

On the realized risk of high-dimensional Markowitz portfolios

Noureddine El Karoui
Department of Statistics, UC Berkeley*

Abstract

We study the realized risk of Markowitz portfolio computed using parameters estimated from data and generalizations to similar questions involving the out-of-sample risk in quadratic programs with linear equality constraints.

We do so under the assumption that the data is generated according to an elliptical model, which allows us to study models where we have heavy-tails, tail dependence, and leptokurtic marginals for the data. We place ourselves in the setting of high-dimensional inference where the number of assets in the portfolio, p , is large and comparable to the number of samples, n , we use to estimate the parameters. Our approach is based on random matrix theory. We consider both the impact of the estimation of the mean and of the covariance.

Our work shows that risk is underestimated in this setting, and further, that in the class of elliptical distributions, the Gaussian case yields the least amount of risk underestimation. The problem is more pronounced for genuinely elliptical distributions and Gaussian computations give an overoptimistic view of the situation.

We also propose a robust estimator of realized risk and investigate its performance in simulations.

1 Introduction

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 of course part of the standard textbook literature on these issues (Ruppert (2006), Campbell et al. (1996)).

Naturally, many variants exist now, involving various notions of risk. The most common ones seem to involve Value-At-Risk (VaR) and conditional Value-At-Risk (cVaR) as alternatives to variance. We discuss here only the classical problem.

In the ideal (or, in statistical parlance, population) solution, the covariance and the mean of the returns are known. The mathematical formulation is then the following simple quadratic program: we wish to find the weights w by solving 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, μ is the vector of mean returns, Σ is the covariance between the returns of the assets, and μ_P is the level of expected returns the investor wishes to achieve. 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*.

*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 an Alfred P. Sloan research Fellowship, the France-Berkeley Fund, 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, quadratic programs, multivariate statistical analysis, high-dimensional inference, random matrix theory, concentration of measure, Markowitz problem, elliptical distributions. **Contact :** nkaroui@stat.berkeley.edu

We note that the problem is sometimes formulated slightly differently, i.e with the constraint $w'\mu \geq \mu_P$ instead of $w'\mu = \mu_P$. However, this has minimal consequences in our setting since knowing the solution of the problem we consider (for all μ_P) will yield the solution of the problem involving the inequality constraint. If needed, another motivation for studying the above mentioned problem for all μ_P 's is that in the case where we want to incorporate a riskless asset in the portfolio, the shape of the efficient frontier changes and becomes a straight line. If the variance of the portfolio is on the x -axis and the level of expected returns is on the y -axis, this straight line goes through $(0, r)$ and is tangent to the efficient frontier computed above, which is a parabola. The line touches the parabola at a point called the *tangency portfolio* and we naturally need the whole efficient frontier to compute it (see e.g Ruppert (2006)).

Going back to the original problem, in practice, we of course do not know μ and Σ and we need to estimate them. An interesting question is therefore to understand what happens in the Markowitz problem when we replace population quantities by corresponding estimators. In this paper we will be especially concerned with the following risk management question: if we choose our strategy by solving the Markowitz problem with μ (resp. Σ) replaced by the sample mean $\hat{\mu}$ (resp. the sample covariance matrix $\hat{\Sigma}$), what is the risk of our portfolio? Related interesting questions are naturally the respective contributions of $\hat{\mu}$ and $\hat{\Sigma}$ in our measures of risk, and the problems they may create.

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 in the form presented here being a particular instance of such a problem. This more general question is relevant for various statistical problems where we are interested in out-of-sample measures of risk.

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 has 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.

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 $\hat{\Sigma}$, 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 $\hat{\Sigma}$ is unbiased for Σ , the eigenvalues and eigenvectors of $\hat{\Sigma}$ behave very differently from those of Σ . We refer the interested reader to Johnstone (2001), El Karoui (2007), El Karoui (2008a), Bickel and Levina (2007), Rothman et al. (2008), El Karoui (2009a) for a partial introduction to these problems.

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 classic paper Jobson and Korkie (1980) is concerned, in the Gaussian case, with issues similar to the ones we will be investigating. However, it does not seem that so far there has been much interest in this high-dimensionality question in the finance literature. For instance, a book-length treatment of asset allocation questions (Meucci, 2005), gives only a rather cursory one page discussion of these issues.

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 (2009a) 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. (As we will see, these geometric implications have a strong impact on the results we will present.)

In the simplest setting, we consider data $\{X_i\}_{i=1}^n$, which are p -dimensional. In a financial context, these vectors are 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 $\mathcal{N}(0, \text{Id}_p)$ - the normality assumption for the data being close to assuming a Black-Scholes model for the underlying diffusion of stock prices. 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 $\rho = \lim_{n \rightarrow \infty} p/n$ in the following manner: if $p < n$, the density of F is

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

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

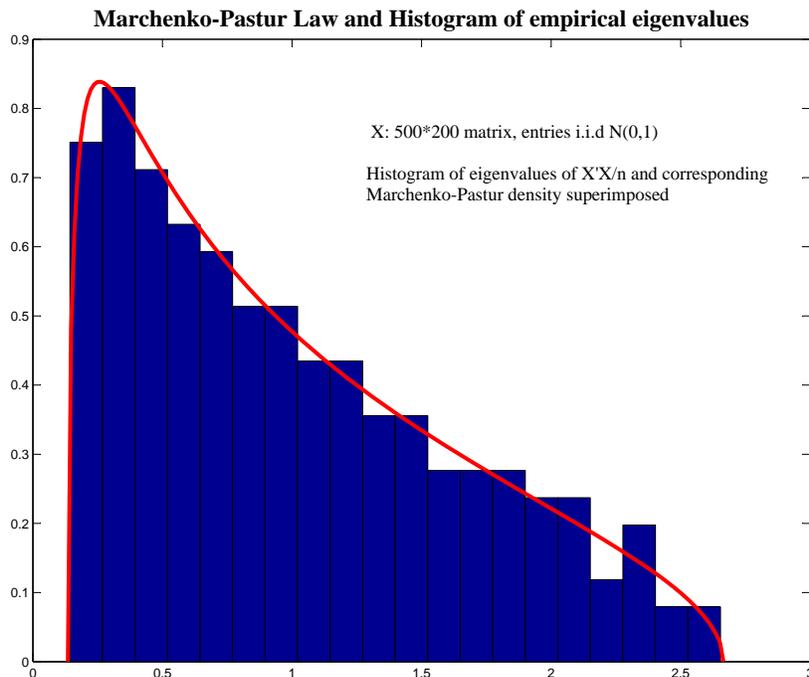


Figure 1: Illustration of Marčenko-Pastur law, $n=500$, $p=200$. The red curve is the density of the Marčenko-Pastur -law for $\rho = 2/5$. The simulation was done with i.i.d Gaussian data. The histogram is the histogram of eigenvalues of $X'X/n$

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 $X_{n+1} \sim \mathcal{N}(0, \text{Id}_p)$, $\text{var}(v'X_{n+1}) = 1$, for all v , since $v'X_{n+1} \sim \mathcal{N}(0, 1)$. However, if we do not know Σ and estimate it by $\widehat{\Sigma}$, a naive (and wrong) reasoning 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_{n+1} independent of $\{X_i\}_{i=1}^n$, the variance in the direction of v_p , $\text{var}(v_p'X_{n+1})$, 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. Naturally, this will also lead us to choose a portfolio that is suboptimal and should therefore have higher realized risk than the optimal portfolio.

This simple argument suggests that severe problems might arise in the high-dimensional Markowitz problem and other quadratic programs with linear equality constraints. In particular, risk might be underestimated. While this heuristic argument is probably clear to specialists of random matrix theory, as far as we know, the problem has not been investigated at a mathematical level of rigor in that literature until the very recent El Karoui (2009b). 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 note that there are claims of universality in the physics literature, in other words, the model for the data would not matter, only the population parameters (and n and p) would, and we show below that there is no universality in these problems. Finally, we personally found the physics literature very hard to read because of the lack of detailed and rigorous proofs and its reliance on a vocabulary that is different from the one in mathematical random matrix theory.

In this paper, which is a companion to El Karoui (2009b), we propose a theoretical analysis of the problem in an elliptical framework (which incorporates the Gaussian case as a subcase) for general quadratic programs with linear equality constraints, one of them involving the parameter μ . We treat the problem at this level of generality because, beyond the finance settings, the results should be interesting in statistics and several areas of applied mathematics where one need to solve optimization problems based on estimated parameters. Our results are several-fold. We relate the realized risk of the portfolios to the theoretical efficient frontier that is key to the Markowitz theory. We quantify this realized risk as a function of the population parameters, n , p and a quantity characterizing the ellipticity of the data. Finally, we propose an estimator of this realized risk that is easy to compute. We show its performance in some simulations.

The elliptical framework for modeling returns of financial stocks has been advocated in the literature for some time now (see Frahm and Jaekel (2005)) and is still, as a random matrix problem, quite interesting to theoreticians (see El Karoui (2009a)). One of its many benefits is that it allows us to incorporate heavy-tailed modeling in our analysis, and it yields marginal distributions for the returns of individual assets that are all leptokurtic. Finally, elliptical distributions have non-zero coefficient of tail dependence (see McNeil et al. (2005)), whereas the Gaussian distribution has zero dependence in the tail.

Interestingly, there seems to be a consensus in the finance literature that estimation of covariance is “easy”, and the more difficult aspect of the Markowitz problem (and other portfolio optimization problems) comes from estimating the mean. By contrast, statisticians working in high-dimensional inference have

recently devoted a lot of efforts to improving covariance estimation, which is thought to be a hard task. We show here (and showed in El Karoui (2009b)) that estimating the mean and the covariance matrix both matter and create quantifiable (and first-order) problems and biases.

The paper is divided into three main parts and a conclusion. In Section 2, we give some preliminaries on the problem we are tackling here and the key results of El Karoui (2009b) that will prove useful in the paper. In Section 3, we state the main technical results of the paper and discuss them briefly. Because the proofs are rather technical and self contained, they are given in an Appendix. In Section 4, we compare the results in the Gaussian and genuinely elliptical setting. The main conclusion there is that the Gaussian case gives an over-optimistic view of risk underestimation and that risk underestimation is more pronounced in the elliptical setting. This shows in particular that no “universality” should be expected in these questions. We also present some simulations that illustrate both the accuracy and potential limitations of our work. We discuss our results and possible extensions in the conclusion.

2 Preliminaries

In this section we remind the reader of some classical and well-known results concerning quadratic programs with linear equality constraints. We also briefly reminds the reader of the key results in El Karoui (2009b) we will need later on.

Setup of the problem

We get to observe data X_1, \dots, X_n , with distribution

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

where Σ is a $p \times p$ covariance matrix and the Y_i are i.i.d $\mathcal{N}(0, \text{Id}_p)$. Here we assume that the λ_i 's are independent of the Y_i 's, but they might be correlated with one another. We note that $\mathbf{E}(X_i) = \mu$ and $\text{cov}(X_i) = \mathbf{E}(\lambda_i^2) \Sigma$. If $\mathbf{E}(\lambda_i^2) = 1$, all these models lead to the same covariance for the data, Σ . We note that X_i has as many moments as λ_i , and in particular can have much heavier tails than Gaussian data. It is also easy to see that the marginals of X_i are leptokurtic. By a slight abuse of language, we call data distributed according to the model stipulated in Equation (1) elliptical, though the model is slightly different than the standard model for elliptical data in statistics.

Known results

We are interested in the solution of the following problem, which is a generalization of the Markowitz problem:

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

We remind the reader of the following fact:

Fact 2.1. *The solution of Problem (QP-eqc-Pop) is given by*

$$w_{\text{optimal}} = \Sigma^{-1} V M^{-1} U,$$

where V is the $p \times k$ matrix containing the v_i 's, U is the $k \times 1$ vector containing the u_i 's, and $M = (V' \Sigma^{-1} V)^{-1}$, provided all these quantities exist.

Throughout, we will assume that the number of constraints k stays fixed in the asymptotics we consider. Our assumptions will also guarantee that Σ^{-1} and M^{-1} exists. Unless otherwise noted, these assumptions will be made implicitly throughout the paper.

In the situation we care most about, $v_k = \hat{\mu}$ and $\Sigma = \hat{\Sigma}$ are estimated from data. We will call \hat{V} the $p \times k$ matrix containing the v_i 's. In other word, we will seek the solution of the problem

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

We call w_{emp} the vector of weights obtained by solving the problem (QP-eqc-Emp). A very important question is to understand how w_{emp} and functions of this vector relate to w_{optimal} and functions of w_{optimal} . We will do so in the setting where p and n are large and do asymptotic computations when $p \rightarrow \infty$ and $n \rightarrow \infty$, while $p/n \rightarrow \rho \in (0, 1)$. A reason for doing this sort of double asymptotic computations is that they might yield better insights than standard (i.e fixed p , large n) asymptotics when p and n are moderately large, i.e in the few 100's. This is verified in our simulations.

In El Karoui (2009b), we focused on the issue of relating the naive estimator of risk $w'_{\text{emp}} \hat{\Sigma} w_{\text{emp}}$ to the population risk $w'_{\text{optimal}} \Sigma w_{\text{optimal}}$. By contrast, in this paper for reasons that are explained below we will be focusing on $w'_{\text{emp}} \Sigma w_{\text{emp}}$.

On the realized risk of portfolios

From a risk management standpoint, a natural quantity to estimate is the realized risk (or out-of-sample risk) of a vector obtained by solving Problem (QP-eqc-Emp). We first place ourselves in the setting where the X_i 's are independent. By realized risk, we mean

$$\text{RRisk} = \text{var} (w'_{\text{emp}} X_{n+1} | X_1, \dots, X_n) = w'_{\text{emp}} \Sigma w_{\text{emp}}, \quad (2)$$

namely the risk that we will be subjected to in the future if we chose w_{emp} as our allocation today, and the returns are independent. The previous result naturally holds under the milder assumption that X_{n+1} is independent of $\{X_i\}_{i=1}^n$. If the λ_i 's are dependent, while the Y_i 's are independent, we have

$$\text{var} (w'_{\text{emp}} X_{n+1} | X_1, \dots, X_n) = \mathbf{E} (\lambda_{n+1}^2 | \{\lambda_i\}_{i=1}^n) w'_{\text{emp}} \Sigma w_{\text{emp}}.$$

So most of our work will focus on understanding the random variable $w'_{\text{emp}} \Sigma w_{\text{emp}}$, which we will call the realized risk, while keeping in mind that our analysis would also give us (through a simple modification) results concerning the case where the λ_i 's are dependent.

In particular, we might want to compare this risk to the naive estimator of risk, $w'_{\text{emp}} \hat{\Sigma} w_{\text{emp}}$, and to the actual optimal risk, $w'_{\text{optimal}} \Sigma w_{\text{optimal}}$. These latter two estimators were considered in El Karoui (2009b), therefore we only need to compare RRisk and $w'_{\text{optimal}} \Sigma w_{\text{optimal}}$. Clearly to do so, we will only need to focus on understanding $w'_{\text{emp}} \Sigma w_{\text{emp}}$. Note that this is not such an easy quantity to estimate since Σ is unknown, and as we said earlier, the sample covariance matrix is a poor estimator of Σ in high-dimension, the setting we will consider here.

From now on we will assume throughout that $\hat{\mu}$ is the sample mean and $\hat{\Sigma}$ is the sample covariance matrix. In other words, X is the data matrix whose i -th row is X_i , $\hat{\mu}' = \mathbf{e}' X / n$ and $\hat{\Sigma} = (X - \mathbf{e} \hat{\mu}')' (X - \mathbf{e} \hat{\mu}') / n - 1$. We will sometime refer to the $n \times p$ matrix $\mathbf{e} \hat{\mu}'$ as \bar{X} . The sample covariance matrix could also be rescaled by $1/n$ instead of $1/(n-1)$, in which case it is not unbiased but since this might lead to rather less cumbersome expressions, we will sometime choose this normalization. Note that the normalization will have no effect on our asymptotic results.

A simple but key observation in what follows is the following simple fact.

Fact 2.2. *Suppose the observed data can be written in matrix form*

$$X = \mathbf{e} \mu' + \Sigma_1 Y \Sigma = \mathbf{e} \mu' + Y_1 \Sigma.$$

Then, if $H = \text{Id}_n - \mathbf{e} \mathbf{e}' / n$, and $\hat{M} = \hat{V}' \hat{\Sigma}^{-1} \hat{V}$,

$$\text{RRisk} = w'_{\text{emp}} \Sigma w_{\text{emp}} = U' \hat{M}^{-1} \hat{V}' \Sigma^{-1/2} \left(\frac{Y_1' H Y_1}{n-1} \right)^{-2} \Sigma^{-1/2} \hat{V} \hat{M}^{-1} U. \quad (3)$$

The simple fact above is an immediate consequence of the fact that $\widehat{\Sigma} = X'HX/(n-1)$ and therefore, $\widehat{\Sigma} = \Sigma^{1/2} \left(\frac{Y_1'HY_1}{n-1} \right) \Sigma^{1/2}$.

The expression we give in Equation (3) might look relatively nasty. However, it considerably simplifies the problem. As a matter of fact, U and \widehat{M} are finite dimensional objects, which are now well understood (see El Karoui (2009b)). For instance we showed in El Karoui (2009b), that under technical assumptions similar to the ones we will be making in this paper, if $\kappa = \rho/(1-\rho)$,

$$\widehat{M} \simeq \mathfrak{s}V'\Sigma^{-1}V + \kappa e_k e_k',$$

where \mathfrak{s} is defined below in Equation (4). Our assumptions guarantee that the approximate equality above holds in probability and we can also take the inverses of the two matrices on both sides of the approximate equality and have approximate equality for the inverses.

So the only real difficulty we will have to deal with is to understand the $k \times k$ matrix

$$\widehat{V}'\Sigma^{-1/2} \left(\frac{Y_1'HY_1}{n-1} \right)^{-2} \Sigma^{-1/2}\widehat{V}.$$

On closer inspections, and with insights coming from El Karoui (2009b), it turns out that we only need to understand mainly two things: first, for certain well-chosen deterministic vectors α , we need to grapple with

$$\alpha' \left(\frac{Y_1'HY_1}{n-1} \right)^{-2} \alpha.$$

We will see that the ideas we developed in El Karoui (2009b) will be extremely helpful in that context. Second, we will have to consider

$$\widehat{\mu}'\Sigma^{-1/2} \left(\frac{Y_1'HY_1}{n-1} \right)^{-2} \Sigma^{-1/2}\widehat{\mu}.$$

This quantity will require substantially more work.

The key insight from random matrix theory we will need in this context is the fact that these random quantities converge to (deterministic) constants in the asymptotic setting we consider. Hence, it turns out that RRisk will be the product of 5 essentially deterministic matrices, and we will be able to relate this product to the ‘‘population’’ quantity (or theoretical efficient frontier) $w'_{\text{optimal}}\Sigma w_{\text{optimal}} = U'M^{-1}U$.

In Section 3, we present our main results and apply them to compare the Gaussian and elliptical case in Section 4.

Notations Before we start presenting our results, let us describe some notations we will be using. $\|v\|$ is by default the Euclidian norm of the vector v . We sometime also write $\|v\|_2$. \succeq represents the positive semi-definite ordering for matrices: so if $A \succeq B$, $A - B$ is positive semi-definite. $\|A\|_2$ is the operator norm or largest singular value of the matrix A . $o_P(1)$ means that the corresponding random variable goes to zero in probability. $a \vee b$ stands for $\max(a, b)$. The vector \mathbf{e} is a vector whose entries are all equal to 1; it is generally of dimension n .

3 Main results

In this section, we state our main results so as to extract them from the technical details of the proofs. We wish to note that we will work with the assumption that Y_i 's (see Equation (1)) are Gaussian. We chose to do so to limit the technical details and to bypass standard methods of random matrix theory, in the hope that our proofs would show more clearly the phenomena at play. As explained in El Karoui (2009a), most of the results of random matrix theory we will rely on depend strongly on the geometry that the purported model implies on the data. By working with our model we are able to capture all the richness of elliptical models from a random matrix theoretic point of view, in particular the different models induce differences in the geometry of the data, while keeping proofs relatively clear. It is very likely possible to use other well known techniques (based on Stieltjes transforms, so more specialized and less broadly accessible) to

weaken the assumptions on the Y_i 's and replace the current ones by assumptions concerning concentration of convex Lipschitz functions of the Y_i 's (see Ledoux (2001) and El Karoui (2009a) for concrete examples in the present context). However, this would change essentially nothing to the geometry of the data - the key driver of the results in our opinion - and hence we expect to get the same results under these weaker assumptions. Work in this direction is currently under way. In the present paper we really want to focus on the key phenomena and not on what is now for serious practitioners of random matrix theory essentially manageable technical details.

As said earlier, we will need two main results to draw conclusions about the realized risk of high-dimensional Markowitz portfolios. We now present them.

3.1 On $V'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}V$

The first question is to get a good understanding of the quantity

$$V'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}V,$$

where V is $p \times k$ deterministic and given matrix.

We have the following results, proven in Appendix A.

Theorem 3.1. *Suppose we observe n i.i.d observations X_i , where X_i has the form $X_i = \mu + \lambda_i Y_i$, with $Y_i \stackrel{iid}{\sim} \mathcal{N}(0, \text{Id}_p)$ and $\{\lambda_i\}_{i=1}^n$ is independent of $\{Y_i\}_{i=1}^n$. We assume that $\mathbf{E}(\lambda_i^2) = 1$. We call X the $n \times p$ data matrix containing the X_i 's.*

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

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

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

Under these assumptions, if α is a (sequence of) deterministic vectors with norm 1,

$$\alpha' \left(\frac{X'HX}{n-1} \right)^{-2} \alpha \rightarrow \xi \text{ in probability}$$

where if \mathfrak{s} satisfies

$$\int \frac{dG(\tau)}{1 + \rho\tau\mathfrak{s}} = 1 - \rho, \quad (4)$$

ξ is defined as

$$\xi = \frac{1}{\frac{1}{\mathfrak{s}^2} - \rho \int \frac{\tau^2 dG(\tau)}{(1 + \tau\rho\mathfrak{s})^2}}. \quad (5)$$

We note that if v is a given deterministic vector and $\alpha = \Sigma^{-1/2}v/\sqrt{v'\Sigma^{-1}v}$, the previous theorem means that, under the appropriate technical conditions,

$$\frac{v'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}v}{v'\Sigma^{-1}v} \rightarrow \xi \text{ in probability}.$$

It is in this latter form that the theorem is going to be used.

To flesh out a little bit the result, let us point out that in the Gaussian case, $G = \delta_1$, so $\mathfrak{s} = 1/(1 - \rho)$ and $\xi = (1 - \rho)^{-3}$.

Perhaps remarkably, even though it is stated with the condition $\mathbf{E}(\lambda_i^2) = 1$, this theorem does not actually require any assumptions on the moments of λ_i , nor do we need to assume that the λ_i 's are

independent of each other. All that really matters is the existence of a deterministic limiting empirical distribution for $\tau_i = \lambda_i^2$. (Assumption-BB) essentially means that the distribution of the λ_i 's does not put "too much" mass near 0, which will guarantee that the smallest singular values of certain random matrices appearing in our computations are bounded away from 0. This is very important in the proof.

To get a sense of differences between the elliptical and Gaussian cases, let us mention the following fact:

Fact 3.1. *We have*

$$\xi \geq \frac{\mathfrak{s}^2}{1 - \rho} .$$

This latter value corresponds to the Gaussian case.

Let us now turn to the more general situation. Let us call *P1* the following condition on the population parameters:

$$\forall 1 \leq i \neq j \leq k, \frac{v_i' \Sigma^{-1} v_j}{(v_i + v_j)' \Sigma^{-1} (v_i + v_j)} \text{ and } \frac{v_i' \Sigma^{-1} v_j}{(v_i - v_j)' \Sigma^{-1} (v_i - v_j)} \text{ stay bounded away from } 0 . \quad (\text{Condition-P1})$$

As an immediate corollary of Theorem 3.1, we have

Corollary 3.1. *Suppose now that $X_i = \mu + \lambda_i \Sigma^{1/2} Y_i$ and the conditions of Theorem 3.1 hold. With the definitions above, and if the $p \times k$ deterministic matrix V and Σ are such that Condition-P1 is satisfied, we have asymptotically*

$$V' \widehat{\Sigma}^{-1} \Sigma \widehat{\Sigma}^{-1} V = \xi V' \Sigma^{-1} V + o_P(V' \Sigma^{-1} V) = \xi M + o_P(M) .$$

We note that since M is $k \times k$ matrix and k is held fixed in our asymptotics, all norms on M are equivalent. So $o_P(M)$ just means that the maximal entry of $V' \widehat{\Sigma}^{-1} \Sigma \widehat{\Sigma}^{-1} V - \xi M$ is negligible compared to the largest entry of M , or equivalently here that the largest singular value of the difference of the two matrices is negligible compared to that of M .

3.2 Quadratic forms in $\widehat{\mu}$ and $\widehat{\Sigma}^{-1} \Sigma \widehat{\Sigma}^{-1}$

Because we need to estimate the mean of the data, we also have to deal with forms of the type $\widehat{\mu}' \widehat{\Sigma}^{-1} \Sigma \widehat{\Sigma}^{-1} \widehat{\mu}$ and $\widehat{\mu}' \widehat{\Sigma}^{-1} \Sigma \widehat{\Sigma}^{-1} v$, for fixed v . Our main result in this direction is the following. The result is proven in Appendix B.

Theorem 3.2. *Suppose that $X_i = \mu + \lambda_i \Sigma^{1/2} Y_i$, where Y_i are i.i.d $\mathcal{N}(0, \text{Id}_p)$ and $\{\lambda_i\}_{i=1}^n$ are random variables, independent of $\{Y_i\}_{i=1}^n$ and with $\mathbf{E}(\lambda_i^2) = 1$. 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

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

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

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

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 $\mathcal{S} = W'W/n$. Finally, we use the notation $\widehat{m} = W'\mathbf{e}/n$, $\widetilde{\mu} = \Sigma^{-1/2}\mu$.

Then, we have, for ξ defined as in Equation (5), and \mathfrak{s} defined as in Equation (4),

$$\frac{\widehat{\mu}'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}v}{\sqrt{v'\Sigma^{-1}v}} = \frac{\mu'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}v}{\sqrt{v'\Sigma^{-1}v}} + o_P(1) = \xi \frac{\mu'\Sigma^{-1}v}{\sqrt{v'\Sigma^{-1}v}} + o_P\left(1 \vee \frac{\mu'\Sigma^{-1}v}{\sqrt{v'\Sigma^{-1}v}}\right). \quad (6)$$

Also,

$$\widehat{\mu}'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}\widehat{\mu} = \mu'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}\mu + \frac{\rho_n}{(1-\rho_n)^2}\mathfrak{s} + 2\widetilde{\mu}'(\mathcal{S} - \widehat{m}\widehat{m}')^{-2}\widehat{m} + o_P(1) \quad (7)$$

and we recall that $\widetilde{\mu}'(\mathcal{S} - \widehat{m}\widehat{m}')^{-2}\widehat{m}/\|\widetilde{\mu}\| = o_P(1)$.

The following remarks should help with the use of Equation (7) in practice. We can consider three cases, having to do with the size of $\mu'\Sigma^{-1}\mu = \|\widetilde{\mu}\|_2^2$.

1. If $\mu'\Sigma^{-1}\mu \rightarrow 0$, then, $\widehat{\mu}'\widehat{\Sigma}^{-1}\widehat{\mu} = \frac{\rho_n}{(1-\rho_n)^2}\mathfrak{s} + o_P(1)$.
2. If $\mu'\Sigma^{-1}\mu \rightarrow \infty$, then $\widehat{\mu}'\widehat{\Sigma}^{-1}\widehat{\mu} \sim \xi\mu'\Sigma^{-1}\mu$.
3. Finally, if $\mu'\Sigma^{-1}\mu$ stays bounded away from 0 and infinity,

$$\widehat{\mu}'\widehat{\Sigma}^{-1}\widehat{\mu} = \xi\mu'\Sigma^{-1}\mu + \frac{\rho}{(1-\rho)^2}\mathfrak{s} + o_P(1).$$

We note that when the λ_i 's are independent, (Assumption-BL) will be satisfied as soon as λ_i^2 has two moments, by the Marcinkiewicz-Zygmund strong law of large numbers (see Chow and Teicher (1997), p. 125). Since two moments are required for X_i to have a covariance, the existence of a second moment is also necessary for the population quantities to exist. (Assumption-BL) is here to guarantee that certain quadratic forms involving random projections are asymptotically deterministic.

4 Applications to computing the realized risk of portfolios

We now combine our results to reach conclusions about the realized risk of portfolios selected by solving the problem (QP-eqc-Emp).

As a matter of notation, all of our approximation statements hold with high-probability asymptotically, unless otherwise noted. We will carry out our work assuming that the data is generated from the models described in Section 2 and our theorems and under the following assumptions:

1. Assumption A0: $p/n \rightarrow \rho \in (0, 1)$, and the empirical distribution of the λ_i 's converge weakly in probability to a deterministic limit G .
2. Assumption A1: for all $i \in \{1, \dots, k\}$, $v_i'\Sigma^{-1}v_i$ stays bounded away from 0. v_k is assumed to be equal to μ .
3. Assumption A2: the smallest eigenvalue of $M = V'\Sigma^{-1}V$ stays bounded away from 0 and the condition number of M remains bounded. Also, the smallest entry of M in absolute value is bounded away from 0.
4. Assumption A3: if $\epsilon = \pm 1$, for all (i, j) , $(v_i + \epsilon v_j)'\Sigma^{-1}(v_i + \epsilon v_j)$ stay bounded away from infinity.
5. Assumption A4: (Assumption-BB) and (Assumption-BL) hold. (See Theorem 3.2 for definitions.)
6. Assumption A5: The operator norm of Σ , $\|\Sigma\|_2$, remains bounded.

We note that under these assumptions, the conclusions of our main theorems above are immediately applicable. In particular, A2 and A3 imply that Condition-P1 is satisfied. Condition A5 could be relaxed and will simply be needed for estimation purposes later. Also, the part of A2 concerning the smallest off-diagonal entry of M could also likely be relaxed.

Let us now take a moment to recall some key relevant results from El Karoui (2009b). Under similar assumptions as the ones we are now operating under, it was shown there that:

- $\mu'w_{\text{emp}}$, the realized returns of our portfolio, was not a consistent estimator of μ_P , the target returns for our portfolio. We proposed in El Karoui (2009b) an estimator of $\mu'w_{\text{emp}}$ which seems to perform well in (perhaps limited) simulations. The corrections we proposed there (or others) should be used if one wants to plot efficient frontier graphs that reflect the correct level of returns of portfolios.
- We showed in El Karoui (2009b) that $\mathfrak{s}^{(E)} \geq \mathfrak{s}^{(G)} = (1 - \rho)^{-1}$. In other words, the \mathfrak{s} corresponding to genuinely elliptical models is greater than the \mathfrak{s} corresponding to Gaussian models.

4.1 Theoretical predictions

We recall the notations $\rho = \lim p/n$ and $\kappa = \rho/(1 - \rho)$. Applying the results of our theorems above, we have the following fact:

Fact 4.1. *Let us call $M = V'\Sigma^{-1}V$. When Assumptions A0-A5 are satisfied, we have*

$$\begin{aligned}\widehat{V}'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}\widehat{V} &\simeq \xi V'\Sigma^{-1}V + \frac{\kappa}{1 - \rho} \mathfrak{s} e_k e_k', \\ \widehat{M} &= \widehat{V}'\widehat{\Sigma}^{-1}\widehat{V} \simeq \mathfrak{s} V'\Sigma^{-1}V + \kappa e_k e_k' .\end{aligned}$$

Also, under our assumptions, as shown in El Karoui (2009b),

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

Before we proceed, let us also recall that

$$\frac{\xi}{\mathfrak{s}^2} \geq \frac{1}{1 - \rho} ,$$

and that in the Gaussian case, $\xi = (1 - \rho)^{-3}$ and $\mathfrak{s} = (1 - \rho)^{-1}$. So the right hand side of the previous inequality is achieved in the Gaussian setting.

Fact 4.1 allows us to give the following characterization of the realized risk of ‘‘Markowitz’’ portfolios.

Theorem 4.1. *Recall that in matrix form, the optimal risk for (QP-eqc-Pop), $w'_{\text{theo}}\Sigma w_{\text{theo}}$ is equal to $U'M^{-1}U$.*

Suppose that a portfolio is chosen by allocating weights w_{emp} to each asset according to the solution of (QP-eqc-Emp). Under assumptions A0 – A5, we have

$$w'_{\text{emp}}\Sigma w_{\text{emp}} \simeq \frac{1}{\mathfrak{s}^2} \left(\xi U'(M + \frac{\kappa}{\mathfrak{s}} e_k e_k')^{-1} U + \frac{\kappa}{\mathfrak{s}} \left(\frac{\mathfrak{s}^2}{1 - \rho} - \xi \right) \left[U'(M + \frac{\kappa}{\mathfrak{s}} e_k e_k')^{-1} e_k \right]^2 \right) . \quad (8)$$

Furthermore, in the situation where μ is assumed to be known or equivalently, if all the elements of V are deterministic and given,

$$w'_{\text{emp}}\Sigma w_{\text{emp}} \simeq \frac{\xi}{\mathfrak{s}^2} U'M^{-1}U \simeq \frac{\xi}{\mathfrak{s}^2} w'_{\text{theo}}\Sigma w_{\text{theo}} .$$

Before we give the proof, which is now just linear algebra, we remind the reader that $\xi \geq \mathfrak{s}^2/(1 - \rho)$, with equality in the Gaussian case. So the second term in the approximation to $w'_{\text{emp}}\Sigma w_{\text{emp}}$ is negative, except in the Gaussian case where it is zero.

Furthermore, in cases where μ does not need to be estimated, $w'_{\text{emp}}\Sigma w_{\text{emp}}$ is at least $1/(1 - \rho)$ times as large as the population optimum $w'_{\text{theo}}\Sigma w_{\text{theo}}$, the coefficient $1/(1 - \rho)$ corresponding to the Gaussian case.

We also recall one of the main results of El Karoui (2009b): under the same assumptions as those of Theorem 4.1, we have, for the (extremely) naive estimate of risk

$$w'_{\text{emp}}\widehat{\Sigma} w_{\text{emp}} \simeq \frac{1}{\mathfrak{s}} U' \left(M + \frac{\kappa}{\mathfrak{s}} e_k e_k' \right)^{-1} U .$$

Furthermore, if μ does not need to be estimated, we then have $w'_{\text{emp}}\widehat{\Sigma} w_{\text{emp}} \simeq \frac{1}{\mathfrak{s}} w'_{\text{theo}}\Sigma w_{\text{theo}}$.

These remarks combined with Theorem 4.1 allow to us quantify the difference between the naive estimate of risk of our portfolio and its realized risk. We will later discuss an estimator of this realized risk.

Proof. Let us call $\widehat{N} = \widehat{V}'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}\widehat{V}$. The analyses we performed show that

$$\begin{aligned}\widehat{N} &\simeq \xi M + \frac{\kappa}{1-\rho} \mathfrak{s} e_k e_k' \\ &\simeq \xi \left(M + \frac{\kappa}{\mathfrak{s}} \right) + \frac{\kappa}{\mathfrak{s}} \left(\frac{\mathfrak{s}^2}{1-\rho} - \xi \right) e_k e_k'\end{aligned}$$

Therefore, since $\widehat{M}^{-1} \simeq 1/\mathfrak{s} \left(M + \frac{\kappa}{\mathfrak{s}} e_k e_k' \right)^{-1}$, we have

$$\widehat{M}^{-1} \widehat{N} \simeq \frac{1}{\mathfrak{s}} \left(\xi \text{Id}_p + \frac{\kappa}{\mathfrak{s}} \left(\frac{\mathfrak{s}^2}{1-\rho} - \xi \right) \left(M + \frac{\kappa}{\mathfrak{s}} e_k e_k' \right)^{-1} e_k e_k' \right).$$

Consequently, we also have

$$\widehat{M}^{-1} \widehat{N} \widehat{M}^{-1} \simeq \frac{1}{\mathfrak{s}^2} \left(\xi \left(M + \frac{\kappa}{\mathfrak{s}} e_k e_k' \right)^{-1} + \frac{\kappa}{\mathfrak{s}} \left(\frac{\mathfrak{s}^2}{1-\rho} - \xi \right) \left(M + \frac{\kappa}{\mathfrak{s}} e_k e_k' \right)^{-1} e_k e_k' \left(M + \frac{\kappa}{\mathfrak{s}} e_k e_k' \right)^{-1} \right).$$

Since $w'_{\text{emp}} \Sigma w_{\text{emp}} = U' \widehat{M}^{-1} \widehat{N} \widehat{M}^{-1} U$, we finally conclude that

$$w'_{\text{emp}} \Sigma w_{\text{emp}} \simeq \frac{1}{\mathfrak{s}^2} \left(\xi U' \left(M + \frac{\kappa}{\mathfrak{s}} e_k e_k' \right)^{-1} U + \frac{\kappa}{\mathfrak{s}} \left(\frac{\mathfrak{s}^2}{1-\rho} - \xi \right) \left[U' \left(M + \frac{\kappa}{\mathfrak{s}} e_k e_k' \right)^{-1} e_k \right]^2 \right).$$

This proves Equation (8).

Let us now turn to the second part of the theorem. When $\widehat{V} = V$, $\widehat{M} = V' \widehat{\Sigma}^{-1} V \simeq \mathfrak{s} M$. Also, $\widehat{N} = V' \widehat{\Sigma}^{-1} \Sigma \widehat{\Sigma}^{-1} V \simeq \xi M$. Therefore, in this situation where μ does not need to be estimated,

$$w'_{\text{emp}} \Sigma w_{\text{emp}} \simeq \frac{\xi}{\mathfrak{s}^2} U' M^{-1} U.$$

Now $U' M^{-1} U$ is equal to $w'_{\text{theo}} \Sigma w_{\text{theo}}$, the population optimum and efficient frontier. \square

Theorem 4.2 (Comparison Gaussian-Elliptical Case). *The realized risk of a portfolio computed by solving (QP-eqc-Emp) where the data is elliptical is greater than that of its Gaussian counterpart.*

If $w_{\text{emp}}^{(E)}$ is the solution of (QP-eqc-Emp) when X_i are elliptical and $w_{\text{emp}}^{(G)}$ is the solution when X_i are Gaussian,

$$(w_{\text{emp}}^{(E)})' \Sigma w_{\text{emp}}^{(E)} \geq (w_{\text{emp}}^{(G)})' \Sigma w_{\text{emp}}^{(G)}.$$

The inequality is strict if X_i is genuinely elliptical and not Gaussian (i.e $\lambda_i \neq 1$ with strictly positive probability).

It is interesting to compare this theorem to its counterpart in El Karoui (2009b), Theorem 5.1 there. That theorem shows that the (very) naive estimator of risk, $(w_{\text{emp}}^{(E)})' \widehat{\Sigma} w_{\text{emp}}^{(E)}$ underestimates the true risk, and that this underestimation is more pronounced in the elliptical case than in the Gaussian case. So we conclude that

$$\frac{(w_{\text{emp}}^{(E)})' \Sigma w_{\text{emp}}^{(E)}}{(w_{\text{emp}}^{(E)})' \widehat{\Sigma} w_{\text{emp}}^{(E)}} \geq \frac{(w_{\text{emp}}^{(G)})' \Sigma w_{\text{emp}}^{(G)}}{(w_{\text{emp}}^{(G)})' \widehat{\Sigma} w_{\text{emp}}^{(G)}} \simeq \frac{1}{(1-\rho)^2}.$$

We note that the analysis presented in the proof below actually shows that

$$(w_{\text{emp}}^{(E)})' \Sigma w_{\text{emp}}^{(E)} \geq \frac{1}{1-\rho} U' \left(M + \frac{\kappa}{\mathfrak{s}} e_k e_k' \right)^{-1} U,$$

and therefore, asymptotically and with high-probability,

$$\frac{(w_{\text{emp}}^{(E)})' \Sigma w_{\text{emp}}^{(E)}}{(w_{\text{emp}}^{(E)})' \widehat{\Sigma} w_{\text{emp}}^{(E)}} \geq \frac{\mathfrak{s}}{1-\rho}.$$

Since $\mathfrak{s} \geq 1/(1-\rho)$, with equality in the Gaussian case, this inequality gives a sharper notion of the impact of ellipticity on risk underestimation.

Finally, let us say that

$$\frac{(w_{\text{emp}}^{(E)})' \Sigma w_{\text{emp}}^{(E)}}{(w_{\text{emp}}^{(E)})' \widehat{\Sigma} w_{\text{emp}}^{(E)}}$$

is a measure of how accurate the (very) naive estimator of risk $(w_{\text{emp}}^{(E)})' \widehat{\Sigma} w_{\text{emp}}^{(E)}$ is at predicting the actual risk of our strategy (in the setting of i.i.d data), $(w_{\text{emp}}^{(E)})' \Sigma w_{\text{emp}}^{(E)}$. What our computations show is that it is never terribly accurate and it is least inaccurate in the Gaussian case. This also suggests that doing corrections or predictions based on Gaussian computations will yield poor results (and still risk underestimation!) in the class of elliptical distributions considered here.

We now turn to the proof.

Proof of Theorem 4.2 : Recall that $\mathfrak{s}^{(E)} \geq \mathfrak{s}^{(G)}$. Recall also that $\xi^{(E)}/(\mathfrak{s}^{(E)})^2 \geq 1/(1-\rho)$. Finally, our theorems show that

$$\widehat{N} = \widehat{V}' \widehat{\Sigma}^{-1} \Sigma \widehat{\Sigma}^{-1} \widehat{V} \simeq \xi M + \frac{\kappa}{1-\rho} \mathfrak{s} e_k e_k'.$$

Therefore,

$$\frac{1}{\mathfrak{s}^2} \widehat{N} \simeq \frac{\xi}{\mathfrak{s}^2} M + \frac{\kappa}{1-\rho} \frac{1}{\mathfrak{s}} e_k e_k' \succeq \frac{1}{1-\rho} \left(M + \frac{\kappa}{\mathfrak{s}} e_k e_k' \right)$$

Now, $\widehat{M} = \widehat{V}' \widehat{\Sigma}^{-1} \widehat{V} \simeq \mathfrak{s} \left(M + \frac{\kappa}{\mathfrak{s}} e_k e_k' \right)$. So

$$\widehat{M}^{-1} \widehat{N} \widehat{M}^{-1} \simeq \left(M + \frac{\kappa}{\mathfrak{s}} e_k e_k' \right)^{-1} \left[\frac{1}{\mathfrak{s}^2} \widehat{N} \right] \left(M + \frac{\kappa}{\mathfrak{s}} e_k e_k' \right)^{-1}.$$

Therefore, asymptotically with high-probability,

$$\widehat{M}^{-1} \widehat{N} \widehat{M}^{-1} \succeq \left(M + \frac{\kappa}{\mathfrak{s}} e_k e_k' \right)^{-1} \left[\frac{1}{1-\rho} \left(M + \frac{\kappa}{\mathfrak{s}} e_k e_k' \right) \right] \left(M + \frac{\kappa}{\mathfrak{s}} e_k e_k' \right)^{-1} = \frac{1}{1-\rho} \left(M + \frac{\kappa}{\mathfrak{s}} e_k e_k' \right)^{-1}.$$

Now, $M + \frac{\kappa}{\mathfrak{s}^{(E)}} e_k e_k' \preceq M + \frac{\kappa}{\mathfrak{s}^{(G)}} e_k e_k'$, so

$$\left(M + \frac{\kappa}{\mathfrak{s}^{(E)}} e_k e_k' \right)^{-1} \succeq \left(M + \frac{\kappa}{\mathfrak{s}^{(G)}} e_k e_k' \right)^{-1}.$$

We conclude that asymptotically with high-probability,

$$\widehat{M}^{-1} \widehat{N} \widehat{M}^{-1} \succeq \frac{1}{1-\rho} \left(M + \frac{\kappa}{\mathfrak{s}^{(G)}} e_k e_k' \right)^{-1}.$$

Now recall that in the Gaussian case, $\xi^{(G)}/(\mathfrak{s}^{(G)})^2 = 1/(1-\rho)$, so that

$$\left(\widehat{M}^{-1} \widehat{N} \widehat{M}^{-1} \right)^{(G)} \simeq \frac{1}{1-\rho} \left(M + \frac{\kappa}{\mathfrak{s}^{(G)}} e_k e_k' \right)^{-1}$$

Since $w_{\text{emp}}' \Sigma w_{\text{emp}} = U' \widehat{M}^{-1} \widehat{N} \widehat{M}^{-1} U$, we conclude that, asymptotically with high-probability

$$(w_{\text{emp}}^{(E)})' \Sigma w_{\text{emp}}^{(E)} \geq (w_{\text{emp}}^{(G)})' \Sigma w_{\text{emp}}^{(G)}.$$

The proof shows that in the case where X_i are genuinely elliptical, since $\mathfrak{s}^{(E)} > \mathfrak{s}^{(G)}$, we have a strict inequality in the conclusion. \square

4.2 Improved estimation of the realized risk

We now seek a robust estimator - in the class of elliptical distributions considered here - for the realized risk of our Markowitz portfolios. From a practical standpoint, it could be useful to help assess the actual risk of a portfolio constructed using all the data available, i.e the n observations $\{X_i\}_{i=1}^n$. Naturally, statistically, one might want to use techniques like cross-validation to make this assessment empirically, but this would reduce the effective number of samples of the procedure and hence yield even less optimal allocations than the ones we could get by using all the data.

For the purposes of the discussion that follows, we now assume that the λ_i 's are independent.

Recall that Equation (8) showed that

$$\text{RRisk} \simeq \frac{1}{\mathfrak{s}^2} \left(\xi U'(M + \frac{\kappa}{\mathfrak{s}} e_k e_k')^{-1} U + \frac{\kappa}{\mathfrak{s}} \left(\frac{\mathfrak{s}^2}{1 - \rho} - \xi \right) \left[U'(M + \frac{\kappa}{\mathfrak{s}} e_k e_k')^{-1} e_k \right]^2 \right) .$$

Furthermore, $\widehat{M}^{-1} \simeq \frac{1}{\mathfrak{s}} (M + \frac{\kappa}{\mathfrak{s}} e_k e_k')^{-1}$, so it turns out that

$$\text{RRisk} \simeq \frac{\xi}{\mathfrak{s}} U' \widehat{M}^{-1} U + \frac{\kappa}{\mathfrak{s}} \left(\frac{\mathfrak{s}^2}{1 - \rho} - \xi \right) \left[U' \widehat{M}^{-1} e_k \right] .$$

Now we recall that in El Karoui (2009b), we proposed an estimator of \mathfrak{s} and the λ_i^2 's: our proposal, motivated by concentration of measure results for Gaussian random vectors (see Ledoux (2001)) was to

1. Estimate λ_i^2 by

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

2. If we denote $\rho_n = p/n$, we then proposed to estimate \mathfrak{s} by $\widehat{\mathfrak{s}}$, the positive solution of

$$g(x) = 1 - \rho_n ,$$

where $g(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{1 + x \widehat{\lambda}_i^2 \rho_n} .$

In light of Equation (5), we propose to estimate ξ by $\widehat{\xi}$ with

$$\frac{1}{\widehat{\xi}} = \frac{1}{\widehat{\mathfrak{s}}^2} - \rho_n \frac{1}{n} \sum_{i=1}^n \frac{\widehat{\tau}_i^2}{(1 + \widehat{\tau}_i \rho_n \widehat{\mathfrak{s}})^2} . \quad (9)$$

Note that $U' \widehat{M}^{-1} U$ is the plug-in (and very naive) estimator of risk. Let us denote it by f_{emp} . We propose to use

$$\widehat{\text{RRisk}} = \frac{\widehat{\xi}}{\widehat{\mathfrak{s}}} f_{\text{emp}} + \frac{\kappa}{\widehat{\mathfrak{s}}} \left(\frac{\widehat{\mathfrak{s}}^2}{1 - \rho_n} - \widehat{\xi} \right) \left[U' \widehat{M}^{-1} e_k \right] . \quad (10)$$

The simulation work that follows illustrate the performance of this estimator.

4.3 Simulation results

To investigate the quality of our proposed estimator in practice, we now present some simulations. We compare two situations, one where the data generated are normally distributed and one where they have a distribution close to a multivariate t_6 distribution. In this latter case, the λ_i 's are i.i.d and have a univariate t_6 distribution, scaled to have second moment equal to 1. We note that the t_6 distribution has only 5 moments, so the corresponding X_i 's also have only 5 moments and are much more heavy-tailed than in the situation where they have a Gaussian distribution.

Before we present pictures of our simulations, let us briefly summarize our findings. The estimator presented in Equation (10) seems to work quite or reasonably well in expectation, when n and p are in the low 100's. However, in this situation, the variance is still quite large. We also note that the "true" realized

risk of our portfolios, which is itself a conditional variance, and hence vary from one simulation to another, is also quite variable. Furthermore, in this situation, the difference in realized risks between the Gaussian and the t_6 cases are considerable, as our theoretical results suggest. We make this remark to emphasize how assuming Gaussianity would lead to vastly over-optimistic conclusions.

When we increase dimensionality to the low 1000's for both p and n (something that is probably unrealistic at this point in financial applications but might be relevant in other areas of applications), the variance issue becomes less important or significant (as we would expect since we are closer to the limiting setting and therefore our predictions should be more accurate), and the quality of our predictions of realized risk is even better. Of course, the advantage of having an estimator that is robust in the class of elliptical distributions is that we do not need to specify which distribution likely generated the data - as long as it is elliptical, we will be fine. We note that again here the difference between the realized risks in the Gaussian and t_6 cases is very large.

We present simulations in the situation where Σ is a Toeplitz matrix with $\sigma(i, j) = .4^{|i-j|}$. Two dimensionality settings are investigated. In the first case, illustrated in Figure 2, p. 34, $n = 250$ and $p = 100$. In the second case, illustrated in Figure 3, p. 35, $n = 2500$, $p = 1000$. The first constraint vector, v_1 , was chosen to be the eigenvector associated with the $.9 * p$ largest eigenvalue of Σ . We also relied on the vector v_2 , chosen to be eigenvector associated with the $.15 * p$ largest eigenvalue of Σ . The second constraint vector, which played the role of μ , was chosen as $\sqrt{.3}v_1 + \sqrt{.7}v_2$.

We also present simulations that are closer to real data. In this second set of simulations, we took the daily returns of 48 Fama-French industry portfolios, for the year 2005. We computed the corresponding sample mean and sample covariance matrix, and took them as our new population parameters. From them, as in a parametric bootstrap, we generated 1000 datasets, with $n = 252$ and $p = 48$, under the Gaussian and t_6 models. In this situation, the matrix M had a relatively large condition number, equal to 40, and the setting is more difficult for our estimators. As we show in Figure 4, p. 36, the variability of our estimator is quite large and the average performance is not as good (the performance of the medians are quite similar) as the one we observed on the other synthetic problems. The smaller p and relatively large condition number of M might explain some of these problems.

5 Conclusion

We have analyzed in this paper the impact of high-dimensionality and ellipticity of the data vectors on the risk of Markowitz portfolios obtained by seeking the solution of the generalized Markowitz Problem (QP-eqc-Emp), a quadratic program with linear equality constraints. One of our main result is that we have provided an estimator of this realized risk that is robust in the class of elliptical distributions and appears to work reasonably well in practice, in the limited simulations we have investigated. An interesting by-product of our results is that they show that there is no “universality” in the random matrix sense of the word for this problem. The details of the model do matter and we cannot limit ourselves to specifying the population parameters if we want to get a general and robust answer. Our results therefore suggest that claims of universality found in the physics literature are unfounded.

Elliptical models are a relatively rich class of models and allow us to incorporate in the modeling features of financial data that are often found in practice. For instance, our models have leptokurtic marginals, can have heavy-tails (only two moments are needed for our results to be valid in the i.i.d case) and tail-dependence. However, it seems to us that what drives the key results are global geometric features of the data and not details about the marginals. These global geometric features play an important role in random matrix theory (see El Karoui (2009a)) and it is not surprising that they play a key role here. We expect that the results obtained in our paper can be shown to be valid under weaker and less restrictive distributional assumptions, something we are currently working on, as long as the geometry of the data implied by the model is conserved.

Interestingly, we show that both estimation of the mean and the covariance of the data vectors is important and creates its share of problems. Both yield first order effects and biases which cannot be ignored. Naturally, if extra information is available and used, these problems could be partially alleviated by, for instance, shrinking the sample mean and covariance matrix toward appropriate targets. But an analysis that takes into account both sources of problems is indeed necessary.

We note that our work could also be adapted to study the performance of portfolios obtained by using weighted estimators of covariance or by bootstrapping. As shown in El Karoui (2009b) for the bootstrap, these problems are essentially covered by the elliptical framework. We note that the more complicated aspects of the issue come from having to understand $\widehat{\mu}'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}\widehat{\mu}$, which in general will behave differently than it did in the situations we have considered in this paper (see the bootstrap analysis in El Karoui (2009b) for an example in a different but related problem). The possible choice of different weights for the estimation of the mean and the covariance also complicates the situation, though the tools used in this paper seem suitable for analyzing this problem. (We postpone its analysis to another paper, as this one is already quite long.) On the other hand, we note that forms of the type $v'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}v$ should behave as we have described in this paper, after we properly account from the “ellipticity” generated by weighting differently different observations.

Random matrix theory appears to be a convenient and valuable tool for the study of the optimization problem we were concerned with. It has helped shed some light on an otherwise quite difficult problem and might be helpful in the analysis of other optimization problems that are sensitive to spectral properties of the input data.

APPENDIX

Recall that the two main quantities of technical interest in this paper are $v'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}v$ and $\widehat{\mu}'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}\widehat{\mu}$.

If $Y_1 = \Sigma_1 Y$, where Y is an $n \times p$ matrix with i.i.d $\mathcal{N}(0, 1)$ entries, understanding $v'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}v$ is equivalent to understanding $\alpha' \left(\frac{Y_1' H Y_1}{n-1} \right)^{-2} \alpha$, for an appropriately chosen α . This question is the focus of the following section.

A On $\alpha' \left(\frac{Y_1' H Y_1}{n-1} \right)^{-2} \alpha$

The analysis of this quantity is closely connected the work that was done in El Karoui (2009b), Theorem 4.1. We have the following theorem.

Theorem A.1. *Suppose we observe n i.i.d observations X_i , where X_i has the form $X_i = \mu + \lambda_i Y_i$, with $Y_i \stackrel{iid}{\sim} \mathcal{N}(0, \text{Id}_p)$ and $\{\lambda_i\}_{i=1}^n$ is independent of $\{Y_i\}_{i=1}^n$. We assume that $\mathbf{E}(\lambda_i^2) = 1$. We call X the $n \times p$ data matrix containing the X_i 's, which are the rows of X .*

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

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

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

Under these assumptions, if α is a (sequence of) deterministic vectors with norm 1,

$$\alpha' \left(\frac{X' H X}{n-1} \right)^{-2} \alpha \rightarrow \xi$$

where if \mathfrak{s} satisfies,

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

ξ is defined as

$$\xi = \frac{1}{\frac{1}{\mathfrak{s}^2} - \rho \int \frac{\tau^2 dG(\tau)}{(1 + \tau\rho\mathfrak{s})^2}}. \quad (\text{A-2})$$

Let us call $P1$ the following condition on the population parameters:

$$\forall 1 \leq i \neq j \leq k, \frac{v_i' \Sigma^{-1} v_j}{(v_i + v_j)' \Sigma^{-1} (v_i + v_j)} \text{ and } \frac{v_i' \Sigma^{-1} v_j}{(v_i - v_j)' \Sigma^{-1} (v_i - v_j)} \text{ stay bounded away from } 0. \quad (\text{Condition-P1})$$

As an immediate corollary, we have

Corollary A-1. *With the definitions above, and if V and Σ are such that Condition-P1 is satisfied, and if $M = V' \Sigma^{-1} V$, we have asymptotically*

$$V' \widehat{\Sigma}^{-1} \Sigma \widehat{\Sigma}^{-1} V = \xi V' \Sigma^{-1} V + o_P(V' \Sigma^{-1} V) = \xi M + o_P(M).$$

The proof of the corollary follows exactly the same steps as the proof of Lemma 4.1 in El Karoui (2009b).

We also have the following fact:

Fact A.1. *We have*

$$\xi \geq \frac{\mathfrak{s}^2}{1 - \rho}.$$

This latter value corresponds to the Gaussian case.

Proof of the fact. Let us consider the integral

$$I = \int \frac{\rho^2 \tau^2 \mathfrak{s}^2}{(1 + \tau \rho \mathfrak{s})^2} dG(\tau).$$

By writing $\rho \tau \mathfrak{s} / (1 + \rho \tau \mathfrak{s}) = (1 - 1/(1 + \rho \tau \mathfrak{s}))$, we see that

$$I = 1 - 2 \int \frac{dG(\tau)}{1 + \rho \tau \mathfrak{s}} + \int \frac{dG(\tau)}{(1 + \rho \tau \mathfrak{s})^2}.$$

Now by definition of \mathfrak{s} , $\int \frac{dG(\tau)}{1 + \rho \tau \mathfrak{s}} = (1 - \rho)$. On the other hand, by convexity of the square function, we have

$$\int \frac{dG(\tau)}{(1 + \rho \tau \mathfrak{s})^2} \geq \left(\int \frac{dG(\tau)}{1 + \rho \tau \mathfrak{s}} \right)^2 = (1 - \rho)^2.$$

Therefore,

$$I \geq 1 - 2(1 - \rho) + (1 - \rho)^2 = (1 - (1 - \rho))^2 = \rho^2,$$

or

$$\int \frac{\tau^2 \mathfrak{s}^2}{(1 + \rho \tau \mathfrak{s})^2} dG(\tau) \geq 1.$$

Therefore,

$$\frac{1}{\mathfrak{s}^2} \left(1 - \rho \int \frac{\tau^2 \mathfrak{s}^2 dG(\tau)}{(1 + \rho \tau \mathfrak{s})^2} \right) \leq \frac{1 - \rho}{\mathfrak{s}^2}.$$

Now we know that $\xi \geq 0$ by construction (see below). So we conclude that

$$\xi \geq \frac{\mathfrak{s}^2}{(1 - \rho)}.$$

□

Proof of Theorem A.1. Our proof follows closely the proof of Theorem 4.1 in El Karoui (2009b). The matrix X can be written as

$$X = \mathbf{e}\mu' + \Lambda Y ,$$

where Y is an $n \times p$ data matrix having the Y_i 's as its rows and hence i.i.d $\mathcal{N}(0, 1)$ entries under our assumptions, and Λ is the diagonal matrix containing the λ_i 's. Therefore,

$$\mathcal{S} = \frac{1}{n-1} X' H X = \frac{1}{n-1} Y' \Lambda' H \Lambda Y \triangleq \frac{1}{n-1} Y' L Y ,$$

where $L = \Lambda' H \Lambda$. It has been argued in El Karoui (2009b) that $Y' L Y$ is invertible with probability 1 under our assumptions. Note that the quantity we care about is

$$\alpha' \mathcal{S}^{-2} \alpha .$$

We will first get results conditional on Λ and then will argue that we can de-condition and get results unconditionally. We call γ_i the ordered eigenvalues of \mathcal{S} , γ_p being the smallest.

Results conditionally on Λ Let us call $\mathcal{L}_{\epsilon, \delta}$ the set of matrices Λ such that $p/N < 1 - \epsilon$ and $C(N-1)/(n-1) > \delta$. Under (Assumption-BB), for a δ bounded away from 0 (e.g $\delta = 1/2 \liminf C_0 N/n$), $P(\Lambda \in \mathcal{L}_{\epsilon, \delta}) \rightarrow 1$. We assume that the Λ we condition on below belongs to $\mathcal{L}_{\epsilon, \delta}$. In such situation, one can show (see Lemma B-1 in El Karoui (2009b)) that, if P_Λ denotes probability conditional on Λ , and if $\Lambda \in \mathcal{L}_{\epsilon, \delta}$,

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

In other words, the smallest eigenvalue of \mathcal{S} is uniformly bounded away from 0 with very high (conditional) probability. We will see that these uniform bounds on the smallest eigenvalue of \mathcal{S} will eventually allow us to go from the conditional results on Λ to unconditional ones.

Let us write the spectral decomposition of \mathcal{S} :

$$\mathcal{S} = \sum_{i=1}^p \gamma_i v_i v_i' .$$

As was explained in the proof of Theorem 4.1 in El Karoui (2009b), the eigenvalues and eigenvectors of \mathcal{S} are independent, and the matrix of eigenvectors is uniformly distributed on the orthogonal group (after taking proper care of sign indeterminacy - see Chikuse (2003), p.40). Therefore, the random variable we care about can be written as

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

Going through the proof of Theorem 4.1 in El Karoui (2009b), we see that if we consider, for a given function h , the random variable

$$Z_h = \sum_{i=1}^p h(\gamma_i) (\alpha' v_i)^2 ,$$

it is shown there that

$$\text{var}(Z_n | \gamma_i, \Lambda) \leq C \frac{1}{p^2} \sum_{i=1}^p (h(\gamma_i))^2 .$$

In our case here, $h(x) = x^{-2}$. Under our assumptions, according to Lemma B-1 in El Karoui (2009b), we have $\gamma_i^2 \geq \mathfrak{C}_n (1 - \sqrt{p/(N-1)})^2 / 2$, where $\mathfrak{C}_n = C_0(N-1)/(n-1)$, with high ($\{Y_i\}_{i=1}^n$)-probability, so we conclude that

$$\text{var}(\alpha' \mathcal{S}^{-2} \alpha | \{\gamma_i\}, \Lambda) \rightarrow 0 .$$

Since (see e.g El Karoui (2009b))

$$\mathbf{E}(\alpha' \mathcal{S}^{-2} \alpha | \{\gamma_i\}, \Lambda) = \frac{1}{p} \sum_{i=1}^p \frac{1}{\gamma_i^2} ,$$

we conclude that

$$\alpha' \mathcal{S}^{-2} \alpha - \frac{1}{p} \sum_{i=1}^p \frac{1}{\gamma_i^2} \Big| \{\gamma_i\}, \Lambda \rightarrow 0 \text{ in probability .}$$

The same arguments as those used in the proof of Theorem 4.1 of El Karoui (2009b) then show that the same result can be obtained conditionally on Λ only.

Identifying the limit It is now clear that at least conditionally on Λ , our problem reduces to understanding the limit of

$$\frac{1}{p} \sum_{i=1}^p \frac{1}{\gamma_i^2} .$$

The Stieltjes transform (see e.g Bai (1999)) of the spectral distribution of \mathcal{S} is

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

As was explained in El Karoui (2009b), Theorem 4.1, according to results of Marčenko and Pastur (1967), Wachter (1978) and Silverstein (1995), for any fixed $z \in \mathbb{C}^+$, $s_p(z) \rightarrow s(z)$ in probability, where $s(z)$ satisfies if G is the limiting spectral distribution of $L = \Lambda' H \Lambda$,

$$-\frac{1}{s(z)} = z - \int \frac{\tau dG(\tau)}{1 + \tau \rho s(z)} . \quad (\text{A-3})$$

Note that under our assumptions, all the L 's have the same limiting spectral distribution, G . So the result does not depend asymptotically on the sequence of Λ 's we are conditioning on. Because we know that, given $\Lambda \in \mathcal{L}_{\epsilon, \delta}$, the smallest eigenvalue of \mathcal{S} is asymptotically bounded away from 0, and hence the limiting spectral distribution of \mathcal{S} , \mathcal{K} , has support bounded away from 0, we know that s is analytic in a neighborhood of zero. Also, because the pointwise convergence of Stieltjes transforms imply weak convergence of spectral distributions, we see that, if \mathcal{K} is the limiting spectral distribution of \mathcal{S} , by taking $m_\eta(x) = \inf(1/\eta^2, 1/x^2)$ as a test function, for any given η smaller than the left endpoint of the support of \mathcal{K} ,

$$\frac{1}{p} \sum_{i=1}^p m_\eta(\gamma_i) \rightarrow \int m_\eta(x) d\mathcal{K}(x) = \int \frac{1}{x^2} d\mathcal{K}(x) \text{ in probability .}$$

Because the smallest γ_i is bounded away from 0 with high-probability, we also see that if η is small enough, and p and n are large enough,

$$\frac{1}{p} \sum_{i=1}^p m_\eta(\gamma_i) = \frac{1}{p} \sum_{i=1}^p \frac{1}{\gamma_i^2} \text{ with high probability.}$$

Now, because s is analytic in a neighborhood of 0, we have

$$s'(0) = \int \frac{1}{x^2} d\mathcal{K}(x) ,$$

and we have finally established that

$$\frac{1}{p} \sum_{i=1}^p \frac{1}{\gamma_i^2} \rightarrow s'(0) \text{ in probability ,}$$

conditionally on any Λ belonging to $\mathcal{L}_{\epsilon, \delta}$.

Now, according to Equation (A-3) and using the fact that $s(z)$ is analytic in a neighborhood of 0, we have, for z in a neighborhood of zero,

$$\frac{s'(z)}{s^2(z)} = 1 + \rho s'(z) \int \frac{\tau^2 dG(\tau)}{(1 + \rho \tau s(z))^2} .$$

From this equation, and the fact that we know that $s(0) \neq 0$ and is finite, we conclude that $s'(0) \neq 0$, for otherwise we would have $0 = 1$. We finally obtain

$$\frac{1}{s'(0)} = \frac{1}{s^2(0)} - \rho \int \frac{\tau^2 dG(\tau)}{(1 + \rho\tau s(0))^2}.$$

Note that as seen in Theorem 4.1 of El Karoui (2009b), $s(0) = \mathfrak{s}$. We see that $s'(0)$ is the value of ξ announced above. Also, since $s'(0) = \lim \sum \gamma_i^{-2}/p$, $s'(0) \geq 0$, and therefore $\xi \geq 0$.

Getting results unconditionally on Λ So far we have worked with matrices Λ belonging to $\mathcal{L}_{\epsilon, \delta}$. Following exactly the de-conditioning arguments given at the end of the proof of Theorem 4.1 in El Karoui (2009b), we see that the results hold also unconditionally on Λ under the assumptions of the theorem. The theorem is shown. \square

B Quadratic forms in $\hat{\mu}$ and $\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}$

In this section, we will use the definition $\hat{\Sigma} = X'HX/n$, if X is the data matrix. This mild change of scaling has no asymptotic consequences but makes the notation less cumbersome in our proofs and theorems.

B-1 Understanding $\hat{\mu}'\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\mu}$, when $\mu = 0$

The aim of this and the following subsections is to show the following theorem.

Theorem B.1. *Suppose Y is an $n \times p$ matrix whose rows are the vectors Y_i , which are i.i.d $\mathcal{N}(0, \text{Id}_p)$.*

Suppose Λ is a diagonal matrix whose i -th entry is λ_i , which is possibly random and is independent of Y . 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 and

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

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

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

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

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

Then we have

$$\zeta_{n,p} \rightarrow \rho \mathfrak{s}, \text{ in probability.}$$

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

$$\hat{m}' \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} \hat{m} \rightarrow \frac{\kappa}{1 - \rho} \mathfrak{s} \text{ in probability.}$$

We note that under our assumptions $\zeta_{n,p}$ will exist with probability 1, since the λ_i 's are all different from 0 and the Y_i 's have a continuous distribution.

B-1.1 Linear algebraic preliminaries

Before we deal with the central issues of this problem, we need the following preliminary lemma.

Lemma B-1. *If $X = Y_1 \Sigma^{1/2}$ and $\hat{\mu} = X' \mathbf{e}/n$, then*

$$Z_{n,p} = \hat{\mu}' \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} \hat{\mu} = \frac{\mathbf{e}' Y_1}{n} \left(\frac{Y_1' H Y_1}{n} \right)^{-2} \frac{Y_1' \mathbf{e}}{n},$$

and is therefore independent of Σ . Further, if M is an invertible matrix and u is a vector such that $u' M^{-1} u \neq 1$, we have

$$u' (M - uu')^{-2} u = \frac{u' M^{-2} u}{(1 - u' M^{-1} u)^2}.$$

In particular, if $\hat{\mu}_{Y_1} = Y_1' \mathbf{e}/n$, $\mathcal{S} = Y_1' Y_1/n$, $\hat{\Sigma}_{Y_1} = \mathcal{S} - \hat{\mu}_{Y_1} \hat{\mu}_{Y_1}' = Y_1' H Y_1/n$, we have

$$Z_{n,p} = \hat{\mu}_{Y_1}' \hat{\Sigma}_{Y_1}^{-2} \hat{\mu}_{Y_1} = \frac{\hat{\mu}_{Y_1}' \mathcal{S}^{-2} \hat{\mu}_{Y_1}}{(1 - \hat{\mu}_{Y_1}' \mathcal{S}^{-1} \hat{\mu}_{Y_1})^2}.$$

This lemma shows that the problems we are considering do not involve Σ when $\mu = 0$ and that therefore we can assume that $\Sigma = \text{Id}_p$ without loss of generality in our analysis. The central object of our study will be the random variable

$$\zeta_{n,p} = \hat{\mu}'_W \mathcal{S}^{-2} \hat{\mu}_W = \frac{\mathbf{e}' W}{n} \left(\frac{W' W}{n} \right)^{-2} \frac{W' \mathbf{e}}{n}, \quad (\text{B-4})$$

where $W = \Lambda Y$, and Y is a random matrix whose entries are i.i.d $\mathcal{N}(0, 1)$ and Λ is a diagonal random matrix whose entries are independent of Y .

Under the assumptions of Theorem B.1, we showed in El Karoui (2009b) that

$$\hat{\mu}'_W \mathcal{S}^{-1} \hat{\mu}_W \rightarrow \rho$$

in probability. So all that is left to do is understand $\zeta_{n,p}$.

We now prove Lemma B-1

Proof of Lemma B-1. The first part of the lemma is almost immediate after we realize that $\hat{\mu}' = \mathbf{e}' X/n = \mathbf{e}' Y \Sigma^{1/2}/n$. All the matrices Σ cancel and we are left with an expression that does not involve Σ .

For the second part, we use a differentiation trick we will use repeatedly in this paper. Let us call $M_t = M + t \text{Id}$, where $t \geq 0$ and M is assumed to be positive definite (this is all we need for this paper, but the results below hold in more generality). In this case, M_t is also invertible for any $t \geq 0$ (and actually for t in a neighborhood of 0). Simple calculus shows that, for M positive definite, and any $t \geq 0$, M_t^{-1} is differentiable on $[0, \infty)$ and

$$u' (M_t')^{-2} u = -\frac{\partial}{\partial t} u' (M_t')^{-1} u,$$

M being invertible guaranteeing differentiability at 0.

Now, using the classic expansion of the inverse of a rank-1 perturbation of M_t (see Horn and Johnson (1990), p. 19), we have, if $u' M_t^{-1} u \neq 1$,

$$(M_t - uu')^{-1} = M_t^{-1} + \frac{M_t^{-1} u u' M_t^{-1}}{1 - u' M_t^{-1} u}, \text{ and}$$

$$u' (M_t - uu')^{-1} u = \frac{1}{1 - u' M_t^{-1} u} - 1.$$

Differentiating the last equality and multiplying by (-1) , we get, for any $t \geq 0$,

$$u' (M_t - uu')^{-2} u = \frac{u' M_t^{-2} u}{(1 - u' M_t^{-1} u)^2}.$$

Applying the previous equality at $t = 0$ gives the result.

We note that another method of proof would be to use a rank one update and then take squares. This is a bit more tedious than the simple trick we presented here, but also shows that the result apply for any invertible M . \square

B-1.2 Structure of the proof

The proof of convergence of $\zeta_{n,p}$ is based on regularization ideas. In particular, we will focus on

$$\zeta_{n,p}(t) = \widehat{\mu}'_W (\mathcal{S} + t\text{Id}_p)^{-2} \widehat{\mu}_W, \quad (\text{B-5})$$

where t is a positive number. The overall strategy is the following:

1. Show that for any $\epsilon > 0$, we can find t_ϵ such that $|\zeta_{n,p} - \zeta_{n,p}(t_\epsilon)| < \epsilon$ with high-probability.
2. Show, for any given t_ϵ that $\zeta_{n,p}(t_\epsilon)$ converges in probability to $\rho\mathfrak{s}(t_\epsilon)$, where this quantity is deterministic.
3. Show that $\mathfrak{s}(t_\epsilon)$ can be made arbitrarily close to \mathfrak{s} (by picking ϵ small enough), so that in the end one concludes that $\zeta_{n,p} - \rho\mathfrak{s}$ tends to 0 in probability.

Technically, $\zeta_{n,p}(t)$ is a much nicer object to work with than $\zeta_{n,p}$ since it is bounded and amenable to variance computations (at least conditionally on Λ). This is especially useful because to show convergence in probability of $\zeta_{n,p}(t)$ we rely - among other things - on conditional variance computations, based on the Efron-Stein inequality (see Lugosi (2006), Theorem 9).

B-1.3 Approximating $\zeta_{n,p}$ by $\zeta_{n,p}(t_n)$

To show that we can approximate $\zeta_{n,p}$ by $\zeta_{n,p}(t)$, we rely on the following lemma:

Lemma B-2. *Suppose Assumption-BB is satisfied. Then we have, as $n \rightarrow \infty$,*

$$\forall \epsilon > 0, \exists t_\epsilon : P(|\zeta_{n,p} - \zeta_{n,p}(t_\epsilon)| > \epsilon) \rightarrow 0.$$

In plain english, the lemma means that for any $\epsilon > 0$ we can find $\zeta_{n,p}(t_\epsilon)$, which approximates $\zeta_{n,p}$ to within ϵ with high-probability.

In this and other proofs, it will be convenient to work with the matrix $P_2(t)$ defined as

$$P_2(t) = \frac{W}{\sqrt{n}} \left(\frac{W'W}{n} + t\text{Id}_p \right)^{-2} \frac{W'}{\sqrt{n}}. \quad (\text{B-6})$$

Proof. Recall that $\zeta_{n,p} = \zeta_{n,p}(0)$ and that, if $W = \Lambda Y$ and $\nu' = \mathbf{e}'W/n$,

$$\begin{aligned} \zeta_{n,p}(t) &= \nu' \left(\frac{W'W}{n} + t\text{Id}_p \right)^{-2} \nu \\ &= \frac{\mathbf{e}'W}{n} \left(\frac{W'W}{n} + t\text{Id}_p \right)^{-2} \frac{W'\mathbf{e}}{n} \\ &= \frac{\mathbf{e}'}{\sqrt{n}} \left(\frac{W}{\sqrt{n}} \left(\frac{W'W}{n} + t\text{Id}_p \right)^{-2} \frac{W'}{\sqrt{n}} \right) \frac{\mathbf{e}}{\sqrt{n}} \\ &\triangleq \frac{\mathbf{e}'}{\sqrt{n}} P_2(t) \frac{\mathbf{e}}{\sqrt{n}}. \end{aligned}$$

Let us write the singular value decomposition of W/\sqrt{n} as

$$\frac{W}{\sqrt{n}} = UDV'.$$

Then, using the fact that V is orthogonal, we have

$$\left(\frac{W'W}{n} + t\text{Id}_p \right) = V(D^2 + t\text{Id}_p)V'.$$

Therefore,

$$\left(\frac{W'W}{n} + t\text{Id}_p\right)^{-2} = V(D^2 + t\text{Id}_p)^{-2}V', \text{ and}$$

$$P_2(t) = UD(D^2 + t\text{Id}_p)^{-2}DU' = \sum_{i=1}^p \frac{d_i^2}{(d_i^2 + t)^2} u_i u_i'.$$

Recall that our aim is to compare $P_2(0) - P_2(t)$, or more precisely, quadratic forms involving this difference of matrices. Towards this end, we make the following simple remark: suppose that the d_i 's are decreasingly ordered and $t > 0$:

$$0 \leq \frac{1}{d_i^2} - \frac{d_i^2}{(d_i^2 + t)^2} = \frac{2td_i^2 + t^2}{d_i^2(d_i^2 + t)^2} = \frac{2t}{(d_i^2 + t)^2} + \frac{t^2}{d_i^2(d_i^2 + t)^2} \leq \frac{2t}{(d_p^2 + t)^2} + \frac{t^2}{d_p^2(d_p^2 + t)^2}.$$

We can hence conclude that

$$0 \leq \zeta_{n,p} - \zeta_{n,p}(t) \leq \frac{2t}{(d_p^2 + t)^2} + \frac{t^2}{d_p^2(d_p^2 + t)^2} \quad (\text{B-7})$$

since

$$\begin{aligned} 0 \leq \zeta_{n,p} - \zeta_{n,p}(t) &= \frac{\mathbf{e}'}{\sqrt{n}} (P_2(0) - P_2(t)) \frac{\mathbf{e}}{\sqrt{n}} \\ &\leq \left(\frac{2t}{(d_p^2 + t)^2} + \frac{t^2}{d_p^2(d_p^2 + t)^2} \right) \sum_{i=1}^p \left(u_i' \frac{\mathbf{e}}{\sqrt{n}} \right)^2 \leq \frac{2t}{(d_p^2 + t)^2} + \frac{t^2}{d_p^2(d_p^2 + t)^2}, \end{aligned}$$

because $\|\mathbf{e}/\sqrt{n}\|_2 = 1$ and U is an orthogonal matrix.

Let us now explain why d_p is bounded away from 0 under our assumptions. Let us call $\mathcal{L}_{\epsilon_0, \delta}$ the set of matrices Λ such that $p/N < 1 - \epsilon_0$ and $C_0(N - 1)/(n - 1) > \delta$. Under our assumptions, for a δ_0 bounded away from 0 (e.g $\delta_0 = 1/2 \liminf C_0(N - 1)/(n - 1)$), $P(\Lambda \in \mathcal{L}_{\epsilon_0, \delta_0}) \rightarrow 1$. 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 in El Karoui (2009b), if P_Λ denotes probability conditional on Λ ,

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

Hence, when $\Lambda \in \mathcal{L}_{\epsilon_0, \delta_0}$, d_p , the smallest singular value of W/\sqrt{n} , is bounded away from 0 with high-probability.

We conclude that

$$\forall \epsilon > 0, \exists t_\epsilon \text{ such that } P(|\zeta_{n,p} - \zeta_{n,p}(t_\epsilon)| > \epsilon) \rightarrow 0.$$

□

B-1.4 About $\mathbf{E}(\zeta_{n,p}(t))$ and $\mathbf{E}(\zeta_{n,p}(t)|\Lambda)$

Recall that

$$\zeta_{n,p}(t) = \frac{1}{n} \mathbf{e}' P_2(t) \mathbf{e},$$

and is therefore the scaled sum of the entries of $P_2(t)$. Let us call $p_{2,t}(i, j)$ the (i, j) -th entry of $P_2(t)$. We have the following lemma:

Lemma B-3. *When Y_i are i.i.d and have a symmetric distribution (i.e $Y_i \stackrel{\mathcal{L}}{=} -Y_i$), we have, if $i \neq j$,*

$$\mathbf{E}(p_{2,t}(i, j)|\Lambda) = \mathbf{E}(p_{2,t}(i, j)) = 0.$$

Further, for any given $t > 0$,

$$\mathbf{E}(\zeta_{n,p}(t)|\Lambda) - \frac{1}{n} \text{trace}(P_2(t)) \rightarrow 0, \text{ in probability,}$$

and $\frac{1}{n} \text{trace}(P_2(t))$ has a deterministic limit, which depends only on G , the limiting spectral distribution of Λ .

Proof. In all the proof, we assume that $t > 0$ is given. We use an invariance idea similar to the one used in a corresponding situation in El Karoui (2009b). Let us first note that the expectations we are referring to are well-defined. As a matter of fact, because $W'W/n + t\text{Id}_p \succeq t\text{Id}_p$, we have

$$\|P_2(t)\|_2 \leq \frac{1}{t} \|W(W'W)^{-1}W'\|_2 = \frac{1}{t},$$

since the matrix appearing in the right-hand side is an orthogonal projection matrix. Therefore, all the entries of $P_2(t)$ are bounded and less than $1/t$.

We now work conditionally on Λ and focus on the case $i = 1, j \neq 1$. Let us call

$$p_{2,t}(1, j) = \Theta_\Lambda(Y_1, \dots, Y_n) = \lambda_1 \lambda_j Y_1' \left(\frac{1}{n} \sum_{i=1}^n \lambda_i^2 Y_i Y_i' + t\text{Id}_p \right)^{-2} Y_j.$$

We have clearly, $\Theta_\Lambda(Y_1, Y_2, \dots, Y_n) = -\Theta_\Lambda(-Y_1, Y_2, \dots, Y_n)$. In other respects, because $Y_i \stackrel{\mathcal{L}}{=} -Y_i$, and the Y_i 's are independent, we have $\Theta_\Lambda(Y_1, Y_2, \dots, Y_n) \stackrel{\mathcal{L}}{=} \Theta_\Lambda(-Y_1, Y_2, \dots, Y_n)$. Because $\Theta_\Lambda(Y_1, Y_2, \dots, Y_n)$ is bounded, we can take expectations in the previous equality and we have, if $i \neq j$,

$$\mathbf{E}(p_{2,t}(1, j) | \Lambda) = 0. \tag{B-8}$$

We conclude that the same result holds unconditionally and $\mathbf{E}(p_{2,t}(i, j)) = 0$, since under our assumptions P_2 is defined with probability 1.

Hence, for any $t > 0$,

$$\mathbf{E}(\zeta_{n,p}(t)) = \frac{1}{n} \sum_{i=1}^n \mathbf{E}(p_{2,t}(i, i)) = \frac{1}{n} \mathbf{E}(\text{trace}(P_2(t))).$$

Let us now argue that, for any fixed $t > 0$,

$$\zeta_{n,p}(t) - \frac{1}{n} \text{trace}(P_2(t)) \rightarrow 0 \text{ in probability.}$$

If $\mathcal{S} = W'W/n$ and if l_i are the eigenvalues of \mathcal{S} ,

$$\frac{1}{n} \text{trace}(P_2(t)) = \frac{p}{n} \frac{1}{p} \sum_{i=1}^p \frac{l_i}{(l_i + t)^2} = \rho_n \left(\frac{1}{p} \sum_{i=1}^p \frac{1}{l_i + t} - t \frac{1}{(l_i + t)^2} \right).$$

Under our assumptions on the convergence of the spectral distribution of Λ , we have convergence of the spectral distribution of \mathcal{S} in probability (see Marčenko and Pastur (1967), Wachter (1978), Silverstein (1995), El Karoui (2009a) and El Karoui (2009b)) to a deterministic probability measure \mathcal{K} , and therefore

$$\frac{1}{n} \text{trace}(P_2(t)) \rightarrow \rho \mathfrak{s}(t) \text{ in probability.}$$

Though we will not need it later for any explicit computations, let us mention that if \mathcal{K} is the limiting spectral distribution of \mathcal{S} , we have, since $a(x) = 1/(x + t)$ is continuous and bounded on $[0, \infty)$,

$$\mathfrak{s}(t) = \int \frac{d\mathcal{K}(l)}{l + t}. \tag{B-9}$$

Because $\frac{1}{n} \text{trace}(P_2(t)) \leq 1/t$, we also have $\text{var}(\frac{1}{n} \text{trace}(P_2(t))) \rightarrow 0$ and $\mathbf{E}(\frac{1}{n} \text{trace}(P_2(t))) \rightarrow \rho \mathfrak{s}(t)$. Note also that since

$$\mathbf{E}(\zeta_{n,p}(t) | \Lambda) = \frac{1}{n} \mathbf{E}(\text{trace}(P_2(t)) | \Lambda),$$

we also have, under our assumptions, that

$$\text{var} \left(\frac{1}{n} \mathbf{E}(\text{trace}(P_2(t)) | \Lambda) \right) \rightarrow 0, .$$

since $\frac{1}{n}\mathbf{E}(\text{trace}(P_2(t))|\Lambda)$ is a bounded random variable, depending on Λ , which converges in probability to a limit that is independent of Λ .

So for any given t ,

$$\begin{aligned} \mathbf{E}(\zeta_{n,p}(t)|\Lambda) - \rho\mathfrak{s}(t) &\rightarrow 0, \text{ in probability, and} \\ \mathbf{E}(\zeta_{n,p}(t)|\Lambda) - \frac{1}{n}\text{trace}(P_2(t)) &\rightarrow 0 \text{ in probability.} \end{aligned}$$

The first statement is to be understood in Λ -probability. Note that $\mathfrak{s}(t)$ depends only on the limiting spectral distribution of Λ , which is the same for all the Λ 's we consider. In other words, $\mathfrak{s}(t)$ is deterministic, and the random variable $\mathbf{E}(\zeta_{n,p}(t)|\Lambda)$ is asymptotically non-random. The second statement is to be understood with respect to the probability induced by the joint distribution of the λ_i 's and the Y_i 's (so the joint $\Lambda \times Y$ -probability). □

B-1.5 Conditional variance and convergence of $\zeta_{n,p}(t)$

We now place ourselves in the setting where $\mu = 0$. To understand $\widehat{\mu}'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}\widehat{\mu}$, we simply need to understand, if $W = \Lambda Y$,

$$\zeta_{n,p} = \frac{\mathbf{e}'W}{n} \left(\frac{W'W}{n} \right)^{-2} \frac{W'\mathbf{e}}{n}.$$

Lemma B-4. *Suppose that $Y_i \stackrel{iid}{\sim} \mathcal{N}(0, \text{Id}_p)$ and $\zeta_{n,p}(t)$ is defined as above in Equation (B-5). Then we can find a constant C and two (explicit) functions g_1 and g_2 such that for any $t > 0$*

$$\text{var}(\zeta_{n,p}(t)|\Lambda) \leq C \frac{1}{n^2} \sum_{i=1}^n (\lambda_i^4 g_1(t) + g_2(t)).$$

Furthermore, for any $t > 0$, when Assumption-BLa and Assumption-BB are satisfied,

$$\zeta_{n,p}(t) - \frac{1}{n}\text{trace}(P_2(t)) \rightarrow 0 \text{ in probability.}$$

The key to the proof of this lemma is the Efron-Stein inequality (Efron and Stein, 1981), as stated for instance in Theorem 9 of Lugosi (2006): it says that if X is a random variable such that $X = f(\xi_1, \dots, \xi_n)$, where the ξ_i 's are independent random variables, then, if $X_i = f_i(\xi_1, \dots, \xi_{i-1}, \xi_{i+1}, \dots, \xi_n)$,

$$\text{var}(X) \leq \sum_{i=1}^n \text{var}(X - X_i).$$

We will therefore try to approximate $\zeta_{n,p}(t)$ by random variables $\zeta_i(t)$ involving all but one of the Y_i 's to get our conditional variance bound - the key to the lemma.

Proof. We call

$$\mathcal{S} = W'W/n = \frac{1}{n} \sum_{i=1}^n \lambda_i^2 Y_i Y_i'.$$

We call $\mathcal{S}_i = \mathcal{S} - \lambda_i^2 Y_i Y_i'$. We also call $\mathcal{S}(t) = \mathcal{S} + t\text{Id}_p$ and $\mathcal{S}_i(t) = \mathcal{S}(t) - \lambda_i^2 Y_i Y_i'/n$.

Let us call $\widehat{m} = W'\mathbf{e}/n$ and $\widehat{m}_i = W_i'\mathbf{e}/n$, where $W_i = W - \lambda_i e_i Y_i'$ (e_i is the i -th canonical basis vector in \mathbb{R}^n). Note that W_i is simply W with the i -th row replaced by 0.

We call

$$\begin{aligned} Z_{n,p}(t) &\triangleq \widehat{m}'(\mathcal{S}(t))^{-1}\widehat{m} \\ Z_i(t) &\triangleq \widehat{m}_i'(\mathcal{S}_i(t))^{-1}\widehat{m}_i. \end{aligned}$$

Note that

$$\begin{aligned}\frac{\partial Z_{n,p}(t)}{\partial t} &= -\widehat{m}'(\mathcal{S}(t))^{-2} \widehat{m} = -\zeta_{n,p}(t) \\ \frac{\partial Z_i(t)}{\partial t} &= -\widehat{m}'_i(\mathcal{S}_i(t))^{-2} \widehat{m}_i = -\zeta_i(t).\end{aligned}$$

Now, recall the results of El Karoui (2009b), Theorem 4.2 and its proof:

$$\begin{aligned}Z_{n,p}(t) &= Z_i(t) + \frac{1}{n} \left(1 - \frac{(1 - \lambda_i w_i(t))^2}{1 + \lambda_i^2 q_i(t)} \right), \text{ where} \\ w_i(t) &= \widehat{m}'_i(\mathcal{S}_i(t))^{-1} Y_i, \\ q_i(t) &= \frac{Y'_i(\mathcal{S}_i(t))^{-1} Y_i}{n}.\end{aligned}$$

Note that $q_i(t) \geq 0$. Taking derivatives with respect to t in the previous expression, we get:

$$\frac{\partial Z_{n,p}(t)}{\partial t} = \frac{\partial Z_i(t)}{\partial t} - \frac{1}{n} \left[\frac{2(\lambda_i w_i(t) - 1)\lambda_i w'_i(t)}{1 + \lambda_i^2 q_i(t)} - \frac{(1 - \lambda_i w_i(t))^2}{(1 + \lambda_i^2 q_i(t))^2} \lambda_i^2 q'_i(t) \right].$$

Hence,

$$|\zeta_{n,p}(t) - \zeta_i(t)| = \frac{1}{n} \left| \frac{2(\lambda_i w_i(t) - 1)\lambda_i w'_i(t)}{1 + \lambda_i^2 q_i(t)} - \frac{(1 - \lambda_i w_i(t))^2}{(1 + \lambda_i^2 q_i(t))^2} \lambda_i^2 q'_i(t) \right|.$$

We notice that, trivially, $\mathcal{S}(t) \succeq t\text{Id}_p$ and $\mathcal{S}_i(t) \succeq t\text{Id}_p$. Also,

$$\begin{aligned}w'_i(t) &= -\widehat{m}'_i(\mathcal{S}_i(t))^{-2} Y_i \\ q'_i(t) &= -\frac{Y'_i(\mathcal{S}_i(t))^{-2} Y_i}{n}.\end{aligned}$$

Therefore,

$$|q'_i(t)| \leq \frac{\|(\mathcal{S}_i(t))^{-1}\|_2 \|(\mathcal{S}_i(t))^{-1/2} Y_i\|_2^2}{n} = q_i(t) \frac{1}{t}.$$

Consequently,

$$\left| \frac{|1 - \lambda_i w_i(t)|}{(1 + \lambda_i^2 q_i(t))^2} \lambda_i^2 q'_i(t) \right| \leq \frac{1}{t} \frac{|1 - \lambda_i w_i(t)|}{1 + \lambda_i^2 q_i(t)}.$$

Let us write

$$\Delta_i(t) = n(\zeta_{n,p}(t) - \zeta_i(t)) = \frac{\lambda_i w_i(t) - 1}{1 + \lambda_i^2 q_i(t)} \left(2\lambda_i w'_i(t) - \frac{\lambda_i w_i(t) - 1}{1 + \lambda_i^2 q_i(t)} \lambda_i^2 q'_i(t) \right). \quad (\text{B-10})$$

The remarks we made above show that

$$\begin{aligned}|\Delta_i(t)| &\leq \left[\frac{(1 - \lambda_i w_i(t))^2}{t} + 2\lambda_i^2 |w_i(t) w'_i(t)| + 2|\lambda_i| |w'_i(t)| \right] \frac{1}{1 + \lambda_i^2 q_i(t)} \\ &\leq \frac{2}{t} + \lambda_i^2 \left(2\frac{w_i^2(t)}{t} + |w_i(t) w'_i(t)| \right) + 2|\lambda_i| |w'_i(t)|\end{aligned}$$

Using the fact that $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$, as well as the equally well-known $(a + b)^2 \leq 2(a^2 + b^2)$ and $|ab| \leq (a^2 + b^2)/2$, we have

$$\Delta_i(t)^2 \leq 3 \left[\frac{4}{t^2} + \lambda_i^4 \left(\frac{8w_i^4(t)}{t^2} + w_i^4(t) + (w'_i(t))^4 \right) + 4\lambda_i^2 (w'_i(t))^2 \right]$$

Note that, conditional on $Y_{(i)} = (Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n)$ and Λ , $w'_i(t) \sim \mathcal{N}(0, \widehat{m}'_i(\mathcal{S}_i(t))^{-4} \widehat{m}_i)$. Recall also from El Karoui (2009b) that under the same conditioning, $w_i(t) \sim \mathcal{N}(0, \widehat{m}'_i(\mathcal{S}_i(t))^{-2} \widehat{m}_i)$. So in particular, since

$$\widehat{m}'_i(\mathcal{S}_i(t))^{-4} \widehat{m}_i \leq \frac{1}{t^3},$$

because, as explained in El Karoui (2009b), $\widehat{m}'_i(\mathcal{S}_i(t))^{-1}\widehat{m}_i \leq 1$, we have, using the same arguments as in El Karoui (2009b),

$$\mathbf{E}((w'_i(t))^2|\Lambda) \leq \frac{1}{t^3} \text{ and } \mathbf{E}((w'_i(t))^4|\Lambda) \leq \frac{3}{t^6} .$$

Similarly,

$$\mathbf{E}(w_i^4(t)|\Lambda) \leq \frac{3}{t^2} .$$

So we conclude that, for C a constant (independent of p and n),

$$\mathbf{E}((\Delta_i(t))^2|\Lambda) \leq C(\lambda_i^4 f_1(t) + \lambda_i^2 f_2(t) + f_3(t)) .$$

For completeness, let us say that a possible choice of such functions is $f_1(t) = 8t^{-4} + t^{-2} + t^{-6}$, $f_2(t) = t^{-3}$ and $f_3(t) = t^{-2}$.

After using the very coarse bound $\lambda_i^2 \leq 1 + \lambda_i^4$, this inequality can be rewritten as (for appropriate functions g_1 and g_2)

$$\mathbf{E}((\Delta_i(t))^2|\Lambda) \leq C(\lambda_i^4 g_1(t) + g_2(t)) .$$

Applying the Efron-Stein's inequality (see Lugosi (2006), Theorem 9), we see that,

$$\text{var}(\zeta_{n,p}(t)|\Lambda) \leq \sum_{i=1}^n \mathbf{E}((\zeta_{n,p}(t) - \zeta_i(t))^2|\Lambda) \leq C \frac{1}{n^2} \sum_{i=1}^n (\lambda_i^4 g_1(t) + g_2(t)) .$$

Let us now consider $\mathcal{L}_\eta = \{\Lambda : \frac{1}{n^2} \sum_{i=1}^n \lambda_i^4 \leq \eta\}$. If $\Lambda \in \mathcal{L}_\eta$, we have, for all $x > 0$, and P_Λ denotes probability conditional on Λ ,

$$P_\Lambda(|\zeta_{n,p}(t) - \mathbf{E}(\zeta_{n,p}(t)|\Lambda)| > x) \leq \frac{1}{x^2} C \left(\eta g_1(t) + \frac{g_2(t)}{n} \right) .$$

In particular, when n is large enough and $\epsilon > 0$ is given, we see that we can pick an $\eta(t, x, \epsilon)$ such that for any given x and t ,

$$P_\Lambda(|\zeta_{n,p}(t) - \mathbf{E}(\zeta_{n,p}(t)|\Lambda)| > x) \leq 4C\epsilon .$$

Now, given x , t and ϵ , $P(\Lambda \in \mathcal{L}_{\eta(t,x,\epsilon)}) \rightarrow 1$ under our assumptions. Also, we have shown above that $\mathbf{E}(\zeta_{n,p}(t)|\Lambda) \rightarrow \rho \mathfrak{s}(t)$ in Λ -probability, where $\mathfrak{s}(t)$ is deterministic and the same for a set of Λ 's of probability 1. In other words, if we call $\mathcal{L}_{\delta,t} = \{\Lambda : |\mathbf{E}(\zeta_{n,p}(t)|\Lambda) - \rho \mathfrak{s}(t)| \leq \delta\}$, for any $\delta > 0$ and $t > 0$, we have $P(\mathcal{L}_{\delta,t}) \rightarrow 1$. Of course, if $\mathcal{L}_{\eta(t,x,\epsilon),\delta} = \mathcal{L}_{\eta(t,x,\epsilon)} \cap \mathcal{L}_{\delta,t}$, we also have $P(\mathcal{L}_{\eta(t,x,\epsilon),\delta}) \rightarrow 1$. So we conclude, by elementary conditioning arguments (similar to the ones found at the end of the proof of Theorem 4.2 in El Karoui (2009b)) that under our assumptions

$$\text{for any given } t > 0, \quad \zeta_{n,p}(t) - \rho \mathfrak{s}(t) \rightarrow 0 \text{ in probability,}$$

where the last statement is of course to be understood unconditionally on Λ .

Note that because $\frac{1}{n} \text{trace}(P_2(t)) \rightarrow \rho \mathfrak{s}(t)$ in (unconditional) probability, this also implies that, for any given $t > 0$,

$$\zeta_{n,p}(t) - \frac{1}{n} \text{trace}(P_2(t)) \rightarrow 0 \text{ in probability.}$$

□

B-1.6 Convergence of $\zeta_{n,p}$

We now put together all the previous arguments to show the first result announced in Theorem B.1, namely that, under our assumptions,

$$\zeta_{n,p} \rightarrow \rho \mathfrak{s} \text{ in probability .}$$

As a matter of fact, we have, for any $t > 0$,

$$\left| \zeta_{n,p} - \frac{1}{n} \text{trace}(P_2(0)) \right| \leq |\zeta_{n,p} - \zeta_{n,p}(t)| + \left| \zeta_{n,p}(t) - \frac{1}{n} \text{trace}(P_2(t)) \right| + \left| \frac{1}{n} \text{trace}(P_2(t)) - \frac{1}{n} \text{trace}(P_2(0)) \right| .$$

We have already established that, if d_p is the smallest singular value of $\mathcal{S} = W'W/n$,

$$0 \leq \zeta_{n,p} - \zeta_{n,p}(t) \leq \frac{2t}{(d_p^2 + t)^2} + \frac{t^2}{d_p^2(d_p^2 + t)^2} \text{ and}$$

$$\zeta_{n,p}(t) - \frac{1}{n}\text{trace}(P_2(t)) \rightarrow 0 \text{ in probability .}$$

Now, we also have by the same argument as the one we used to bound $\zeta_{n,p} - \zeta_{n,p}(t)$

$$\left| \frac{1}{n}\text{trace}(P_2(t)) - \frac{1}{n}\text{trace}(P_2(0)) \right| \leq \frac{2t}{(d_p^2 + t)^2} + \frac{t^2}{d_p^2(d_p^2 + t)^2}$$

Since under our assumptions d_p is bounded away from 0 with high-probability, for any $\epsilon > 0$, we can find t_ϵ such that, in probability,

$$\begin{aligned} |\zeta_{n,p} - \zeta_{n,p}(t_\epsilon)| &< \epsilon , \\ \left| \zeta_{n,p}(t_\epsilon) - \frac{1}{n}\text{trace}(P_2(t_\epsilon)) \right| &< \epsilon , \text{ and} \\ \left| \frac{1}{n}\text{trace}(P_2(t_\epsilon)) - \frac{1}{n}\text{trace}(P_2(0)) \right| &< \epsilon . \end{aligned}$$

Hence,

$$\left| \zeta_{n,p} - \frac{1}{n}\text{trace}(P_2(0)) \right| \rightarrow 0 \text{ in probability .}$$

Now

$$\frac{1}{n}\text{trace}(P_2(0)) = \rho_n \frac{1}{p} \sum_{i=1}^p d_i^{-1} \rightarrow \rho \mathbf{s} \text{ in probability ,}$$

by the analysis done in El Karoui (2009b) and the result is shown.

B-1.7 On $\hat{\boldsymbol{\mu}}' \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}$

Let us now focus on the second result announced in Theorem B.1. Recall that the statistic we were interested in was

$$Z_{n,p} = \hat{\boldsymbol{\mu}}' \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} .$$

Calling $T_{n,p} = \hat{\boldsymbol{\mu}}' \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}$, we have shown in Lemma B-1 that

$$Z_{n,p} = \frac{\zeta_{n,p}}{(1 - T_{n,p})^2} .$$

In El Karoui (2009b), we showed that $T_{n,p} \rightarrow \rho$ under our assumptions. And we just showed that

$$\zeta_{n,p} \rightarrow \rho \mathbf{s} .$$

Recalling the notation $\kappa = \rho/(1 - \rho)$, we finally have

$$Z_{n,p} \rightarrow \frac{\kappa}{1 - \rho} \mathbf{s} ,$$

as announced in Theorem B.1.

B-2 On $\hat{\mu}'\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}v$

We use here the normalization $\hat{\Sigma} = \frac{1}{n}X'HX$. Note that $\hat{\Sigma}$ is shift-invariant (it does not depend on μ), so we call $\hat{m} = W'e/n$, where $W = \Lambda Y$. \hat{m} is the sample mean in the case where $\mu = 0$. Let us call $\mathcal{S} = W'W/n$. We have

$$\hat{\Sigma} = \Sigma^{1/2} (\mathcal{S} - \hat{m}\hat{m}') \Sigma^{1/2}.$$

Therefore, we have

$$\begin{aligned} \hat{\mu}'\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\mu} &= \mu'\Sigma^{-1/2} (\mathcal{S} - \hat{m}\hat{m}')^{-2} \Sigma^{-1/2}\mu + 2\mu'\Sigma^{-1/2} (\mathcal{S} - \hat{m}\hat{m}')^{-2} \hat{m} \\ &\quad + \hat{m}' (\mathcal{S} - \hat{m}\hat{m}')^{-2} \hat{m}. \end{aligned}$$

On the other hand, if v is a deterministic vector,

$$\hat{\mu}'\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}v = \mu'\Sigma^{-1/2} (\mathcal{S} - \hat{m}\hat{m}')^{-2} \Sigma^{-1/2}v + \hat{m}' (\mathcal{S} - \hat{m}\hat{m}')^{-2} \Sigma^{-1/2}v.$$

The work earlier in this paper gives results concerning quadratic forms of the type $v'\Sigma^{-1/2} (\mathcal{S} - \hat{m}\hat{m}')^{-2} \Sigma^{-1/2}v$, for a fixed sequence of vectors v , and $\hat{m}' (\mathcal{S} - \hat{m}\hat{m}')^{-2} \hat{m}$.

To complete the study, we now need to understand quantities of the type

$$\hat{m}' (\mathcal{S} - \hat{m}\hat{m}')^{-2} v.$$

We now turn to studying these objects and begin with the following lemma:

Lemma B-5. *Let u be a vector and M be a positive definite matrix such that $M - uu'$ is invertible. Then*

$$u'(M - uu')^{-2}v = \frac{u'M^{-2}v}{1 - u'M^{-1}u} + \frac{u'M^{-2}u}{(1 - u'M^{-1}u)^2}u'M^{-1}v. \quad (\text{B-11})$$

In our applications of this lemma we will have $M = \mathcal{S}$, $u = \hat{m}$ and v deterministic. Because we have studied in El Karoui (2009b) quantities of the type $v'\mathcal{S}^{-1}\hat{m}$ and $\hat{m}'\mathcal{S}^{-1}\hat{m}$, and because in light of the results above we now understand $\hat{m}'\mathcal{S}^{-2}\hat{m}$, we will just have to focus on quantities of the type $v'\mathcal{S}^{-2}\hat{m}$ to get a general understanding of statistics of the form $\hat{\mu}'\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\mu}$.

Proof. We use the same trick as above. Because M is invertible, $M_t = M + t\text{Id}$ is such that M_t^{-1} is well-defined and differentiable in a neighborhood of 0. Recall that the rank one update formula gives

$$u'(M_t - uu')^{-1} = \frac{u'M_t^{-1}}{1 - u'M_t^{-1}u}.$$

Let us call $g(t) = -u'(M_t - uu')^{-1}v$. We have

$$g'(0) = u'(M - uu')^{-2}v.$$

However, $g(t) = (u'M_t^{-1}v)/(1 - u'M_t^{-1}u)$. Therefore,

$$g'(0) = \frac{u'M^{-2}v}{1 - u'M^{-1}u} + \frac{u'M^{-1}v}{(1 - u'M^{-1}u)^2}u'M^{-2}u.$$

□

B-2.1 On $\hat{m}'\mathcal{S}^{-2}v$

Our aim is to show the following theorem:

Theorem B.2. *Suppose that v is a deterministic vector with $\|v\|_2 = 1$. Suppose that the assumptions stated in Theorem B.2 hold and also that*

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

Consider, if $W = \Lambda Y$, $\mathcal{S} = W'W/n$,

$$\psi = \frac{1}{n} \mathbf{e}' \Lambda Y \mathcal{S}^{-2} v = \widehat{m}' \mathcal{S}^{-2} v .$$

Then

$$\psi \rightarrow 0 \text{ in probability.} \quad (\text{B-12})$$

Furthermore, we have

$$\widehat{m}' (\mathcal{S} - \widehat{m} \widehat{m}')^{-2} v \rightarrow 0 \text{ in probability.} \quad (\text{B-13})$$

Proof. As before, our proof will rely on an approximation argument and the Efron-Stein inequality. We call, for $t > 0$,

$$\psi(t) = \frac{1}{n} \mathbf{e}' \Lambda Y (\mathcal{S} + t \text{Id}_p)^{-2} v .$$

As before we call $S(t) = \mathcal{S} + t \text{Id}_p$ and we will be using the same notations as in the proof of Theorem B.1. Note that

$$\psi(t) = \frac{\mathbf{e}}{\sqrt{n}} \frac{W}{\sqrt{n}} \left(\frac{W'W}{n} + t \text{Id}_p \right)^{-2} v .$$

We first remark that if W is changed to $-W$, then $\psi(t)$ is changed to $-\psi(t)$. We also note that since $\|S(t)\|_2 \leq 1/t$, $\psi(t)$ has an expectation, conditional on Λ . Since $Y \stackrel{\mathcal{L}}{=} -Y$, we immediately have

$$\mathbf{E}(\psi(t) | \Lambda) = 0 ,$$

since, conditional on Λ , $\psi(t) \stackrel{\mathcal{L}}{=} -\psi(t)$. Let us call $\psi_i(t)$ the random variable obtained by replacing λ_i by zero in the definition of $\psi(t)$. Note that $\psi_i(t)$ does not involve Y_i .

Now the proof of Theorem 4.3 in El Karoui (2009b) shows that if

$$\Psi(t) = \frac{\mathbf{e}}{\sqrt{n}} \frac{W}{\sqrt{n}} \left(\frac{W'W}{n} + t \text{Id}_p \right)^{-1} v ,$$

and $\Psi_i(t)$ is the random variable obtained from Ψ by replacing λ_i by 0,

$$\begin{aligned} \Psi(t) - \Psi_i(t) &= \frac{1}{n} \left(\frac{\lambda_i \theta_i(t) (1 - \lambda_i w_i(t))}{1 + \lambda_i^2 q_i(t)} \right) , \text{ where} \\ \theta_i(t) &= Y_i' (\mathcal{S}_i(t))^{-1} v . \end{aligned}$$

Naturally, we have $\psi(t) = -\frac{\partial \Psi(t)}{\partial t}$ and similarly for $\Psi_i(t)$ and $\psi_i(t)$. Therefore, we have

$$n(\psi_i(t) - \psi(t)) = \lambda_i \left[\frac{\theta_i(t)' (1 - \lambda_i w_i(t)) - \lambda_i w_i'(t) \theta_i(t)}{1 + \lambda_i^2 q_i(t)} - \frac{\theta_i(t) (1 - \lambda_i w_i(t))}{(1 + \lambda_i^2 q_i(t))^2} \lambda_i^2 q_i'(t) \right] .$$

So we conclude that

$$n|\psi_i(t) - \psi(t)| \leq |\lambda_i| \left(|\theta_i'(t)| |1 - \lambda_i w_i(t)| + |\lambda_i w_i'(t) \theta_i(t)| + \frac{|\theta_i(t)| |1 - \lambda_i w_i(t)|}{t} \right) .$$

A key aspect of this equation for our purposes is that λ_i appears in it with powers at most 2. We also note that $\theta_i'(t) | Y_{(i)}, \Lambda \sim \mathcal{N}(0, v' (\mathcal{S}_i(t))^{-4} v)$, and recall that $\theta_i(t) | Y_{(i)}, \Lambda \sim \mathcal{N}(0, v' (\mathcal{S}_i(t))^{-2} v)$. Finally, as seen many times before, $\mathcal{S}_i(t)^{-1} \leq t \text{Id}_p$.

Therefore, $\mathbf{E}((\theta'_i(t))^{-2k}) \leq C_k t^{-4k}$ and $\mathbf{E}((\theta_i(t))^{-2k}) \leq C_k t^{-2k}$. We conclude that

$$\mathbf{E}(n^2(\psi_i(t) - \psi(t))^2 | \Lambda) \leq C(\lambda_i^4 h_1(t) + h_2(t)) .$$

Hence, the Efron-Stein inequality guarantees that

$$\text{var}(\psi(t) | \Lambda) \leq C \frac{1}{n^2} \sum_{i=1}^n (\lambda_i^4 h_1(t) + h_2(t)) .$$

So if Λ is such that $\sum_{i=1}^n \lambda_i^4 / n^2 \rightarrow 0$, we have

$$\psi(t) \rightarrow 0 \text{ conditionally on } \Lambda .$$

Note that under Assumption-BLa, the set of matrices such that $\sum_{i=1}^n \lambda_i^4 / n^2 \rightarrow 0$ has measure 1.

We now recall that under Assumption-BB, we can find sets of matrices $\mathcal{L}_{\epsilon_0, \delta_0}$ whose measures go to 1 asymptotically, for which the smallest singular value of \mathcal{S} is bounded away from 0 with high-probability. Recall also from El Karoui (2009b) that conditionally on Λ , $\widehat{m} = W' \mathbf{e} / n$ is $\mathcal{N}(0, \frac{\sum_{i=1}^n \lambda_i^2}{n^2} \text{Id}_p)$ and therefore, $\|\widehat{m}\|_2^2 \sim \chi_p^2 / n (\sum_{i=1}^n \lambda_i^2 / n)$. Now

$$|\psi - \psi(t)| \leq \|\widehat{m}\| \|\|\mathcal{S}^{-2} - (\mathcal{S}(t))^{-2}\|_2 \|v\| .$$

So if $\Lambda \in \mathcal{L}_{\epsilon_0, \delta_0}$ and is such that $\sum_{i=1}^n \lambda_i^2 / n$ stays bounded, we see that for any $\epsilon > 0$, we can find t_ϵ such that $|\psi - \psi(t_\epsilon)| < \epsilon$ with high-probability. For such a Λ , satisfying also the conditions of Assumption-BLa, we have

$$\psi \rightarrow 0 \text{ in probability, conditionally on } \Lambda .$$

By using deconditioning arguments similar to the ones we presented above, we can conclude that under Assumption-BB, Assumption-BLa and Assumption-BLb, we have

$$\psi \rightarrow 0 \text{ in probability ,}$$

where this last statement is unconditional on Λ . Equation (B-12) is shown.

To show that the result announced in Equation (B-13) holds, we just notice that according to Lemma B-5,

$$\widehat{m}'(\mathcal{S} - \widehat{m}\widehat{m}')^{-2}v = \frac{\widehat{m}'\mathcal{S}^{-2}v}{1 - \widehat{m}'\mathcal{S}^{-1}\widehat{m}} + \frac{\widehat{m}'\mathcal{S}^{-2}\widehat{m}}{(1 - \widehat{m}'\mathcal{S}^{-1}\widehat{m})^2} \widehat{m}'\mathcal{S}^{-1}v .$$

Under our assumptions, results in El Karoui (2009b) show that $\widehat{m}'\mathcal{S}^{-1}\widehat{m} \rightarrow \rho$ in probability. We have also just established that $\widehat{m}'\mathcal{S}^{-2}\widehat{m}$ has a finite limit in probability. And finally, we know from El Karoui (2009b), Theorem 4.3 that $\widehat{m}'\mathcal{S}^{-1}v \rightarrow 0$ in probability. Because $\psi = \widehat{m}'\mathcal{S}^{-2}v$, we conclude that

$$\widehat{m}'(\mathcal{S} - \widehat{m}\widehat{m}')^{-2}v \rightarrow 0 \text{ in probability .}$$

□

B-2.2 Combining all the arguments together

The following theorem summarizes our findings and follows essentially immediately from the previous two, Theorems B.1 and B.2.

Theorem B.3. *Suppose that $X_i = \mu + \lambda_i \Sigma^{1/2} Y_i$, where Y_i are i.i.d $\mathcal{N}(0, \text{Id}_p)$ 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

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

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

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

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

We assume that V and Σ are such that Condition-P1 is satisfied.

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 $\mathcal{S} = W'W/n$. Finally, we use the notation $\hat{m} = W'\mathbf{e}/n$, $\tilde{\mu} = \Sigma^{-1/2}\mu$.

Then, we have, for ξ defined as in Equation (A-2), and \mathfrak{s} defined as in Equation (A-1),

$$\frac{\hat{\mu}'\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}v}{\sqrt{v'\Sigma^{-1}v}} = \frac{\mu'\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}v}{\sqrt{v'\Sigma^{-1}v}} + o_P(1) = \xi \frac{\mu'\Sigma^{-1}v}{\sqrt{v'\Sigma^{-1}v}} + o_P\left(1 \vee \frac{\mu'\Sigma^{-1}v}{\sqrt{v'\Sigma^{-1}v}}\right). \quad (\text{B-14})$$

Also,

$$\hat{\mu}'\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\mu} = \mu'\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\mu + \frac{\rho_n}{(1 - \rho_n)^2}\mathfrak{s} + 2\tilde{\mu}'(\mathcal{S} - \hat{m}\hat{m}')^{-2}\hat{m} + o_P(1) \quad (\text{B-15})$$

and we recall that $\tilde{\mu}'(\mathcal{S} - \hat{m}\hat{m}')^{-2}\hat{m}/\|\tilde{\mu}\| = o_P(1)$.

References

- 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. (2007). Covariance regularization by thresholding. Technical Report 744, Department of Statistics, UC Berkeley.
- 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.
- EFRON, B. and STEIN, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9**, 586–596.
- 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. (2009a). Concentration of measure and spectra of random matrices: with applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability To Appear*.

- EL KAROUI, N. (2009b). High-dimensionality effects in the Markowitz problem and other quadratic programs with linear equality constraints: risk underestimation. Technical Report 781, Department of Statistics, UC Berkeley.
- 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.
- 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.
- 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.
- LEDOUX, M. (2001). *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI.
- 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.
- 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.
- WACHTER, K. W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. *Ann. Probability* **6**, 1–18.

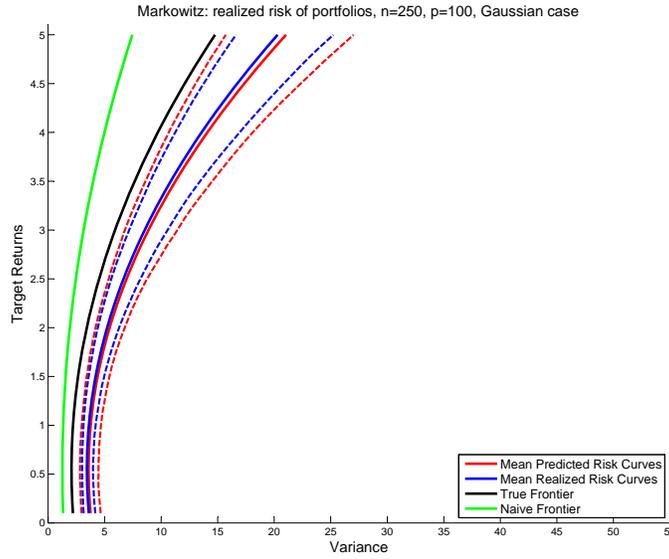
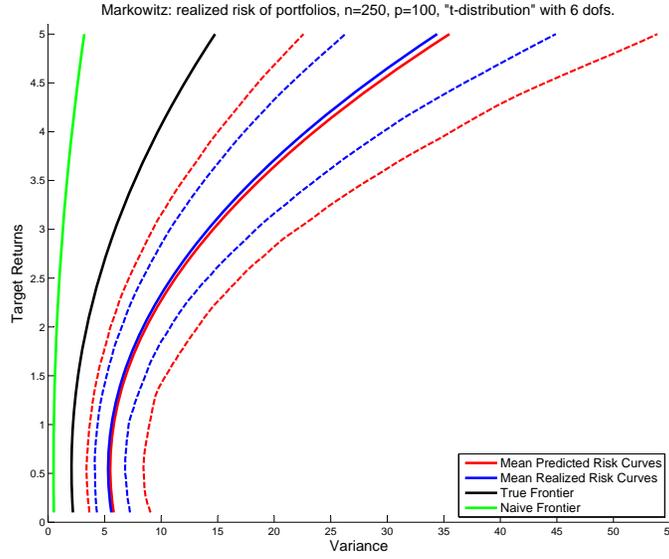


Figure 2: Performance of naive and corrected frontiers, for scaled “ t_6 ” (upper picture) and Gaussian returns. Here, $n = 250$ and $p = 100$. The number of simulations is 1000 in all pictures. The dashed lines represent (empirical) 95% confidence bands. (The confidence bands corresponds are computed for a fixed level of expected returns y .) The x -axis represents our estimate of the realized variance of the optimal portfolios. The y -axis represents the target returns for the portfolios. The plots show both the average realized risk of the naive Markowitz portfolios (blue curves) and the fact that our estimator (in red) is nearly unbiased (red solid curves near the blue curves). They also illustrate the robustness of our corrections. Another striking feature is the lack of robustness of Gaussian computations, since the average realized risk computed with “ t_6 ” returns are very different from the Gaussian ones. The fact that, as our theoretical work predicts, Gaussian computations leads to underestimation of the realized risk in the class of elliptical distributions considered in the paper is illustrated by the fact that the “ t_6 ” curves are to the right of the Gaussian curves. The population (or true) efficient frontier is in black. The green curve is the mean estimated risk, if estimated through the naive estimator $w'_{\text{emp}} \hat{\Sigma} w_{\text{emp}}$.

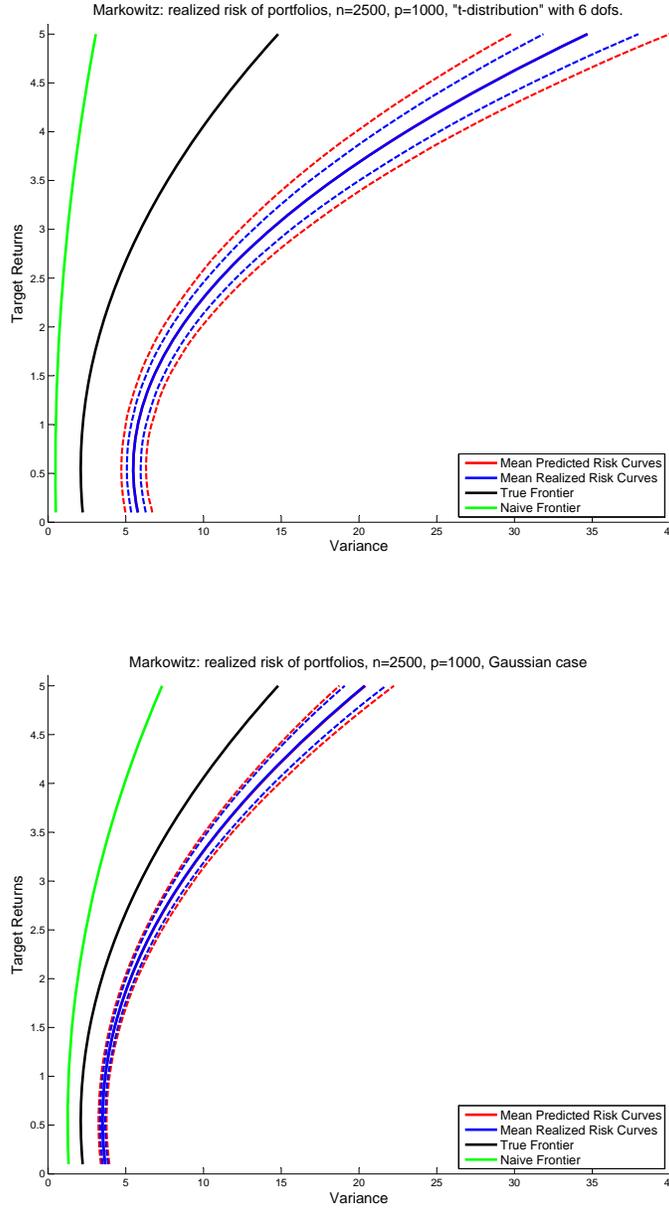


Figure 3: Performance of naive and corrected frontiers, for scaled “ t_6 ” (upper picture) and Gaussian returns. Here, $n = 2500$ and $p = 1000$. The number of simulations is 1000 in all pictures. The dashed lines represent (empirical) 95% confidence bands. (The confidence bands are computed for a fixed level of expected returns y .) The x -axis represents our estimate of the realized variance of the optimal portfolios. The y -axis represents the target returns for the portfolios. The plots show both the average realized risk of the naive Markowitz portfolios (blue curves) and the fact that our estimator (in red) is nearly unbiased (red solid curves below the blue curves and not visible on the plots). They also illustrate the robustness of our corrections. Another striking feature is the lack of robustness of Gaussian computations, since the average realized risk computed with “ t_6 ” returns are very different from the Gaussian ones. The fact that, as our theoretical work predicts, Gaussian computations leads to underestimation of the realized risk in the class of elliptical distributions considered in the paper is illustrated by the fact that the “ t_6 ” curves are to the right of the Gaussian curves. The population (or true) efficient frontier is in black. The green curve is the mean estimated risk, if estimated through the naive estimator $w'_{\text{emp}} \hat{\Sigma} w_{\text{emp}}$.

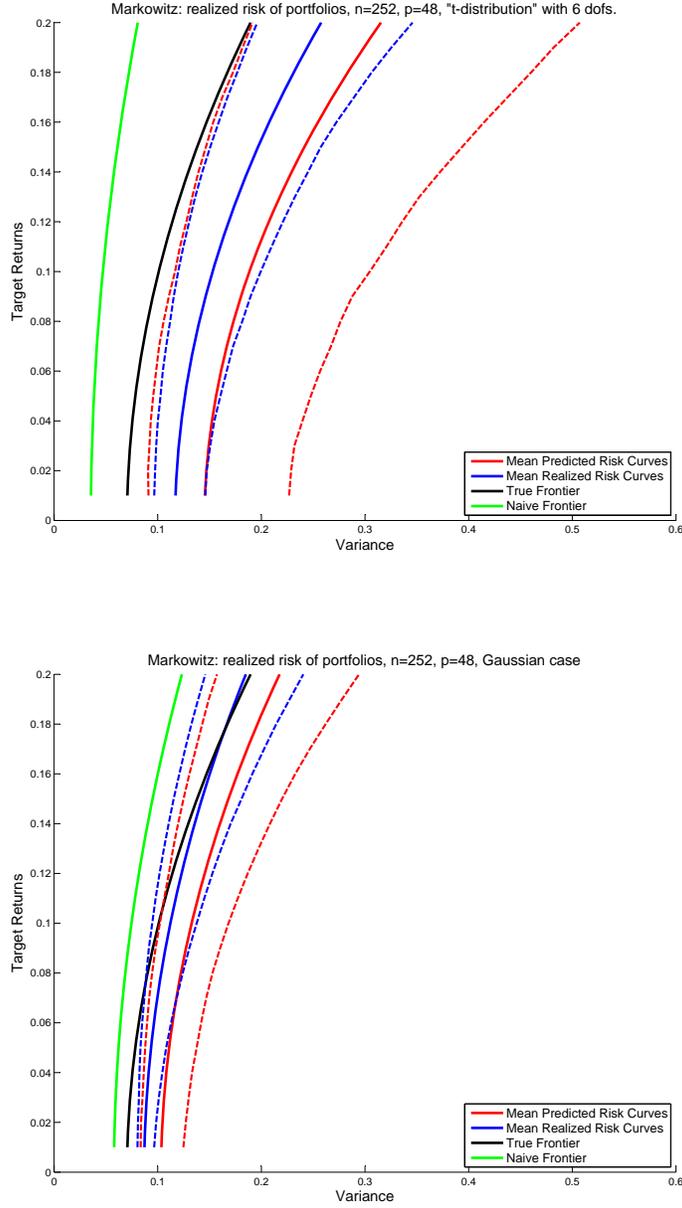


Figure 4: Performance of naive and corrected frontiers, for scaled “ t_6 ” (upper picture) and Gaussian returns. Here, $n = 252$ and $p = 48$ and the population parameters were computed from real data. The number of simulations is 1000 in all pictures. The dashed lines represent (empirical) 95% confidence bands. (The confidence bands are computed for a fixed level of expected returns y .) The x -axis represents our estimate of the realized variance of the optimal portfolios. The y -axis represents the target returns for the portfolios. The plots show both the average realized risk of the naive Markowitz portfolios (blue curves) and the average of our estimators (in red). Here our estimator is still biased. The lack of robustness of Gaussian computations is again highlighted. The fact that, as our theoretical work predicts, Gaussian computations leads to underestimation of the realized risk in the class of elliptical distributions considered in the paper is illustrated by the fact that the “ t_6 ” (blue) curves are to the right of the Gaussian curves. The population efficient frontier is in black. The green curve is the mean estimated risk, if estimated through the naive estimator $w'_{\text{emp}} \widehat{\Sigma} w_{\text{emp}}$. The variance of our corrections can be quite large, as seen in the t_6 simulations.