_{n\to\infty} C\frac{N-1}{n-1}$, we have $\mathfrak{C} > 0$ and, for any $\eta > 0$,

$$\lambda_p\left(\frac{1}{n-1}Y'\Lambda' H\Lambda Y\right) \geq \mathfrak{C}\left(1-\sqrt{p/N}\right)^2 - \eta$$

with high-probability, when n and p are large enough.

More specifically, according to the result of Davidson and Szarek (2001) we have, for $\lambda_p = \lambda_p\left(\frac{1}{n-1}Y'\Lambda' H\Lambda Y\right)$ and $\mathfrak{C}_n = C(N-1)/(n-1)$,

$$P\left(\sqrt{\lambda_p} \leq \sqrt{\mathfrak{C}_n}\left[(1-\sqrt{p/(N-1)})-t\right]\right) \leq \exp\left(-(N-1)t^2\right)\ .$$

In particular, when p/N is such that $p/N \leq 1-\epsilon$,

$$P\left(\sqrt{\lambda_p} \leq \sqrt{\mathfrak{C}_n}\left[(1-\sqrt{1-\epsilon})-t\right]\right) \leq \exp\left(-(N-1)t^2\right)\ .$$

Interestingly, this bound is "quite uniform" in Λ, in the sense that the only characteristics of Λ that matter are $\mathfrak{C}_n = C\frac{N-1}{n-1}$ and N. □

C Generalizations of the proof of Theorem 4.2

This part of the Appendix explains how to appropriately modify the proofs of Theorems 4.1 and Theorems 4.4 to obtain the results we need in the case of correlated observations (Subsection 4.3) and the bootstrap.

C-1 On $v'\widehat{\Sigma}^{-1}v$ when the observations are correlated

We explain in this subsection how to modify the proof of Theorem 4.1 in the case where the vectors of observations X_i and X_j are potentially correlated. The data was assumed to have the following representation, in matrix form:

$$X = \mathbf{e}\mu' + \Lambda Y\Sigma^{1/2}\ ,$$

where Λ is $n\times n$, deterministic but not necessarily diagonal and Y has i.i.d $\mathcal{N}(0,1)$ entries. We also wrote the SVD of Λ as $\Lambda = ADB'$, where A and B are orthogonal.

If we call $H = \mathrm{Id}_n - \mathbf{e}\mathbf{e}'/n$, we have, of course,

$$\widehat{\Sigma} = \frac{1}{n-1}X'HX = \frac{1}{n-1}\Sigma^{1/2}Y'\Lambda' H\Lambda Y\Sigma^{1/2}\ .$$

The orthogonality of B implies $BY \overset{\mathcal{L}}{=} Y$, and we have

$$\widehat{\Sigma} \overset{\mathcal{L}}{=} \frac{1}{n-1}\Sigma^{1/2}Y'D(A'HA)DY\Sigma^{1/2}\ .$$

If we now call $\delta = A'\mathbf{e}$, we see that $\|\delta\|_2^2 = n$, because A is orthogonal. It can also easily be seen that $A'HA = \mathrm{Id}_n - \delta\delta'/n = H_\delta$. Because of the remark we just made on the norm of δ, H_δ is clearly an orthogonal projection matrix. So we have to understand

$$\widehat{\Sigma} \overset{\mathcal{L}}{=} \frac{1}{n-1}\Sigma^{1/2}Y'DH_\delta DY\Sigma^{1/2}\ ,$$

which is extremely close to the situation of Theorem 4.1, where we had to work with

$$\widehat{\Sigma} \overset{\mathcal{L}}{=} \frac{1}{n-1}\Sigma^{1/2}Y'DH_{\mathbf{e}}DY\Sigma^{1/2}\ .$$

D now plays the role Λ played in Theorem 4.1 and the main modification is that $H = H_{\mathbf{e}}$ is now replaced by H_δ.

An examination of the proof of Theorem 4.1 shows that we never relied on the fact that we used specifically $H_{\mathbf{e}}$ (instead of H_δ) in that proof. All we used was the fact that our H there was a rank-1 perturbation of Id_n and an orthogonal projection matrix. Similarly, Lemma B-1, on which we relied in the course of the proof of Theorem 4.1, handles H_δ for general δ with squared norm n without any problems, so it is still usable in the course of the current study.

Because we know that the singular values of Λ satisfy (Assumption-BB), the proof of Theorem 4.1 goes through without further modifications and Proposition 4.1 holds.

C-2 On quadratic forms involving random projection matrices

A recurrent issue in the questions we addressed was the understanding of statistics of the form

$$\frac{1}{n}u'Pu\ ,$$

where P is a random projection matrix and u a (generally deterministic) vector of dimension n. In particular, the projection matrices we dealt with were of the form

$$P = \Lambda Y(Y'\Lambda^2 Y)^{-1}Y'\Lambda\ ,$$

for Λ a (possibly random) $n \times n$ diagonal matrix and Y an $n \times p$ matrix with i.i.d $\mathcal{N}(0,1)$ entries. We also assume that Λ and u are independent of Y. Finally, we assume that $\|u\|_2/\sqrt{n} = 1$.

In the course of the text, we carried out successfully computations when $u = \mathbf{e}$, but relied to do so on properties of trace (P). The case of general u is more involved and is treated here.

Lemma C-1. *Assume that Λ and u (which is deterministic) are such that*

$$\frac{1}{n^2}\sum_{i=1}^{n} u_i^4 \to 0, \ \textit{and}\ \frac{1}{n^2}\sum_{i=1}^{n}\lambda_i^4 \to 0$$

and that (Assumption-BB) *holds for Λ for a certain sequence $N(n)$.*

Under the preceding assumptions, we have, if $Z(u) = \frac{1}{n}u'Pu$,

$$Z(u) - \frac{1}{n}\sum_{i=1}^{n} u_i^2 \mathbf{E}\left(P(i,i)|\Lambda\right) \to 0 \textit{ in probability}$$

conditionally on Λ.

Proof. We simply sketch the modifications to the proof given after the statement of Theorem 4.2. As noted in Lemma 4.2, the off-diagonal elements of P have mean 0 conditionally on Λ. Now, using the same notations as in Theorem 4.2, we have, using Equation (7) there, if $Z_i(u)$ is the quantity obtained by replacing λ_i by 0 in Z, $r_i = W_i\mathcal{S}_i^{-1}Y_i$, $w_i = r_i'u/n$, and u_i is the i-th coordinate of u,

$$Z(u) = Z_i(u) + \frac{1}{n}\frac{1}{1+\lambda_i^2 q_i}\left(-\lambda_i^2 w_i^2 + 2\lambda_i u_i w_i + \lambda_i^2 u_i w_i\right)\ .$$

The expression between the parentheses is easily seen to be equal to $(1+\lambda_i^2 q_i)u_i^2 - (\lambda_i w_i - u_i)^2$. We get an analog of Equation (8):

$$Z(u) = Z_i(u) + \frac{u_i^2}{n} + \frac{1}{n}\frac{(\lambda_i w_i - u_i)^2}{1+\lambda_i^2 q_i}$$

Clearly, from the definition of w_i, $w_i|\left\{Y_{(-i)},\Lambda\right\} \sim \mathcal{N}(0, u'W_i\mathcal{S}_i^{-2}W_i'u/n^2)$. Since by assumption $\|u\|_2 = \sqrt{n}$, we have

$$0 \le u'W_i\mathcal{S}_i^{-1}W_i'u/n^2 = u'W_i(W_i'W_i)^{-1}W_i'u/n \le 1$$

because $W_i(W_i'W_i)^{-1}W_i'$ is an orthogonal projection matrix (hence its eigenvalues are only 0 and 1) and $\|u/\sqrt{n}\|_2 = 1$.

So we are exactly in the situation we were in during the proof of Theorem 4.2, except for a term in u_i^4 that now appears in our bound on the variance. Hence, with our extra assumption on $\|u\|_4^4/n^2$, we conclude similarly (after a regularization step) that $Z(u)$ converges in probability, conditional on Λ to its conditional mean which is simply

$$\frac{1}{n}u_i^2\mathbf{E}\left(P(i,i)|\Lambda\right) .$$

□

We remark that to get an analog of Theorem 4.3, where now

$$\zeta = \frac{1}{n}u'\Lambda Y\mathcal{S}^{-1}v ,$$

one just need to go through the proof and replace the w_i appearing there by the "new" $w_i = u'W_i\mathcal{S}_i^{-1}Y_i/n$. Exactly the same arguments go through when $\sum_{i=1}^n u_i^2\lambda_i^2/n$ remains bounded. So under this condition, ζ tends to zero in probability.

With the help of the previous lemma, we can now prove the gist of Proposition 4.2.

Fact C.1. *Proposition 4.2 holds*

Proof. We note that Proposition 4.2 is essentially an application of the previous Lemma, with appropriate change of notation. Recall the notations from the proposition. We have $\widetilde{X} = \Lambda Y\Sigma^{1/2}$ and Λ, which is $n \times n$, has singular value decomposition ADB'. Also, $\mathcal{S} = \widetilde{X}'\widetilde{X}/n$, $\widetilde{Y} = B'Y$, $F = \widetilde{Y}'D^2\widetilde{Y}/n$. Hence, in the language of the proposition,

$$\widehat{m}\mathcal{S}^{-1}\widehat{m} = \frac{1}{n^2}\mathbf{e}'AD\widetilde{Y}F^{-1}\widetilde{Y}'DA\mathbf{e} = \omega'P\omega ,$$

where $P = D\widetilde{Y}\left(\widetilde{Y}'D^2\widetilde{Y}\right)^{-1}\widetilde{Y}'D$ and $\omega = A'\mathbf{e}$. When the assumptions of the proposition are in force, Λ is deterministic, Lemma C-1 applies, from which we conclude

$$\widehat{m}\mathcal{S}^{-1}\widehat{m} - \frac{1}{n}\sum_{i=1}^n \omega_i^2\mathbf{E}\left(P(i,i)\right) \to 0 \text{ in probability.}$$

This gives us the analog of Theorem 4.2.

To get the analog of Theorem 4.3, we just need $\sum_{i=1}^n \omega_i^2 d_i^2/n$ to remain bounded, which is an assumption stated in Proposition 4.2. □

C-3 Bootstrap specific results

Bootstrapping Gaussian data Our analysis of the bootstrap problem requires an analysis similar to the one we performed in the previous subsection. In particular, there we have $u = \Lambda^{1/2}\mathbf{e}$, where Λ contains the bootstrap weights. Since those add-up to n, the assumption $\|u\|_2^2 = n$ was clearly satisfied. Also, in the situation where $p/n \to \rho \in (0, 1 - 1/e)$, we are guaranteed that

$$P^* = \Lambda^{1/2}Y(Y'\Lambda Y)^{-1}Y'\Lambda^{1/2}$$

is well defined with high-probability. When conditioning on Λ, we see that we can work only with the submatrix Λ^* (of size n^*) whose diagonal entries are non-zero. This submatrix has its diagonal entries bounded away from 0 as they are at least equal to 1. Also, using arguments similar to those given in the proof of Lemma B-1, we see that we can get a uniform (in Λ) lower bound on the smallest singular value of ΛY, which holds with probability exponentially (in $(n^* - p)$) close to 1.

So now we assume that we are dealing with Λ such that $n^* - p$ tends to ∞, the empirical distribution of Λ goes to Po(1) and $\sum \lambda_i^2/n^2 \to 0$. We also assume that (Assumption-BB) are satisfied for this

Λ. We call the corresponding set of matrices $\mathcal{G}_{B_n}$. When the diagonal entries of Λ are drawn from a Multinomial$(\frac{1}{n}, \ldots, \frac{1}{n}, n)$ it is clear that these conditions are satisfied with probability going to 1.

The main question that we still have to address is that of the behavior of

$$\frac{1}{n}\sum_{i=1}^{n} u_i^2 \mathbf{E}\left(P^*(i,i)|\Lambda\right)$$

when $u_i^2 = \lambda_i$. By definition,

$$P^*(i,i) = \frac{1}{n}\lambda_i Y_i' \left(\frac{1}{n}\sum_{i=1}^{n}\lambda_i Y_i Y_i'\right)^{-1} Y_i = 1 - \frac{1}{1+\lambda_i Y_i' \mathcal{S}_i^{-1} Y_i/n} ,$$

where $\mathcal{S}_i = \frac{1}{n}\sum_{j\neq i}\lambda_j Y_j Y_j'$. Now the concentration arguments given in El Karoui (2009) show that

$$P\left(\left|\frac{Y_i'\mathcal{S}_i^{-1}Y_i}{p} - \frac{\text{trace}\left(\mathcal{S}_i^{-1}\right)}{p}\right| > t \middle| \mathcal{S}_i^{-1}\right) = \mathrm{O}(\exp(-pt^2/\sigma_p(\mathcal{S}_i))) .$$

We also know that with overwhelming probability (measured over $Y_{(-i)} = \{Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n\}$), $\sigma_p(\mathcal{S}_i)$ is bounded away from 0, conditionally on Λ, when Λ is such that (Assumption-BB) holds. Hence, we conclude that

$$\frac{Y_i'\mathcal{S}_i^{-1}Y_i}{p} \simeq \frac{\text{trace}\left(\mathcal{S}_i^{-1}\right)}{p} \simeq \mathfrak{s} ,$$

where $\mathfrak{s}$ corresponds to the situation where $G \Longrightarrow \mathrm{Po}(1)$ (i.e it is defined by Equation (11)). Hence, conditionally on Λ,

$$P^*(i,i) \simeq 1 - \frac{1}{1+\lambda_i \frac{p}{n}\frac{\text{trace}\left(\mathcal{S}_i^{-1}\right)}{p}} ,$$

with very high-probability, i.e the probability that the difference between the two is greater than t is $\mathrm{O}(\exp(-C(n^*-p)t^2))$ for a fixed C. In other respects, we note that rank-1 perturbation arguments give, if $\mathcal{S} = \frac{1}{n}Y'\Lambda Y$,

$$\text{trace}\left(\mathcal{S}_i^{-1}\right) - \text{trace}\left(\mathcal{S}^{-1}\right) = \frac{\lambda_i}{n}\frac{Y_i'\mathcal{S}_i^{-2}Y_i}{1+\lambda_i Y_i'\mathcal{S}_i^{-1}Y_i/n} .$$

In particular, when Λ is such that (Assumption-BB) holds,

$$P\left(\max_{i=1,\ldots,n}\left|\frac{\text{trace}\left(\mathcal{S}_i^{-1}\right) - \text{trace}\left(\mathcal{S}^{-1}\right)}{p}\right| > \epsilon \middle| \Lambda\right) \to 0.$$

We also note that $\text{trace}\left(\mathcal{S}^{-1}\right)/p \to \mathfrak{s}$ conditionally on Λ, if Λ is such that its empirical distribution goes to Po(1).

Therefore, since $\sum_{i=1}^{n}\lambda_i = n$, we also have by a simple union bound argument, conditional on Λ, and assuming that Λ is such that its empirical distribution goes to Po(1),

$$\frac{1}{n}\sum_{i=1}^{n}\lambda_i P^*(i,i) \simeq 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{\lambda_i}{1+\lambda_i\rho\mathfrak{s}} .$$

Now when $\Lambda \Longrightarrow \mathrm{Po}(1)$, which we write G,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\lambda_i}{1+\lambda_i\rho\mathfrak{s}} \to \int\frac{\tau dG(\tau)}{1+\tau\rho\mathfrak{s}} .$$

But in light of the Marčenko-Pastur equation, we have, under these circumstances,

$$\frac{1}{n}\sum_{i=1}^{n}\lambda_i P^*(i,i) \to 1 - \frac{1}{\mathfrak{s}} = \frac{\mathfrak{s}-1}{\mathfrak{s}} .$$

We finally conclude that conditional on Λ being in the set described above (whose probability goes to 1),

$$(\widehat{\mu}^*)'(\widehat{\Sigma}^*)^{-1}\widehat{\mu}^* \to \frac{\frac{\mathfrak{s}-1}{\mathfrak{s}}}{1-\frac{\mathfrak{s}-1}{\mathfrak{s}}} = \mathfrak{s}-1 \geq \frac{\rho}{1-\rho} \ ,$$

since we know that $\mathfrak{s} \geq 1/(1-\rho)$ when G is Po(1), since its mean is 1.

Similar arguments as the ones used in the proofs in the main body of the paper show that the same convergence in probability result holds unconditionally on Λ - the problem being to get bounds that are uniform in Λ, when $\Lambda \in \mathcal{G}_{B_n}$.

Hence, an analog of Theorem 4.2 follows (with $\mathcal{P}_n$ probability going to 1), where the ratio $\rho/(1-\rho)$ is replaced by $\mathfrak{s}-1$. The analog of Theorem 4.3 follows from the arguments given in Appendix C-2, if we can show, in the notation used there that $\sum_{i=1}^n (u_i d_i)^2/n$ remains bounded with probability going to 1. Note that $u_i = d_i = \sqrt{\lambda_i}$ here, where λ_i are the bootstrap weights, so we just need to show that $\sum_{i=1}^n \lambda_i^2/n$ remains bounded. The mean of this quantity clearly goes to 2, using the marginal distribution of λ_i. On the other hand, the arguments we gave in Proposition 4.4 show that its variance goes to 0, so this quantity goes to 2 in probability and therefore remains bounded with probability going to 1.

We therefore have an analog of Theorem 4.3 and also of Theorem 4.4 when bootstrapping Gaussian data.

Bootstrapping elliptically distributed data Finally, let us say a few words about what would happen if we replaced the normality assumption for the X_i's by an elliptical distribution assumption. We focus on the case where $X_i = \lambda_i \Sigma^{1/2} Y_i$, i.e the mean of the X_i's is 0. The previous analyses make clear that the key questions concern $v'(\widehat{\Sigma}^*)^{-1}v$ and $(\widehat{\mu}^*)'(\widehat{\Sigma}^*)^{-1}\widehat{\mu}^*$.

The questions concerning $v'(\widehat{\Sigma}^*)^{-1}v$ fall pretty much directly under the study we have made of elliptical distributions, since we know, according to the proof of Theorem 4.5, that

$$\widehat{\Sigma}^* = \frac{1}{n-1}\Sigma^{1/2}Y'\Lambda' D^{1/2}(\mathrm{Id}_n - \delta\delta'/n)D^{1/2}\Lambda Y \Sigma^{1/2} \ ,$$

where D is the diagonal matrix containing the bootstrap weights and $\delta = D^{1/2}\mathbf{e}$. So, as long as $D^{1/2}\Lambda$ satisfies (Assumption-BB), results similar to Theorem 4.5 will hold.

The questions dealing with $(\widehat{\mu}^*)'(\widehat{\Sigma}^*)^{-1}\widehat{\mu}^*$ are more involved. Analyses similar to the ones performed above show that the key quantity to understand is now

$$\frac{1}{n}\mathbf{e}' D\Lambda Y (Y'\Lambda' D \Lambda Y)^{-1} Y'\Lambda' D\mathbf{e} = \frac{1}{n} u' P_{D^{1/2}\Lambda, Y} u \ ,$$

where $P_{D^{1/2}\Lambda,Y} = D^{1/2}\Lambda Y (Y'\Lambda' D\Lambda Y)^{-1} Y'\Lambda D^{1/2}$ and $u = D^{1/2}\mathbf{e}$. The analysis of this quadratic form can be carried out just like we did above in the Gaussian case, i.e $\Lambda = \mathrm{Id}_n$. However, the remarks we made to get simplified expressions for the limit do not seem to apply anymore: quantities of the type

$$\frac{1}{n}\sum_{i=1}^n \frac{d_i}{1+\lambda_i^2 d_i \rho \mathfrak{s}} \ ,$$

appear, where $\mathfrak{s}$ is the solution of Equation (4) with G being the limit (if it exists) of the empirical distribution of the random variables $\lambda_i^2 d_i$. These quantities do not appear to simplify any further to yield a clearer and more exploitable expression.

References

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition.

Bai, Z. D. (1999). Methodologies in spectral analysis of large-dimensional random matrices, a review. *Statist. Sinica* **9**, 611–677. With comments by G. J. Rodgers and Jack W. Silverstein; and a rejoinder by the author.

Bickel, P. J. and Levina, E. (2007a). Covariance regularization by thresholding. Technical Report 744, Department of Statistics, UC Berkeley.

Bickel, P. J. and Levina, E. (2007b). Regularized estimation of large covariance matrices. *The Annals of Statistics* To Appear.

Black, F. and Litterman, R. (1990). Asset allocation: combining investor views with market equilibrium. *Golman Sachs Fixed Income Research* .

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization.* Cambridge University Press, Cambridge.

Campbell, J., Lo, A., and MacKinlay, C. (1996). *The Econometrics of Financial Markets.* Princeton University Press, Princeton, NJ.

Chikuse, Y. (2003). *Statistics on special manifolds*, volume 174 of *Lecture Notes in Statistics.* Springer-Verlag, New York.

Chow, Y. S. and Teicher, H. (1997). *Probability theory.* Springer Texts in Statistics. Springer-Verlag, New York, third edition. Independence, interchangeability, martingales.

Davidson, K. R. and Szarek, S. J. (2001). Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces, Vol. I*, pp. 317–366. North-Holland, Amsterdam.

Eaton, M. L. (1983). *Multivariate statistics.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York. A vector space approach.

El Karoui, N. (2007). Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *The Annals of Probability* **35**, 663–714.

El Karoui, N. (2008a). Operator norm consistent estimation of large dimensional sparse covariance matrices. *The Annals of Statistics* **36**, 2717–2756.

El Karoui, N. (2008b). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics* **36**, 2757–2790.

El Karoui, N. (2009). Concentration of measure and spectra of random matrices: with applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability* To Appear.

Fang, K. T., Kotz, S., and Ng, K. W. (1990). *Symmetric multivariate and related distributions*, volume 36 of *Monographs on Statistics and Applied Probability.* Chapman and Hall Ltd., London.

Frahm, G. and Jaekel, U. (2005). Random matrix theory and robust covariance matrix estimation for financial data. *arXiv:physics/0503007* .

Horn, R. A. and Johnson, C. R. (1990). *Matrix analysis.* Cambridge University Press, Cambridge. Corrected reprint of the 1985 original.

Horn, R. A. and Johnson, C. R. (1994). *Topics in matrix analysis.* Cambridge University Press, Cambridge. Corrected reprint of the 1991 original.

Jobson, J. D. and Korkie, B. (1980). Estimation for Markowitz efficient portfolios. *J. Amer. Statist. Assoc.* **75**, 544–554.

Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal component analysis. *Ann. Statist.* **29**, 295–327.

Kan, R. and Smith, D. R. (2008). The distribution of the sample minimum-variance frontier. *Management Science* **54**, 13641380.

Lai, T. L. and Xing, H. (2008). *Statistical Models and Methods for Financial Markets.* Springer Texts in Statistics. Springer, New York.

Laloux, L., Cizeau, P., Bouchaud, J.-P., and Potters, M. (2000). Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance* **3**, 391–397.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88**, 365–411.

Lugosi, G. (2006). Concentration of measure inequalities. Lecture notes available online.

Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)* **72 (114)**, 507–536.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis.* Academic Press [Harcourt Brace Jovanovich Publishers], London. Probability and Mathematical Statistics: A Series of Monographs and Textbooks.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance* **7**, 77–91. URL `http://www.jstor.org/stable/2975974`.

McNeil, A. J., Frey, R., and Embrechts, P. (2005). *Quantitative risk management.* Princeton Series in Finance. Princeton University Press, Princeton, NJ. Concepts, techniques and tools.

Meucci, A. (2005). *Risk and asset allocation.* Springer Finance. Springer-Verlag, Berlin.

Meucci, A. (2008). Enhancing the Black-Litterman and related approaches: views and stress-test on risk factors. Available at SSRN, http://ssrn.com/abstract=1213323.

Michaud, R. O. (1998). *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation.* Oxford University Press, USA.

Pafka, S. and Kondor, I. (2003). Noisy covariance matrices and portfolio optimization. II. *Phys. A* **319**, 487–494.

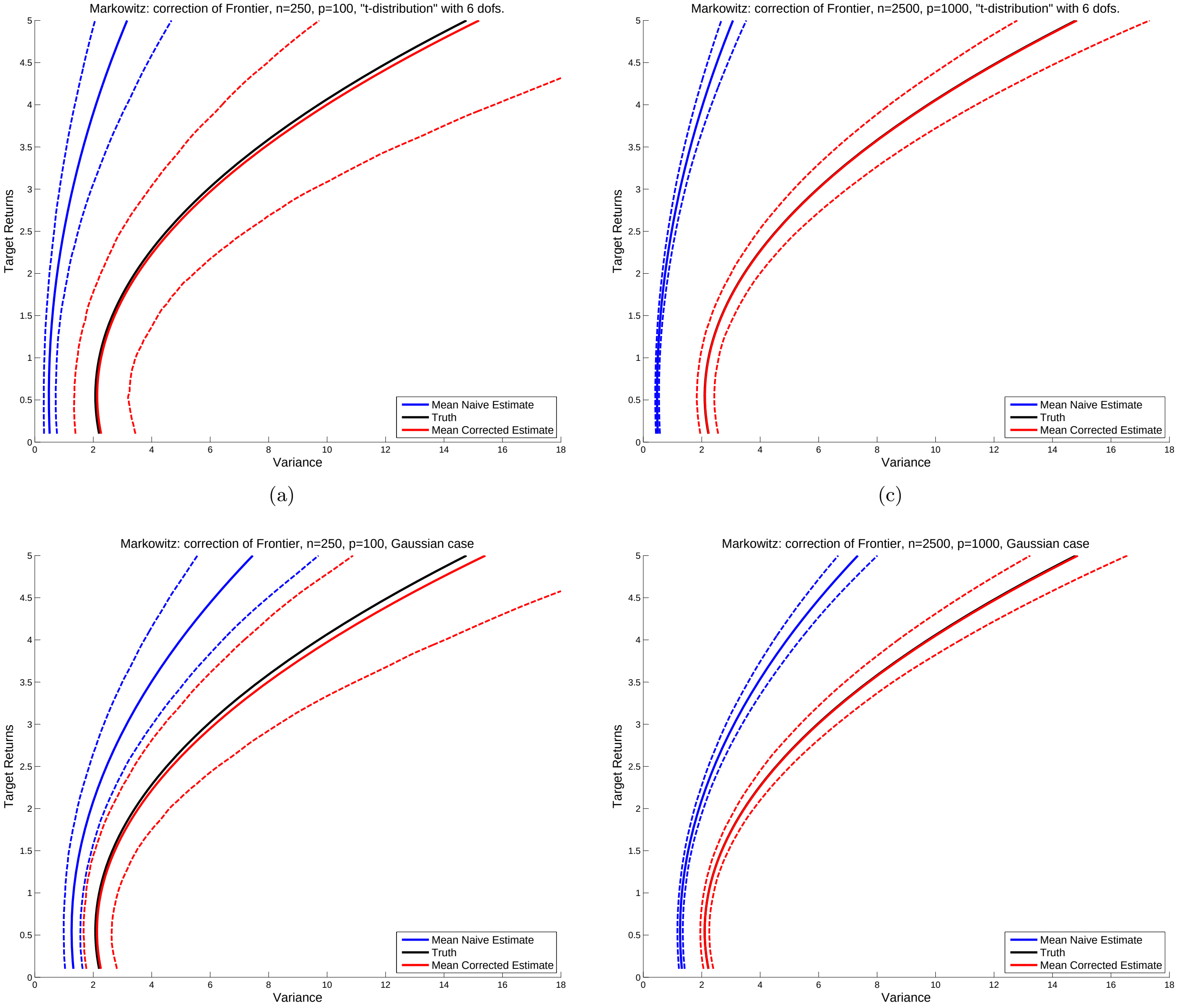
Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2**, 494–515. (electronic). DOI: 10.1214/08-EJS176.

Ruppert, D. (2006). *Statistics and finance.* Springer Texts in Statistics. Springer, New York. An introduction, Corrected second printing of the 2004 original.

Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.* **55**, 331–339.

van der Vaart, A. W. (1998). *Asymptotic statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Wachter, K. W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. *Ann. Probability* **6**, 1–18.



(a)

(c)

(b)

(d)

Figure 4: Performance of naive and corrected frontiers, for scaled "t_6" (upper pictures, (a) and (c)) and Gaussian returns ((b) and (d)). Here, in the left column $n = 250$ and $p = 100$. In the right column, $n = 2500, p = 1000$. The number of simulations is 1000 in all pictures. The dashed lines represent (empirical) 95% confidence bands. (The confidence bands corresponds are computed for a fixed y.) The x-axis represents our estimate of variance of the optimal portfolio. The y-axis represents the target returns for the portfolio. The plots show both the bias in the naive solution (blue solid curves) and the fact that our estimator is nearly unbiased (red solid curves near, or covering the black curve, the population solution). They also illustrate the robustness of our corrections. Another striking feature is the lack of robustness of Gaussian computations, since the "efficient frontiers" computed with "t_6" returns are different from the Gaussian ones. The fact that, as our theoretical work predicts, Gaussian computations underestimate risk-underestimation in the class of elliptical distributions considered in the paper is illustrated by the fact that the "t_6" curves are to the left of the Gaussian curves. Note the narrower confidence bands in the larger dimensional simulations ((c) and (d)). The black line is essentially hidden under the red line in (c) and (d), showing a near perfect correction (on average).