Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods

Noureddine El Karoui[*]   Holger Kösters[†]

Abstract

Shrinkage estimators of covariance are an important tool in modern applied and theoretical statistics. They play a key role in regularized estimation problems, such as ridge regression (aka Tykhonov regularization), regularized discriminant analysis and a variety of optimization problems

In this paper, we bring to bear the tools of random matrix theory to understand their behavior, and in particular, that of quadratic forms involving inverses of those estimators, which are important in practice.

We use very mild assumptions compared to the usual assumptions made in random matrix theory, requiring only mild conditions on the moments of linear and quadratic forms in our random vectors. In particular, we show that our results apply for instance to log-normal data, which are of interest in financial applications.

Our study highlights the relative sensitivity of random matrix results (and their practical consequences) to geometric assumptions which are often implicitly made by random matrix theorists and may not be relevant in data analytic practice.

1 Introduction

Modern multivariate statistics is increasingly high-dimensional. It is now easy to collect many samples (n) with a large number of covariates (p) for each sample. In this paper, we will therefore study multivariate statistical problems in the "large n, large p" setting that is increasingly popular in theoretical statistics. By this we mean that we will study certain statistics in the asymptotic setting where n, the number of observations, is going to infinity, and p, the number of predictors, is also going to infinity. Our focus will be on the situation where p/n remains bounded.

The paper is mostly concerned with forms involving the inverse of a shrunken covariance matrix, or powers of this inverse as they play a key role in several important statistical problems that we review later in this introduction. As a matter of fact, these objects, in one form or another, are central in many aspects of classical regularized methods in statistics and other fields of applied mathematics. The purpose of this paper is to explain how these regularized estimators behave in the "large p, large n" setting and derive some understanding and insights about the behavior of widely used methods that rely on them.

In classical statistics, when $p \ll n$, one can get a good estimate of the spectral properties of Σ, the population covariance matrix, by using its "naive" counterpart, the sample covariance matrix $\widehat{\Sigma}$, with, if $\widehat{\mu}$ is the sample mean of our vectors,

$$\widehat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \widehat{\mu})(X_i - \widehat{\mu})' \ .$$

As is now well-known, this is not the case when p is comparable to n, which we denote by $p \asymp n$. In that setting, even though the central limit theorem and a little bit of concentration of measure guarantee under broad assumptions that

$$\max_{i,j} |\widehat{\Sigma}(i,j) - \Sigma(i,j)| \to 0 \, ,$$

(even when $p \gg n$), the eigenvalues of $\widehat{\Sigma}$ tend to be very different from those of Σ (see Johnstone (2001) or the reviews Johnstone (2007), El Karoui (2011)). Hence, it is important to understand the performance of our standard techniques in this new asymptotic setting.

Recent papers concerned with these types of problems and their implications for concrete applications are for instance El Karoui (2009b) and El Karoui (2009c), where the author showed that somewhat surprisingly for a broad class of covariance matrices, means and distributions for the data, one could characterize the performance of estimators as a function of the ratio p/n, and hence get consistent estimators for parameters, such as the efficient frontier in classical portfolio theory, that appear difficult to estimate without structural assumptions on the population parameters. In these papers, the regularization came under the form of linear constraints on the vector of interest.

As mentioned before, shrinkage estimators of covariance are fundamental objects in modern statistics, partly because of James-Stein type phenomena (Haff (1980)) and they are very widely used. Here are a few examples.

1. Classification (LDA, RDA): when we observe data coming from two Gaussian populations, with different means μ_1 and μ_2, priors π_1 and π_2 but same covariance matrix Σ, the optimal classification rule is known to be Fisher's linear discriminant analysis rule: classify an observation x to class 2, if

$$x'\Sigma^{-1}(\mu_1 - \mu_2) > T(\mu_1, \mu_2, \Sigma, \pi_1, \pi_2) \, ,$$

where $T(\mu_1, \mu_2, \Sigma, \pi_1, \pi_2)$ is a known threshold. Naturally, we do now know Σ in practice, so a natural method is to replace it by $\widehat{\Sigma}$. This is what is usually done in LDA (see Hastie et al. (2009)). In Friedman (1989), concerned by, among other things variance issues in LDA, Friedman proposed to use RDA, regularized discriminant analysis, where instead of using $\widehat{\Sigma}$ as an estimate of Σ, one uses $\widehat{\Sigma} + A$ or $(1-\theta)\widehat{\Sigma} + \theta A$, i.e a shrinkage estimator of covariance. This estimator has also been proposed by Ledoit and Wolf (2004) in another context. It is natural to ask what happens when using these estimators in high-dimension.

2. Shrinkage estimators of covariance: a classic paper on the topic is Haff (1980); we also refer to Anderson (2003), for explanations concerning the benefit of skrinkage. In portfolio optimization, at least in the traditional mean-variance framework, similar issues arise. Hence partly motivated by this problem, Ledoit and Wolf (2004) proposed to use a shrinkage estimator to solve the portfolio optimization problem and get regularized solutions. In the finance literature, there are "finance-driven" shrinkage estimators, like the one arising in the Black-Litterman model (see Meucci (2005)).

3. Regression problems: in ridge regression, where one seeks β to optimize $\|Y - X\beta\| + \lambda\beta'\Gamma\beta$, one also encounters matrices of the form $\widehat{\Sigma} + \lambda\Gamma$, which is a shrunken version of $\widehat{\Sigma}$. The Γ that is usually taken is Id, this regularization amounts to modifying the eigenvalues of $\widehat{\Sigma}$.

In the analysis of all these methods, one needs to understand the behavior of the matrix $(\widehat{\Sigma} + A)^{-1}$ (entrywise and/or globally) as well as similar quantities involving $(\widehat{\Sigma} + A)^{-1}\Sigma_\epsilon(\widehat{\Sigma} + A)^{-1}$ (where Σ_ϵ is positive semidefinite) and this will be one of the focuses of the paper. It is tantalizing to use random matrix theory to do so, a program we got started on in El Karoui (2009b) and El Karoui (2009c). However, as documented in these papers, random matrix theory has several potential pitfalls: standard random matrix models, though in appearance general, put implicitly very strong geometric constraints on the datasets they are supposed to model. In light of this, one might be wary that the remarkable results that come out of it are just consequences of this geometry, which may or may not be similar to the one a practitioner encounters in practice. Hence we feel that any analysis that is not doing a meaningful robustness analysis is sorely lacking.

As we have documented before, the geometric constraints put by classical random matrix theory on the datasets modeled by it are due to manifestations of the concentration of measure phenomenon. Hence, it seems to us that a good starting point for the analysis of shrunken covariance matrices and their applications is that of generalized elliptical distributions, where the data is modeled as

$$\mathfrak{X}_i = \mu + R_i X_i \ ,$$

where R_i is a random variable independent of X_i and X_i has some (mild) concentration properties. (This will be made clear and precise later.)

The advantage of this class of models is that it contains the Gaussian model that is popular with many researchers, though now understood to be lacking in many fundamental ways. When $\mathbf{E}\left(R_i^2\right) = 1$, then $\mathrm{cov}\left(X_i\right) = \mathrm{cov}\left(\mathfrak{X}_i\right)$, so we can study robustness of our results in this class, since all the population parameters (which will depend on covariance and mean) will be the same.

However, by studying the model at this level of generality, we will not be able to rely on various invariance properties of the Gaussian distribution, and hence will really use only the geometric/concentration properties of the random variables of interest. One advantage of such an approach is that these properties are somewhat checkable in practice, through simple histograms for e.g norms and scalar products of points in the dataset, as has been explained before in some of the works cited above. Crucially, by showing that the results depend on the properties of $\{R_i\}_{i=1}^n$, we will able to show that even in our simple setting the geometry is key (change in R_i's may mean change in the geometry) and a major contributing factor in the robustness of the results. Finally, it should be noted (see El Karoui (2009b)) that one can sometimes study the bootstrap properties of various estimators by studying the class of elliptical distributions. Hence our analysis could be used to gain insight into bootstrap properties of various estimators.

The focus of our paper will mostly be on entrywise properties of $(\widehat{\Sigma} + A)^{-1}$ or $(\widehat{\Sigma} + A)^{-1}\Sigma_\epsilon(\widehat{\Sigma} + A)^{-1}$ in the class of models we consider, which naturally appear in the study of the risk of certain procedures. Quadratic forms involving the sample mean are also important in practice and will be studied. Random matrix theory already handles well things like $\mathrm{trace}\left((\widehat{\Sigma} + A)^{-1}\right)$, and other questions concerning only eigenvalues, so we will not spend too much time on this, though they are potentially important in the study of the risk of various estimators.

Beside shedding light on central statistical questions in multivariate analysis, our analysis also proposes what we think is a good and generic technical framework for carrying them out: namely we will do our work through invariance principles and mild concentration work. We will show that the statistics we are considering are asymptotically non-random, by showing that they are concentrated around their mean. And then we will show that the mean is the "same" in a broad class of models by using techniques akin to the Lindeberg method. A main difficulty is then to compute the mean (in many problems it is much harder to compute the mean of a statistic than to show that e.g its variance goes to zero), but our analysis will show that it can be done for favorable distributions in the class considered, and the Gaussian distribution will then be heavily used. Importantly, our analysis is very general and shows robustness even in classes where we have not or cannot at this point compute a limit for the quantity of interest.

We should also point out that our concentration requirements on X_i have purposely been kept to a minimum and hence our results extend way beyond the traditional "linear combination of i.i.d" framework which has been popular in random matrix theory following the nice work of Bai and Silverstein (see e.g Silverstein (1995), Silverstein and Bai (1995)). In particular, we will be able to handle (multivariate) log-normal distributions and other non-linear deformations of Gaussian random variables. Also, conditions on i.i.d-ness are essentially replaced by conditions on the mean and covariance of the random variables we deal with, as well as a little bit of concentration for linear and quadratic forms involving them. Our aim was also to show that these "universality" results could get obtained rather simply so an effort has been made to make the proofs as simple as possible. The paper is a bit long because we treat many cases in details and at what we think is the right level of generality.

Finally, it will be noted by researchers interested in probability that some of our results can be seen as strong versions of classic random matrix results: where classic results gave results about normalized traces of certain random matrices, we will be able to have statements valid for each element of the diagonal of the matrix of interest.

In section 2, we present some of our main technical results and heuristic justification for some of the main results, which should be helpful for statisticians wanting to get a sense of where the results come from. Section 3 contains most proofs and the core technical work. Section 4 discusses some potential applications to statistics, where at this point our main results shed light on existing procedures and "what they really do". We conclude in Section 5 and present a result of independent interest on Stieltjes transforms in the Appendix.

2 Strategy and exposition of some results

Our strategy is to make use of invariance principles and concentration inequalities throughout the paper. Practically, this translates into showing that the statistics we care about are concentrated around their means, that is the concentration part. In a second step, we show that this mean does not depend of the distribution of the data, as long as certain moment conditions are satisfied. To do so, we employ techniques very similar to the Lindeberg method (Stroock (1993) and let us note that it has been perhaps "re-popularized" by the nice work of Chatterjee in this direction, e.g Chatterjee (2005)).

Throughout the paper, we will focus on model of an elliptical type, namely we observe i.i.d observations

$$\mathfrak{X}_i = \mu + R_i X_i \ ,$$

where the X_i's are independent and independent of R_i. The R_i's are allowed to be dependent. Our efforts will go into relaxing distributional assumptions on X_i, while assuming only two moments on R_i - the justification for these choices coming from applications discussed at the end of the paper. In particular, this means that we will be able to handle data with relatively heavy tails.

A main tool in our work will be a simple extension of the Efron-Stein inequality - which will allow us to characterize higher moments of the statistics we care about. This extension is likely known in martingale theory but we present a proof in the appendix for the convenience of the reader. We delay it statement and presentation to the proof section and start by highlighting some of our main results.

2.1 A generalized version of the Efron-Stein inequality

We will make repeated use of the following lemma, which follows from Burkholder's inequality (see Burkholder (1973)).

Lemma. *Suppose $W = h(X_1, \ldots, X_n)$, where the X_i's are independent. We call $\mathcal{F}_j = \sigma(X_1, \ldots, X_j)$. We also denote by Z_m a (measurable) function of $(X_1, \ldots, X_{m-1}, X_{m+1}, \ldots, X_n)$.*

Then, we have, for a constant c that depends only on k, and for $k \geq 2$,

$$\mathbf{E}\left(|W - \mathbf{E}(W)|^k\right) \leq c\left(\mathbf{E}\left(\left[\sum_{m=1}^{n} \mathbf{E}\left((W - W_m)^2 | \mathcal{F}_{m-1}\right)\right]^{k/2}\right) + \sum_{m=1}^{n} \mathbf{E}\left(|W - W_m|^k\right)\right) \ . \quad (1)$$

The classic Efron-Stein inequality corresponds to the case where $k = 2$. The advantage of using higher k's is that it will for instance allow us to control $\max_{j \in J} |W_j - \mathbf{E}(W_j)|$ for J's of higher cardinalities. For instance, if we can show that $\mathbf{E}\left(|W_j - \mathbf{E}(W_j)|^k\right) \leq C n^{-k/2}$ for a certain k, a simple union bound gives us

$$P(\max_{j \in J} |W_j - \mathbf{E}(W_j)| > t) \leq \frac{C|J|}{(n^{1/2}t)^k} \ .$$

Hence a bound valid of $k > 2$ will allow us to handle greater J's. A number of applications (involving for instance thresholding) also require control of higher moments, which will be provided by our methods.

We also note that we purposely tried to avoid deriving central limit theorems. While those are definitely interesting, we wanted to have finite sample bounds and have them be relatively robust with respect to distributional assumptions, in keeping with what we view as their potential practical usefulness.

2.2 Quadratic forms in inverse of shrunken sample covariance matrices are essentially deterministic

We now state an application of the previous Lemma to forms which are at the center of our study.

Theorem. *Suppose $X_1, \ldots, X_n \in \mathbb{R}^p$ are independent. Suppose further that $\mathbf{E}(X_i) = 0$ and, if v is such that $\|v\| = 1$, $\mathbf{E}\left(|X_i'v|^k\right) \leq b_L(k; X_i)$, where $b_L(k; X_i)$ is a deterministic function depending only on the distribution of X_i and k. Call*

$$\mathcal{S} = \frac{1}{n} \sum_{i=1}^{m} R_i^2 X_i X_i',$$

where R_i are deterministic.

Call $M(t) = \mathcal{S} + A$, and assume that for some $t > 0$, A is positive definite, with $A \succeq t\mathrm{Id}_p$. Then, if $\|x\| = 1$,

$$\mathbf{E}\left(|x'[\mathcal{S}]^{-1}x - \mathbf{E}\left(x'[\mathcal{S}]^{-1}x\right)|^k\right) \leq \frac{c_k}{t^{2k}}\left[\left(\sum_{i=1}^{n}\left[\frac{R_i^4}{n^2}b(4; X_i) \wedge t^2\right]\right)^{k/2} + \left(\sum_{i=1}^{n}\left[\frac{R_i^{2k}}{n^k}b(2k; X_i) \wedge t^k\right]\right)\right].$$

It is perhaps instructive to give an example at this point. Here are two.

- Suppose that X_i satisfies $P(|X_i'v| > t) \leq C\exp(-ct^b)$, and X_i has mean 0. Then

$$b_L(k; X_i) \leq \frac{C}{c^{k/b}}\frac{k}{b}\Gamma\left(\frac{k}{b}\right).$$

- Suppose that X_i satisfies $P(|X_i'v| > t) \leq Ct^{-b}$. Then if $b > (k+1)$,

$$b_L(k; X_i) \leq C\left(1 + \frac{1}{b - (k+1)}\right).$$

We note that the condition on the X_i's is rather minimal: all we need is some concentration of linear forms in X_i, something that might seem surprising at first.

The exponential deviation inequality in our first example might look like a strong assumption. However, it is satisfied by many distributions, with quite non-linear structures which would be difficult to analyze if one did not resort to concentration of measure statements. The (centered) Gaussian copula is a good example. We give specific examples in Subsubsection 3.2.1.

The result also gives us a reasonable understanding of the size of the fluctuations behavior of the quadratic forms we are interested in. Note that using the Gaussian case (at $t = 0$) as a comparison, the fluctuation size of $n^{-1/2}$ seems to be the right one.

General strategy The general strategy is now clear. In light of the previous theorem, if we can get a good deterministic approximation to $\mathbf{E}\left(S(t)^{-1}\right)$, we will be able to get an approximation of $x'S(t)^{-1}x$. Note that the considerable simplification here is that we are not dealing with random variables anymore. Fortunately, we can approximate this expectation using variant of methods that have been developed in the random matrix literature (specifically the part of the theory concerned with understanding limiting spectral distributions). Also, it will be possible to show that these expectations do not vary much when we change some details of the distributions - this is the essence of Lindeberg-style ideas. Hence, all we will have to do is show that the expectations in question do not change much when we replace X_i's by Y_i's with a different distribution (but the same covariance and mean). And then compute the expectation in a favorable case, for instance when X_i's are Gaussian.

2.3 Heuristics

To help readers unfamiliar with random matrix theory understand better the results, we now present heuristics that help us guess the results. Formal proofs essentially start from these conjectures and proceed to verify that they are indeed correct.

We will focus on two types of quantities:

$$v'(\mathcal{S} + A)^{-1}v \text{ and } v'(\widehat{\Sigma} + A)^{-1}B(\widehat{\Sigma} + A)^{-1}v \ ,$$

where A and B are positive definite matrices.

Also, $\mathcal{S} = \frac{1}{n}\sum_{i=1}^n R_i^2 X_i X_i'$, where $X_i = \Sigma^{1/2}Y_i$, where Y_i has covariance Id_p, and X_i (or Y_i) satisfies mild concentration inequalities - the details are given when we undertake a rigorous proof. At this point, the reader can safely assume that X_i is $\mathcal{N}(0,\Sigma)$ (so Y_i is $\mathcal{N}(0,\mathrm{Id}_p)$). In other words, $\mathcal{S}$ is the "sample" covariance matrix we would use if we knew the mean of the data.

We have the following heuristic result:

Heuristic 2.1. *Under regularity conditions, we have*

$$v'(\mathcal{S} + A)^{-1}v \simeq v'(\gamma(A)\Sigma + A)^{-1}v \ ,$$

where if

$$\alpha(A) = \frac{1}{n}trace\left(\Sigma(\mathcal{S} + A)^{-1}\right) \ ,$$

$\alpha(A)$ *has an asymptotically deterministic equivalent and*

$$\gamma(A) \simeq \frac{1}{n}\sum_{i=1}^n \frac{R_i^2}{1 + R_i^2\alpha(A)} \ .$$

Argument: The key element of this argument is really the concentration of quadratic forms in Y_i, which allow us to replace quantities of the type $Y_i'MY_i/p$ by $\mathrm{trace}\,(M)\,/p = \mathbf{E}\left(Y_i'MY_i\right)/p$.

The fact that $\frac{1}{n}\mathrm{trace}\left(\Sigma(\mathcal{S} + A)^{-1}\right)$ has an asymptotically deterministic equivalent comes from standard arguments in random matrix theory (for some that rely on concentration and are just a few lines, see El Karoui (2009a); see also Subsection 3.5). Let us write $\mathcal{S} = \sum_{i=1}^n r_i r_i'$, where r_i are independent. Now, we have (using an idea akin to some in Silverstein (1995) and now classic in random matrix theory)

$$\mathcal{S}(\mathcal{S} + A)^{-1} = \mathrm{Id} - A(\mathcal{S} + A)^{-1} \ ,$$

and hence, using the fact that $(r_i r_i' + M_i)^{-1} = M_i^{-1} - \frac{M_i^{-1}r_i r_i' M_i^{-1}}{1 + r_i' M_i^{-1} r_i}$,

$$A(\mathcal{S} + A)^{-1} = \mathrm{Id} - \sum_{i=1}^n \frac{r_i r_i' M_i^{-1}}{1 + r_i' M_i^{-1} r_i} \ ,$$

where $M_i = \mathcal{S} + A - r_i r_i'$.

Therefore, if v and u are two vectors,

$$v'A(\mathcal{S} + A)^{-1}u = v'u - \sum_{i=1}^n \frac{v' r_i r_i' M_i^{-1} u}{1 + r_i' M_i^{-1} r_i} \ .$$

Now because Y_i satisfies a dimension-free concentration inequality, we have, if M is a matrix independent of Y_i, $Y_i'MY_i/p \simeq \mathrm{trace}\,(M)\,/p$. Applying this heuristic in each term of the previous sum, we get,

$$v'A(\mathcal{S} + A)^{-1}u = v'u - \frac{1}{n}\sum_{i=1}^n \frac{R_i^2 v'\Sigma M_i^{-1} u}{1 + R_i^2\frac{1}{n}\mathrm{trace}\left(\Sigma M_i^{-1}\right)} \ .$$

Now not much is lost by replacing M_i by $\mathcal{S} + A$ everywhere in the previous expression. Hence, we have heuristically,

$$v'A(\mathcal{S}+A)^{-1}u = v'u - \left[\frac{1}{n}\sum_{i=1}^{n}\frac{R_i^2}{1+R_i^2\frac{1}{n}\text{trace}\left(\Sigma(\mathcal{S}+A)^{-1}\right)}\right]v'\Sigma(\mathcal{S}+A)^{-1}u ,$$

$$= v'u - \gamma(A)v'\Sigma(\mathcal{S}+A)^{-1}u .$$

Another way of rewriting this equation is simply

$$v'(\mathcal{S}+A)^{-1}u = v'A^{-1}u - \gamma(A)v'A^{-1}\Sigma(\mathcal{S}+A)^{-1}u .$$

Now, let us call $v_k = (A^{-1}\Sigma)^k v$. Applying the previous heuristic to $v = v_k$ and $u = v$, we have if $\beta_k = v_k'(\mathcal{S}+A)^{-1}v$, and $\alpha_k = v_k'A^{-1}v$,

$$\beta_k \simeq \alpha_k - \gamma(A)\beta_{k+1} .$$

Assuming that we can use the previous approximation many times, we get

$$\beta_0 \simeq \sum_{j=0}^{n}(-\gamma(A))^j\alpha_j + (-\gamma(A))^{n+1}\beta_{n+1} .$$

Now assuming that we can sum the series and that $(\gamma(A))^{n+1}\beta_{n+1} \to 0$, we get

$$\beta_0 \simeq \sum_{j=0}^{\infty}(-\gamma(A))^j\alpha_j = v'\left[\sum_{j=0}^{\infty}(-\gamma(A))^j(A^{-1}\Sigma)^j\right]A^{-1}v$$

$$= v'(\text{Id} + \gamma(A)A^{-1}\Sigma)^{-1}A^{-1}v = v'(A + \gamma(A)\Sigma)^{-1}v .$$

Note that $\beta_0 = v'(\mathcal{S}+A)^{-1}v$. Hence, it is perhaps reasonable to conjecture that

$$v'(\mathcal{S}+A)^{-1}v \simeq v'(A + \gamma(A)\Sigma)^{-1}v .$$

Note that the heuristic also gives us conjectures for approximating the value of $v'(\mathcal{S}+A)^{-1}(A^{-1}\Sigma)^k v$, for any given k, as this is what we called earlier β_k. $\qquad\square$

For dealing with higher powers of $(\mathcal{S}+A)^{-1}$, we also need the following heuristic.

Heuristic 2.2. *Under regularity assumptions, we have*

$$v'(\mathcal{S}+A)^{-1}B(\mathcal{S}+A)^{-1}v \simeq v'(A+\gamma(A)\Sigma)^{-1}(B+\xi(A,B)\Sigma)(A+\gamma(A)\Sigma)^{-1}v ,$$

where $\gamma(A)$ is defined in Heuristic 2.2 and

$$\xi(A,B) = \left[\frac{1}{n}\sum_{i=1}^{n}\frac{R_i^4}{(1+R_i^2\alpha(A))^2}\right]\frac{1}{n}\text{trace}\left(\Sigma(\mathcal{S}+A)^{-1}B(\mathcal{S}+A)^{-1}\right) .$$

Furthermore, $\xi(A,B)$ has an asymptotically deterministic equivalent.

Argument : Let us call $f(t) = v'(\mathcal{S}+A(t))^{-1}v$. Then, since $([M(t)]^{-1})' = [M(t)]^{-1}M'(t)[M(t)]^{-1}$, we have

$$f'(t) = -v'(\mathcal{S}+A(t))^{-1}A'(t)(\mathcal{S}+A(t))^{-1}v .$$

Now, if we consider $A(t) = A + tB$, we see that $A'(t) = B$, and therefore,

$$f'(0) = -v'(\mathcal{S}+A)^{-1}B(\mathcal{S}+A)^{-1}v ,$$

which is the quantity we seek to approximate.

Now recall that from Heuristic 2.1, we gathered that

$$v'(\mathcal{S} + A)^{-1}v \simeq v'(A + \gamma(A)\Sigma)^{-1}v \, .$$

We might be tempted to look at this approximate equality as valid for any $A(t)$ and take the derivative with respect to t. Doing so, we would get, if $g(t) = v'(\mathcal{S} + A(t))^{-1}v$,

$$g'(0) = -v'(\mathcal{S} + A)^{-1}(B + \gamma(A(t))'(0)\Sigma)(\mathcal{S} + A)^{-1}v \, .$$

Now,

$$\gamma(A(t)) = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i^2}{1 + R_i^2\alpha(A(t))} \, .$$

Hence, if $h(t) = \gamma(A(t))$ and $k(t) = \alpha(A(t)) = \frac{1}{n}\mathrm{trace}\left(\Sigma(\mathcal{S} + A(t))^{-1}\right)$, we have

$$h'(0) = -k'(0)\frac{1}{n}\sum_{i=1}^{n}\frac{R_i^4}{(1 + R_i^2\alpha(A))^2} \, .$$

Now, $k'(t) = -\frac{1}{n}\mathrm{trace}\left(\Sigma(\mathcal{S} + A(t))^{-1}B(\mathcal{S} + A(t))^{-1}\right)$. Hence,

$$-k'(0) = \frac{1}{n}\mathrm{trace}\left(\Sigma(\mathcal{S} + A)^{-1}B(\mathcal{S} + A)^{-1}\right) \, ,$$

and we conclude that

$$h'(0) = \left[\frac{1}{n}\mathrm{trace}\left(\Sigma(\mathcal{S} + A)^{-1}B(\mathcal{S} + A)^{-1}\right)\right]\left[\frac{1}{n}\sum_{i=1}^{n}\frac{R_i^4}{(1 + R_i^2\alpha(A))^2}\right] = \xi(A, B) \, .$$

The fact that $\xi(A, B)$ is asymptotically non-random comes from the same ideas as described in Heuristic 2.1. □

In our applications, we will also need to understand quantities of the type $\widehat{\mu}'(\widehat{\Sigma} + A)^{-1}\widehat{\mu}$ (where $\widehat{\Sigma} = \mathcal{S} - \widehat{\mu}\widehat{\mu}'$) and $\widehat{\mu}'(\widehat{\Sigma} + A)^{-1}v$. We naturally treat those cases below and refer the reader to that part of the paper for information about these forms. The main issue is that when dealing with $\widehat{\Sigma}$ and $\widehat{\mu}$, a non-negligible interaction term between the two occurs (it is related to $\widehat{\mu}'(\mathcal{S} + A)^{-1}\widehat{\mu}$) and one needs to be a bit careful to treat it.

3 Results and proofs

This section contains the main technical aspects of the paper. In subsection 3.1, we discuss a simple extension of the Efron-Stein inequality. The rest of this section is devoted to showing concentration and invariance of the forms we care about. The method of proof is systematic: we first show concentration (i.e control of the variance or higher moments), and then show that the mean value to which we can reduce the problem does not depend on "details" of the distribution of the data through a Lindeberg-like argument.

Notations Before we proceed, let us set some notations. We denote by $|||M|||_2$ the operator norm (i.e largest singular value) of a matrix M. When dealing with several independent random variables $(X_1, \ldots, X_n)$, we use $\mathbf{E}_i\,()$ to denote expectation with respect to X_i only. We often use the abbreviation psd for positive semi-definite.

3.1 A simple extension of the Efron-Stein inequality

The strategy for our approach is to first show that the quadratic forms we care about, namely

$$v'(\mathcal{S} + A)^{-1}v \ , \ A \succeq t\mathrm{Id}_p \ ,$$

(and variants) are essentially deterministic asymptotically. Modern techniques can be adapted to then get (in simple cases compared to the generality level at which we will work) deterministic approximations of $v'(\mathcal{S} + A)^{-1}v$ and we can then use those to actually compute the limit of the aforementioned quadratic form. But it is important to get a systematic way of showing that for a certain class of random matrices $\mathcal{S}$,

$$v'(\mathcal{S} + A)^{-1}v \simeq v'\mathbf{E}\left((\mathcal{S} + A)^{-1}\right)v \ .$$

To do so, we propose to use (essentially) a martingale difference argument, which is not unknown in random matrix theory (Bai (1999), Girko (1990), and several others), but whose role may not have been as emphasized as it perhaps should have. However, at the level of generality at which we are working, our proofs become easier if we quickly branch away from standard methods. The following lemma is essentially an L^p variant of the Efron-Stein inequality (see Efron and Stein (1981), Theorem 2, and also Lugosi (2006), Theorem 9). It is surely known in martingale theory but we give a simple proof here for the convenience of the reader.

Lemma 3.1. *Suppose $W = h(X_1, \ldots, X_n)$, where the X_i's are independent. We call $\mathcal{F}_j = \sigma(X_1, \ldots, X_j)$. We also denote by W_m a (measurable) function of $(X_1, \ldots, X_{m-1}, X_{m+1}, \ldots, X_n)$.*

Then, we have, for a constant c that depends only on k, and for $k \geq 2$,

$$\mathbf{E}\left(|W - \mathbf{E}(W)|^k\right) \leq c\left(\mathbf{E}\left(\left[\sum_{m=1}^{n}\mathbf{E}\left((W - W_m)^2|\mathcal{F}_{m-1}\right)\right]^{k/2}\right) + \sum_{m=1}^{n}\mathbf{E}\left(|W - W_m|^k\right)\right) \ . \quad (2)$$

Note that in the case $k = 2$, we recover the Efron-Stein inequality

$$\mathrm{var}\,(W) \leq \sum_{m=1}^{n}\mathbf{E}\left((W - W_m)^2\right) \ ,$$

with a possibly worse constant.

In the applications we have in mind, through rank-1 update of inverses of matrices, we will easily get an approximation of Z by a function that does not involve the m-th variable and these results will come in particularly handy.

Proof of Lemma 3.1. We can clearly write $Z - \mathbf{E}(Z)$ as a sum of martingale differences: if

$$V_m = \mathbf{E}(Z|\mathcal{F}_m) - \mathbf{E}(Z|\mathcal{F}_{m-1}) \ ,$$

$$Z - \mathbf{E}(Z) = \sum_{m=1}^{n} V_m \ .$$

Note also that if Z_m is a (measurable) function of all the X_i's except X_m,

$$V_m = \mathbf{E}(Z - Z_m|\mathcal{F}_m) - \mathbf{E}(Z - Z_m|\mathcal{F}_{m-1}) \ ,$$

since $\mathbf{E}(Z_m|\mathcal{F}_m) = \mathbf{E}(Z_m|\mathcal{F}_{m-1})$.

Now let us call $s(Z) = [\sum_{m=1}^{n}\mathbf{E}\left(V_m^2|\mathcal{F}_{m-1}\right)]^{1/2}$. Recall that Burkholder's inequality implies (see Equation 21.5 in Burkholder (1973)) that, if Φ is a non-decreasing function on $[0, \infty]$ with $\Phi(0) = 0$ and $\Phi(2\lambda) \leq c_1\Phi(\lambda)$, then

$$\mathbf{E}(\Phi(Z)) \leq c\left(\mathbf{E}(\Phi(s(Z))) + \sum_{k=1}^{n}\mathbf{E}(\Phi(|V_m|))\right) \ .$$

As noted in Burkholder (1973), $\Phi(x) = x^k$ satisfies the conditions needed for the inequality to hold. Let us remind the reader that it is well known (see Lugosi (2006), p.16) that

$$V_m^2 \leq \mathbf{E}\left((Z - \mathbf{E}_m(Z))^2 | \mathcal{F}_m\right) ,$$

where $\mathbf{E}_m(\cdots)$ is expectation with respect to X_m only, i.e $\mathbf{E}_m(Z) = \mathbf{E}(Z | X_1, \ldots, X_{m-1}, X_{m+1}, \ldots, X_n)$. Also, as noted for instance in Lugosi (2006),

$$\mathbf{E}_m\left((Z - \mathbf{E}_m(Z))^2\right) \leq \mathbf{E}_m\left((Z - Z_m)^2\right) ,$$

where Z_m is any measurable function of $X_1, \ldots, X_{m-1}, X_{m+1}, \ldots, X_n$. We note that

$$\mathbf{E}(\cdot | \mathcal{F}_{m-1}) = \mathbf{E}(\mathbf{E}_m(\cdot) | \mathcal{F}_{m-1}) .$$

Therefore,

$$\mathbf{E}\left(V_m^2 | \mathcal{F}_{m-1}\right) \leq \mathbf{E}\left([Z - \mathbf{E}_m(Z)]^2 | \mathcal{F}_{m-1}\right) \leq \mathbf{E}\left(\mathbf{E}_m\left([Z - \mathbf{E}_m(Z)]^2\right) | \mathcal{F}_{m-1}\right) \leq \mathbf{E}\left((Z - Z_m)^2 | \mathcal{F}_{m-1}\right) ,$$

and we have

$$s(Z) \leq \sqrt{\sum_{m=1}^n \mathbf{E}\left((Z - Z_m)^2 | \mathcal{F}_{m-1}\right)} .$$

Hence, because Φ is non decreasing,

$$\mathbf{E}(\Phi(s(Z))) \leq \mathbf{E}\left(\Phi\left[\sqrt{\sum_{m=1}^n \mathbf{E}\left((Z - Z_m)^2 | \mathcal{F}_{m-1}\right)}\right]\right)$$

Now let us turn our attention to $\mathbf{E}(\Phi(|V_m|))$, specifically when $\Phi(x) = x^k$. Since $V_m = \mathbf{E}(Z - Z_m | \mathcal{F}_m) - \mathbf{E}(Z - Z_m | \mathcal{F}_{m-1})$,

$$|V_m|^k \leq 2^{k-1}\left(|\mathbf{E}(Z - Z_m | \mathcal{F}_m)|^k + |\mathbf{E}(Z - Z_m | \mathcal{F}_{m-1})|^k\right) .$$

Also, when $k \geq 1$, $|x|^k$ is convex, so Jensen's inequality implies that

$$|\mathbf{E}(Z - Z_m | \mathcal{F}_m)|^k \leq \mathbf{E}\left(|Z - Z_m|^k | \mathcal{F}_m\right) .$$

Therefore,

$$\mathbf{E}\left(|V_m|^k\right) \leq 2^k \mathbf{E}\left(|Z - Z_m|^k\right)$$

Equation (2) now follows easily. $\qquad\square$

We note that if we were willing to make stronger assumptions on the data that the ones we will make, we could rely on other concentration inequalities to obtain for instance Gaussian concentration for some of the statistics we are interested in. However, since our study is a robustness study, we made the choice of making weaker assumptions and consequently to have possibly worse concentration inequalities - though of course this allows us to show that our first order results hold for a wider class of distributions.

3.2   Setup of our study

In all that follows we make the following assumptions, which we will casually call "our usual assumptions".

- We assume that p/n remains bounded away from 0 and ∞, i.e $p \sim n$.

- the random variables X_j and Y_j which will appear below have the same covariance matrix, Σ_j, and same mean, 0.

- Y_j's are independent and so are X_j's.

- Y_j's are independent of X_j's

- If v is any fixed vector with norm 1, we have, for $k \geq 1$,

$$\mathbf{E}\left(|X_i'v|^k\right) \leq b_L(k; X_i) \tag{3}$$

- If M is any deterministic and positive semidefinite matrix with $|||M|||_2 \leq 1$,

$$\mathbf{E}\left(|X_j'MX_j - \mathbf{E}\left(X_j'MX_j\right)|^k\right) \leq b_{Q_2}(k; X_j) . \tag{4}$$

- The matrix towards which we shrink, A, is such that $A \succeq t\mathrm{Id}_p$.

Let us note that by Jensen's inequality, there is no loss in generality in assuming that $b_L(k, X_i) \leq \sqrt{b_L(2k; X_i)}$. We will assume this throughout this paper, as this will occasionally be needed to merge certains bounds arising in our estimates, and thus to shorten our formulas.

Also we note that if $A \succeq t\mathrm{Id}$ and $\Sigma_0 \succeq 0$, for any $x \in \mathbb{R}^p$, we have

$$x'(A + \Sigma_0)^{-2}x \leq \frac{1}{t}x'A^{-1}x ,$$

which is easily seen since $M \mapsto M^{-1}$ is monotone (and decreasing with respect to the Loewner's order), so $(A + \Sigma_0)^{-1} \preceq t^{-1}\mathrm{Id}$; now multiplying on both sides by $(A + \Sigma_0)^{-1/2}$, the inequality (and its order) is preserved and we conclude that $(A + \Sigma_0)^{-2} \preceq t^{-1}(A + \Sigma_0)^{-1} \preceq t^{-1}A^{-1}$.

Finally, let us give some order of magnitude bounds. b_L will generally be very easy to control, as it is a linear form in X_i. For instance, if $X_i \sim \mathcal{N}(0, \mathrm{Id}_p)$, we have $X_i'v \sim \mathcal{N}(0, \|v\|)$, so $b_L(X_i; k)$ is of order 1 for all (finite) k. When X_i is $\mathcal{N}(0, \mathrm{Id}_p)$, $X_i'MX_i$ is a weighted χ^2, since $X_i'MX_i \overset{\mathcal{L}}{=} \sum_{k=1}^p \xi_k^2 \lambda_k(M)$ where ξ_k are $\mathcal{N}(0, 1)$ and independent. Hence, we conclude that $b_{Q_2}(k; X_i)$ is of order at most $p^{k/2}$ in this case. The informal bounds we will have in mind are therefore

$$b_L(k; X_i) = \mathrm{O}(1) ,$$
$$\frac{b_{Q_2}(k; X_i)}{p^{k/2}} = \mathrm{O}(1) \left(= \frac{b_{Q_2}(k; X_i)}{n^{k/2}}\right) ,$$

where the last statement comes from the fact that $p \sim n$.
We further note that if Σ is a covariance matrix,

$$b_L(k; \Sigma^{1/2}X_i) \leq |||\Sigma|||_2^{k/2} b_L(k; X_i) ,$$
$$b_{Q_2}(k; \Sigma^{1/2}X_i) \leq |||\Sigma|||_2^k b_{Q_2}(k; X_i) .$$

To bound b_{Q_2} in certain situations, it will be simpler to work through an auxiliary quantity, b_{Q_1}. Let us define it as, if M is any deterministic (psd) matrix with $|||M|||_2 \leq 1$,

$$\mathbf{E}\left(|\sqrt{Y_j'MY_j} - \mathbf{E}\left(\sqrt{Y_j'MY_j}\right)|^k\right) \leq b_{Q_1}(k; Y_j) .$$

Connection between b_{Q_1} and b_{Q_2}. b_{Q_1} and b_{Q_2} are of course very closely related. Also, in a concentration context, because $y \mapsto \sqrt{y'My}$ is Lipschitz with respect to Euclidian norm and convex, it is possible to derive b_{Q_1} for many distributions for which it would be otherwise difficult. For instance Gaussian concentration immediately implies deviation bounds and hence bounds on b_{Q_1} for e.g. centered Gaussian copulas.

Let us now elaborate on the relationship between b_{Q_1} and b_{Q_2}. Let us call $Q_M(Y) = Y'MY$, $q_M(Y) = \sqrt{Q_M(Y)}$, $\Delta_M(Y) = Q_M(Y) - \mathbf{E}\left(Q_M(Y)\right)$ and $\delta_M(Y) = \sqrt{Q_M(Y)} - \mathbf{E}\left(\sqrt{Q_M(Y)}\right)$, i.e $\delta_M(Y) = q_M(Y) - \mathbf{E}\left(q_M(Y)\right)$. Clearly,

$$\Delta_M(Y) = (q_M^2(Y) - [\mathbf{E}\left(q_M(Y)\right)]^2) + [\mathbf{E}\left(q_M(Y)\right)]^2 - \mathbf{E}\left(Q_M(Y)\right)$$
$$= \delta_M(Y)\left[\delta_M(Y) + 2\mathbf{E}\left(q_M(Y)\right)\right] + [\mathbf{E}\left(q_M(Y)\right)]^2 - \mathbf{E}\left(Q_M(Y)\right)$$
$$= \delta_M(Y)\left[\delta_M(Y) + 2\mathbf{E}\left(q_M(Y)\right)\right] - \mathrm{var}\left(q_M(Y)\right) .$$

Using convexity of $x \mapsto |x|^k$, we conclude that

$$|\Delta_M(Y)|^k \le 3^{k-1} \left[|\delta_M(Y)|^{2k} + 2^k |\delta_M(Y)|^k [\mathbf{E}\,(q_M(Y))]^k + [\mathrm{var}\,(q_M(y))]^k \right]$$

$$\le 3^{k-1} \left[|\delta_M(Y)|^{2k} + 2^k |\delta_M(Y)|^k [\mathbf{E}\,(Q_M(Y))]^{k/2} + [\mathrm{var}\,(q_M(y))]^k \right]$$

Now note that $\mathbf{E}\,(Q_M(Y)) = \mathrm{trace}\,(M\Sigma)$ and that $\mathrm{var}\,(q_M(y)) = b_{Q_1}(2; Y)$. So after taking expectations, we have shown that

$$b_{Q_2}(k; Y) \le 3^{k-1} \left[b_{Q_1}(2k; Y) + 2^k b_{Q_1}(k; Y) \left[\mathrm{trace}\,(M\Sigma) \right]^{k/2} + [b_{Q_1}(2; Y)]^k \right] .$$

Also, it is instructive to have a sense of the parameters that impact these bounds and how they grow. In the case of normality distributed random variables, $Q_M(Y)$ is a weighted χ^2 with p degrees of freedom, the weights being the eigenvalues of $\Sigma^{1/2} M \Sigma^{1/2}$. In this case, we have $b_{Q_2}(2; Y) = \sup_{M:|||M|||_2=1} 2\mathrm{trace}\left((\Sigma M)^2\right)$. When $|||M|||_2 = 1$, it is easy to see that $\mathrm{trace}\left((\Sigma M)^2\right) \le \mathrm{trace}\left(\Sigma^2\right)$, since if $A \preceq B$, and both are positive semi-definite, then $\mathrm{trace}\left(A^2\right) \ge \mathrm{trace}\left(B^2\right)$. Hence, $b_{Q_2} = 2\mathrm{trace}\left(\Sigma^2\right)$.

At this point, one might be concerned about the fact that these quantities will be dependent on extreme eigenvalues of Σ. However, in some situations, we can mitigate this problem. For instance, in the case where we assume that the data are i.i.d with the same covariance Σ, it will sometime be possible to work with Y having covariance Id, by simply replacing the shrinkage factor A by $\Sigma^{-1/2} A \Sigma^{-1/2}$, and the vector x at which we evaluate the shrunken matrix by $\Sigma^{-1/2} x$. This is the case for instance when considering $x'(\widehat{\Sigma} + A)^{-1} x$.

3.2.1 Meaningfulness of the assumptions and applicability

It is of course important to check that the assumptions we make can be applied to a wide variety of situations. It is therefore instructive to give examples at this point. Here are two.

- Suppose that X_i satisfies $P(|X_i'v| > t) \le C \exp(-ct^b)$, and X_i has mean 0. Then

$$b_L(k; X_i) \le \frac{C}{c^{k/b}} \frac{k}{b} \Gamma\left(\frac{k}{b}\right) .$$

- Suppose that X_i satisfies $P(|X_i'v| > t) \le C t^{-b}$. Then if $b > (k+1)$,

$$b_L(k; X_i) \le C \left(1 + \frac{1}{b - (k+1)}\right) .$$

We note that the condition on the $b_L(k; X_i)$'s is rather minimal: all we need is some concentration of linear forms in X_i.

The exponential deviation inequality might look like a strong assumption. However, it is satisfied by many distributions, with quite non-linear structures which would be difficult to analyze if one did not resort to concentration of measure arguments (see Ledoux (2001) for a very thorough reference, and see for instance El Karoui (2009a) for spelled-out examples). For the convenience of the reader, here are some examples taken from this last reference (justifications can be found there):

- Gaussian random variables, with $|||\Sigma|||_2$ bounded for instance. (Note that this can be relaxed considerably.)

- Vectors of the type $\sqrt{p}r$ where r is uniformly distributed on the unit (ℓ_2-) sphere is dimension p.

- Vectors $X = \Gamma \sqrt{p}r$, with r uniformly distributed on the unit (ℓ_2-)sphere in $\mathbb{R}^p$ and with $\Gamma\Gamma' = \Sigma$ with e.g. $|||\Sigma|||_2$ bounded.

- Vectors of the type $X = p^{1/b}r$, $1 \le b \le 2$, where r is "uniformly" sampled in the 1-ℓ^b ball or sphere in $\mathbb{R}^p$. (See Ledoux (2001), Theorem 4.21, which refers to Schechtman and Zinn (2000) as the source of the theorem and explains the details of the sampling.)

- Vectors X with log-concave density of the type $\mathrm{e}^{-U(x)}$, with the Hessian of U satisfying, for all x, $\mathrm{Hess}(U) \geq c\mathrm{Id}_p$ (see Ledoux (2001), Theorem 2.7.) For simplicity, though it may not be needed, one can assume that $|||\Sigma|||_2$ remains bounded.

- Vectors (X) distributed according to a (centered) Gaussian copula, with corresponding correlation matrix, Σ, having $|||\Sigma|||_2$ bounded. In other words, if $Z \sim \mathcal{N}(0, R)$, $X = \Phi(Z) - 1/2$, where Φ is the cdf of the standard Gaussian random variables.

- Vectors $X = \Sigma^{1/2}Y$, where Y has i.i.d bounded entries . See Corollary 4.10 in Ledoux (2001) for the concentration part. Here we crucially need the fact that the concentration of measure results we rely on are valid for convex 1-Lipschitz function (and we do not need them for all Lipschitz functions).

- More "exotic" examples involving vectors sampled uniformly from certain Riemannian submanifolds of $\mathbb{R}^p$. We refer to Ledoux (2001) Theorems 2.4 and 3.1 for the concentration aspects for these questions.

Bounding of b_{Q_2} can either be done directly or using the connection (and bound) between b_{Q_2} and b_{Q_1} we just made explicit. If X_i satisfies a concentration inequality for convex Lipschitz functions, then bounding b_{Q_1} is rather simple and this gives us a bound on b_{Q_2}. We now work out the details of this problem. The analysis is standard and follows along the lines of work done in e.g. Ledoux (2001), Chapter 1.

An important example: case of concentrated random variables As a matter of fact, suppose that X_i is such that for any convex and 1-Lipschitz function f, if $X \overset{\mathcal{L}}{=} X_i$,

$$P(|f(X) - \mathbf{E}\,(f(X))\,| > t) \leq C \exp(-ct^b) \ \text{ or } \ P(|f(X) - \mathrm{median}\,(f(X))\,| > t) \leq C \exp(-ct^b)$$

Since $f_v(X) = X'v$ is trivially convex and $\|v\|$-Lipschitz, we see that if the concentration inequality is around the mean, we immediately have

$$b_L(k; X_i) \leq \frac{C}{c^{k/b}} \frac{k}{b} \Gamma\left(\frac{k}{b}\right) \ .$$

If we "only" have a concentration bound around the median, then we can simply use

$$\mathbf{E}\left(|X'v|^k\right) \leq 2^{k-1} \left(\mathbf{E}\left(|X'v - \mathrm{median}\,(X'v)\,|^k\right) + |\mathrm{median}\,(X'v)\,|^k\right) \ .$$

The concentration inequality gives us control of the first term, while $|\mathrm{median}\,(X'v)\,| = |\mathrm{median}\,(X'v) - \mathbf{E}\,(X'v)\,|$ which is also controlled (see Proposition 1.9 in Ledoux (2001)) or simply

$$|\mathrm{median}\,(X'v) - \mathbf{E}\,(X'v)\,| \leq \mathbf{E}\left(|X'v - \mathrm{median}\,(X'v)\,|\right) = \int_0^\infty P(|X'v - \mathrm{median}\,(X'v)\,| > t)dt \leq C \int_0^\infty \exp(-ct^b dt) \ .$$

This is of course nothing else than $C\Gamma(1/b)/(bc^{1/b})$, and so we have a uniform bound.

Similarly, when M is a positive definite matrix with $|||M|||_2 \leq 1$, $\sqrt{X_i'MX_i}$ is a convex 1-Lipschitz function (with respect to Euclidian norm for X_i). Using the fact that for a non-negative random variable Z, $\mathbf{E}\left(Z^k\right) = \int_0^\infty kx^{k-1}P(Z \geq x)dx$, we see that, if our concentration result is around the mean,

$$b_{Q_1}(X_i; k) = \mathbf{E}\left(\left|\sqrt{X_i'MX_i} - \mathbf{E}\left(\sqrt{X_i'MX_i}\right)\right|^k\right) \leq C \int_0^\infty kx^{k-1}\exp(-cx^b)dx = \frac{C}{c^{k/b}}\frac{k}{b}\Gamma\left(\frac{k}{b}\right) \ .$$

Hence, when X_i satisfy a dimension-free concentration inequality, $b_{Q_1}(k; X_i)$ remains bounded uniformly in p and n. Therefore, when $\mathrm{trace}\,(\Sigma)/n$ remains bounded as n grows, so does $b_{Q_2}(k; X_i)/n^{k/2}$, thanks to the relationship between b_{Q_1} and b_{Q_2} we have highlighted above.

The conclusion of this short discussion is that random variables satisfying a dimension free concentration inequality and having covariance such that $\{\mathrm{trace}\,(\Sigma_i)/n\}_{i=1}^n$ remains uniformly bounded in n and p will have $b_{Q_2}(2; X_i)/n$ and $b_L(4; X_i)$ uniformly bounded (in n). Because we will express later our various bounds in terms of these quantities, this observation is very important from the point of view of the applicability of our results.

An important distribution in practice (in particular in financial applications) is the log-normal distribution. Getting bounds for b_L and b_{Q_2} here requires work which we now perform.

3.2.2   The case of the log-normal distribution

Let $Z = (Z_1, \ldots, Z_p)$ be a random vector with a normal distribution with parameters $\tilde{\mu} = (\tilde{\mu}_i)$ and $\tilde{\Sigma} = (\tilde{\sigma}_{ij})$. Then the random vector $Y := (Y_1, \ldots, Y_p)$ with $Y_i := \exp(Z_i)$, $i = 1, \ldots, p$, is said to have a log-normal distribution with parameters $\tilde{\mu}$ and $\tilde{\Sigma}$ (see e.g. Mardia, Kent and Bibby (1979), Chapter 2.6). Note that the moments of the log-normal distribution are all finite, and can be obtained from the moment generating function of the normal distribution. Indeed, for any $t = (t_1, \ldots, t_p) \in \mathbb{N}_0^p$, we have

$$\mathbf{E}(Y_1^{t_1} \ldots Y_p^{t_p}) = \mathbf{E}(\exp(t'Z)) = \exp(t'\tilde{\mu} + \tfrac{1}{2}t'\tilde{\Sigma}t). \tag{5}$$

Set $\tilde{\mu}_* := \|\tilde{\mu}\|_2$ and $\tilde{\sigma}_*^2 := |||\tilde{\Sigma}|||_2$. Then, for any $t = (t_1, \ldots, t_p) \in \mathbb{N}_0^p$, we have the estimate

$$\mathbf{E}(Y_1^{t_1} \ldots Y_p^{t_p}) \le \exp(\|t\|_2 \tilde{\mu}_* + \tfrac{1}{2}\|t\|_2^2 \tilde{\sigma}_*^2). \tag{6}$$

Put $X := Y - \mathbf{E}(Y)$ (where the expectation is taken componentwise, of course). In this section we will derive bounds for the constants $b_L(2r, X)$ and $b_{Q_2}(2, X)$ associated with the (centered) log-normal distribution.

In the sequel we always assume that $Z = \tilde{\mu} + \tilde{\Sigma}^{1/2}\bar{Z}$, where $\bar{Z}$ is a p-dimensional Gaussian random vector with zero mean and identity covariance. Our derivation will be based on the following result for the Gaussian distribution (Pisier, 1986, Chapter 2): If F is a continuously differentiable function and ∇F is the gradient of F (which we always regard as a column vector), then, for any $r \ge 1$,

$$\mathbf{E}|F(\bar{Z}) - \mathbf{E}(F(\bar{Z}))|^r \le K_r(\tfrac{\pi}{2})^r \mathbf{E}\|\nabla F(\bar{Z})\|_2^r,$$

where K_r is the rth moment of the standard Gaussian distribution.

For any $z = (z_i) \in \mathbb{R}^p$, let $\exp(z) := (\exp(z_i)) \in \mathbb{R}^p$ (by slight abuse of notation), and note that this vector-valued version of the exponential function is continuously differentiable and its Jacobian matrix $D(z)$ is diagonal with the elements $\exp(z_i)$ on the main diagonal. With this notation, $Y = \exp(Z) = \exp(\tilde{\mu} + \tilde{\Sigma}^{1/2}\bar{Z})$, and we get, for any $r \ge 1$,

$$\mathbf{E}|F(Y) - \mathbf{E}(F(Y))|^r \le K_r(\tfrac{\pi}{2})^r \mathbf{E}\|\nabla F(Y)'D(Z)\tilde{\Sigma}^{1/2}\|^r.$$

We now specialize this result to linear and quadratic forms.

Linear Forms. Consider the linear form $F(y) := v'y$, where $v = (v_i)$ is a deterministic vector with Euclidean norm 1. Then $\nabla F(y) = v$, and we get, for any integer $r \ge 1$,

$$\mathbf{E}|F(Y) - \mathbf{E}(F(Y))|^{2r} \le K_{2r}(\tfrac{\pi}{2})^{2r}|||\tilde{\Sigma}|||_2^r \mathbf{E}(v'D(Z)D(Z)v)^r.$$

Now, using the special structure of the diagonal matrix $D(Z)$ and the bound (6), we find that

$$\mathbf{E}\left(v'D(Z)D(Z)v\right)^r = \sum_{i_1} \cdots \sum_{i_r} v_{i_1}^2 \ldots v_{i_r}^2 \mathbf{E}\left(Y_{i_1}^2 \ldots Y_{i_r}^2\right)$$

$$\le \exp(2r\tilde{\mu}_* + \tfrac{1}{2}(2r)^2\tilde{\sigma}_*^2)\left(\sum_i v_i^2\right)^r = \exp(2r\tilde{\mu}_* + \tfrac{1}{2}(2r)^2\tilde{\sigma}_*^2).$$

Combining these estimates, we conclude that

$$\mathbf{E}|F(Y) - \mathbf{E}(F(Y))|^{2r} \le K_{2r}(\tfrac{\pi}{2})^{2r}\tilde{\sigma}_*^{2r}\exp(2r\tilde{\mu}_* + \tfrac{1}{2}(2r\tilde{\sigma}_*)^2).$$

Since $v'X - \mathbf{E}(v'X) = v'Y - \mathbf{E}(v'Y)$, it follows that

$$b_L(2r, X) \le K_{2r}(\tfrac{\pi}{2})^{2r}\tilde{\sigma}_*^{2r}\exp(2r\tilde{\mu}_* + \tfrac{1}{2}(2r\tilde{\sigma}_*)^2).$$

In particular, if $\tilde{\mu}_*$ and $\tilde{\sigma}_*^2$ are uniformly bounded, this is of the order O(1).

Quadratic Forms. Consider the quadratic form $F(y) := y'My$, where M is a deterministic *symmetric* matrix with operator norm 1. Then $\nabla F(y) = 2My$, and we get, for any integer $r \ge 1$,

$$\mathbf{E}|F(Y) - \mathbf{E}(F(Y))|^{2r} \le K_{2r}\pi^{2r}|||\tilde{\Sigma}|||_2^r \mathbf{E}(Y'MD(Z)D(Z)MY)^r.$$

14

Observing that $Y = D(Z)1$, where 1 is the vector consisting of 1's, and setting $N := D(Z)MD(Z)$, it follows that

$$\mathbf{E}|F(Y) - \mathbf{E}(F(Y))|^{2r} \leq K_{2r}\pi^{2r}|||\tilde{\Sigma}|||_2^r \mathbf{E}(1'N^{2r}1) .$$

Because most of our bounds depend on $b_{Q_2}(2; X_i)$ only, let us now consider the case $r = 1$. Note that

$$N_{i,j} = M_{i,j}e^{Z_i+Z_j} .$$

So

$$N_{k,l}^2 = \sum_j M_{k,j}M_{j,l}e^{Z_k+2Z_j+Z_l} .$$

Now $Z_k + 2Z_j + Z_l = (e_k + 2e_j + e_l)'Z$, so, by (5),

$$\mathbf{E}\left(\exp^{Z_k+2Z_j+Z_l}\right) = \exp((2e_j + e_k + e_l)'\tilde{\mu})\exp(\tfrac{1}{2}(2e_j + e_k + e_l)'\tilde{\Sigma}(2e_j + e_k + e_l))$$
$$= \exp(2\tilde{\mu}_j + \tilde{\mu}_k + \tilde{\mu}_l)\exp(2\tilde{\Sigma}_{j,j} + \tilde{\Sigma}_{k,k}/2 + \tilde{\Sigma}_{l,l}/2 + 2\tilde{\Sigma}_{j,k} + 2\tilde{\Sigma}_{j,l} + \tilde{\Sigma}_{k,l}) .$$

Therefore,

$$\mathbf{E}\left(N_{k,l}^2\right) = e^{\tilde{\Sigma}_{k,l}}e^{-\tilde{\Sigma}_{k,k}/2}e^{-\tilde{\Sigma}_{l,l}/2}$$
$$\times \sum_j \left(M_{k,j}\exp(\tilde{\mu}_j + \tilde{\mu}_k + \tilde{\Sigma}_{j,j} + \tilde{\Sigma}_{k,k} + 2\tilde{\Sigma}_{j,k})\right)\left(M_{j,l}\exp(\tilde{\mu}_j + \tilde{\mu}_l + \tilde{\Sigma}_{j,j} + \tilde{\Sigma}_{l,l} + 2\tilde{\Sigma}_{j,l})\right) .$$

Let us now write $A \circ B$ for the Hadamard product of two matrices A and B and $e^{\circ A}$ for the Hadamard exponential of a matrix A, i.e. the matrix with entries $e^{A_{i,j}}$. Let us call Δ and $\tilde{\Delta}$ the diagonal matrices with entries $e^{\tilde{\Sigma}_{j,j}}$ and $e^{\tilde{\mu}_j + \tilde{\Sigma}_{j,j}}$, respectively. Note that $M_{k,j}\exp(\tilde{\mu}_j + \tilde{\mu}_k + \tilde{\Sigma}_{j,j} + \tilde{\Sigma}_{k,k} + 2\tilde{\Sigma}_{j,k})$ is the k,j entry of the matrix $\tilde{\Delta}(M \circ e^{\circ 2\tilde{\Sigma}})\tilde{\Delta}$. So

$$\mathbf{E}\left(N^2\right) = \left[\Delta^{-1/2}e^{\circ\tilde{\Sigma}}\Delta^{-1/2}\right] \circ (\tilde{\Delta}(M \circ e^{\circ 2\tilde{\Sigma}})\tilde{\Delta})^2 .$$

Now recall that for any vector x, if D_x is the diagonal matrix with x on its diagonal, (see Horn and Johnson (1994), Lemma 5.1.5),

$$x'(A \circ B)x = \text{trace}\left(D_x A D_x B'\right) .$$

Hence,

$$1'\mathbf{E}\left(N^2\right)1 = \text{trace}\left(\text{Id}_n\left[\Delta^{-1/2}e^{\circ\tilde{\Sigma}}\Delta^{-1/2}\right]\text{Id}_n(\tilde{\Delta}(M \circ e^{\circ 2\tilde{\Sigma}})\tilde{\Delta})^2\right)$$
$$= \text{trace}\left(\left[\Delta^{-1/2}e^{\circ\tilde{\Sigma}}\Delta^{-1/2}\right](\tilde{\Delta}(M \circ e^{\circ 2\tilde{\Sigma}})\tilde{\Delta})^2\right) .$$

Now the Hadamard exponential of a psd matrix is psd (see Horn and Johnson (1994), p. 450). Recall also that for A and B psd matrices, $A \circ B$ is psd (Horn and Johnson (1994), p. 309) and

$$|||A \circ B|||_2 = \lambda_{\max}(A \circ B) \leq \max_i a_{ii}\lambda_{\max}(B) ,$$

by theorem 5.3.4 in Horn and Johnson (1994). Therefore, since M is psd and $|||M|||_2 \leq 1$,

$$|||M \circ e^{\circ 2\tilde{\Sigma}}|||_2 \leq \exp(2\max_j \tilde{\Sigma}_{j,j}) .$$

So

$$|||\Delta(M \circ e^{\circ 2\tilde{\Sigma}})\Delta|||_2 \leq \exp(2\max_j \tilde{\mu}_j + 4\max_j \tilde{\Sigma}_{j,j}) .$$

So we have, using the fact that when A and B are psd, $\text{trace}(AB) \leq \lambda_{max}(B)\text{trace}(A)$, because $A^{1/2}BA^{1/2} \preceq \lambda_{\max}(B)A$,

$$\text{trace}\left(\left[\Delta^{-1/2}e^{\circ\tilde{\Sigma}}\Delta^{-1/2}\right](\Delta(M \circ e^{\circ V} \circ e^{\circ 2\tilde{\Sigma}})\Delta)^2\right) \leq p\exp(4\max_j \tilde{\mu}_j + 8\max_j \tilde{\Sigma}_{j,j}) .$$

Combining the preceding estimates, we conclude that

$$\mathbf{E}|F(Y) - \mathbf{E}(F(Y))|^2 \leq K_2 \pi^2 \tilde{\sigma}_*^2 p \exp(4\tilde{\mu}_* + 8\tilde{\sigma}_*^2) .$$

Now set $v := 2M\mathbf{E}(Y)$ and note that $\|v\|_2^2 \leq 4\mathbf{E}\|Y\|_2^2 \leq 4p \exp(2\tilde{\mu}_* + 2\tilde{\sigma}_*^2)$. Since $X'MX - \mathbf{E}(X'MX) = (Y'MY - \mathbf{E}(Y'MY)) - (v'Y - \mathbf{E}(v'Y))$, it follows that

$$b_{Q_2}(2, X) \leq K_2 \, 4\pi^2 \tilde{\sigma}_*^2 p \exp(4\tilde{\mu}_* + 8\tilde{\sigma}_*^2) .$$

In particular, if $\tilde{\mu}_*$ and $\tilde{\sigma}_*^2$ are uniformly bounded, this is of the order $O(p)$.

3.3 On quadratic forms involving $(X'D^2X/n + A)^{-1}$

3.3.1 On forms of the type $x' \left(X'D^2X/n + A\right)^{-1} x$

Throughout the proofs, we will make heavy use of the following notation: call, consistently with the notations used above,

$$\mathcal{S} = \frac{1}{n} \sum_{i=1}^{n} R_i^2 X_i X_i' \triangleq X'D^2X/n ,$$

where D is a diagonal matrix with positive entries containing the R_i's (on its $d_{i,i}$ entry) and X is the $n \times p$ matrix whose i-th line is X_i'. We will use the notations

$$M \triangleq \mathcal{S} + A , A \succeq t\mathrm{Id}_p ,$$
$$f(X) \triangleq x'M^{-1}x .$$

To alleviate the notation, we do not show explicitly in the notations the dependence of M on A (and therefore, implicitly on t). However, our bounds will involve them, to allow us to show the impact of having a small t (a small regularization), and also to show clearly how $x'A^{-1}x$ affects our bounds. Similarly, because we are mostly interested in the impact of the randomness in X_i's on the form $f(X)$ we keep track only of this random variable.

• **Concentration aspects**

Theorem 3.1. *Suppose $X_1, \ldots, X_n \in \mathbb{R}^p$ are independent. Suppose further that $\mathbf{E}(X_i) = 0$ and, if v is such that $\|v\| = 1$, $\mathbf{E}\left(|X_i'v|^k\right) \leq b_L(k; X_i)$, where $b_L(k; X_i)$ is a deterministic function depending only on the distribution of X_i and k. Call*

$$\mathcal{S} = \frac{1}{n} \sum_{i=1}^{n} R_i^2 X_i X_i' ,$$

where R_i are deterministic.

Call $M = \mathcal{S} + A$, and assume that A is positive definite, with $A \succeq t\mathrm{Id}_p$. We also call $f(X) = x'M^{-1}x$. Then, if $\|x\| = 1$,

$$\mathbf{E}\left(|f(X) - \mathbf{E}(f(X))|^k\right) \leq \frac{c_k}{t^{2k}} \left[\left(\sum_{i=1}^{n} \left[\frac{R_i^4}{n^2} b_L(4; X_i) \wedge t^2 \right] \right)^{k/2} + \left(\sum_{i=1}^{n} \left[\frac{R_i^{2k}}{n^k} b_L(2k; X_i) \wedge t^k \right] \right) \right] .$$

We note that the bound given in the proof below shows the actual dependence of this upper bound on $x'A^{-1}x$. Also, it would be easy to handle the situation where R_i's are random but independent on X_i's.

Proof. We naturally apply Lemma 3.1 to tackle this problem. Let us call $M_i = M - \frac{1}{n}R_i^2 X_i X_i'$.

Using the classic rank-1 update formula,

$$M^{-1} = M_i^{-1} - \frac{R_i^2}{n} \frac{M_i^{-1} X_i X_i' M_i^{-1}}{1 + R_i^2 X_i' M_i^{-1} X_i/n} .$$

Therefore, if $Z = x'M^{-1}x$ and $Z_i = x'M_i^{-1}x$,

$$Z - Z_i = -\frac{R_i^2}{n}\frac{(x'M_i^{-1}X_i)^2}{1 + R_i^2 X_i'M_i^{-1}X_i/n} .$$

Hence,

$$|Z - Z_i| \leq \left[\frac{R_i^2}{n}(x'M_i^{-1}X_i)^2\right] \wedge (x'M_i^{-1}x) ,$$

because M_i is positive definite and $(x'M_i^{-1}X_i)^2 \leq (x'M_i^{-1}x)(X_i'M_i^{-1}X_i)$ by the Cauchy-Schwarz inequality.

Let us call $\mathbf{E}_i$ () expectation with respect to X_i only. Clearly, using our assumption on X_i, we have

$$\mathbf{E}_i\left(|X_i'M_i^{-1}x|^k\right) \leq \|M_i^{-1}x\|^k b_L(k; X_i) .$$

Hence,

$$\mathbf{E}_i\left(|Z - Z_i|^k\right) \leq \left(\frac{R_i^2}{n}\right)^k (x'M_i^{-2}x)^k b_L(2k; X_i) \wedge (x'M_i^{-1}x)^k .$$

Now, $M_i \succeq A \succeq t\mathrm{Id}_p$, so $(x'M_i^{-2}x) \leq t^{-1}x'A^{-1}x$ and $(x'M_i^{-1}x) \leq x'A^{-1}x$, using the fact that $B \longmapsto -B^{-1}$ is operator monotone on Hermitian matrices (Bhatia (1997), p. 114). So we finally have the bounds

$$\mathbf{E}\left(|Z - Z_i|^2|\mathcal{F}_{i-1}\right) \leq \left(\frac{R_i^2}{n}\right)^2 t^{-2}(x'A^{-1}x)^2 b_L(4; X_i) \wedge (x'A^{-1}x)^2 ,$$

$$\mathbf{E}\left(|Z - Z_i|^k\right) \leq \left(\frac{R_i^2}{n}\right)^k t^{-k}(x'A^{-1}x)^k b_L(2k; X_i) \wedge (x'A^{-1}x)^k .$$

Now recalling Equation (2), we have

$$\mathbf{E}\left(|Z - \mathbf{E}(Z)|^k\right) \leq c_k \left\{ \left[\sum_{i=1}^n \left(\frac{R_i^2}{n}\right)^2 \frac{(x'A^{-1}x)^2}{t^2} b_L(4; X_i) \wedge (x'A^{-1}x)^2\right]^{k/2} \right.$$
$$\left. + \sum_{i=1}^n \left[\left(\frac{R_i^2}{n}\right)^k \frac{(x'A^{-1}x)^k}{t^k} b_L(2k; X_i) \wedge (x'A^{-1}x)^k\right] \right\} .$$

Using the fact that $A \succeq t\mathrm{Id}_p$ and $\|x\| = 1$, we have $x'A^{-1}x \leq t^{-1}$, and this gives the result announced in the theorem. $\qquad\square$

● **Lindeberg approach and why the limit does not depend on the distribution of X_i** We are now interested in showing that for a broad class of distribution for X_i, the limit of

$$x'(X'D^2X/n + A)^{-1}x$$

or more precisely

$$\mathbf{E}\left(x'(X'D^2X/n + A)^{-1}x\right)$$

does not depend on the distribution of X_i. We have already seen that we can control the fluctuation of $x'(X'D^2X/n + A)^{-1}x$ around its mean for a broad class of distributions, so all we need to show is that they all have the same means.

We have the following theorem.

Theorem 3.2. *Suppose X_i are i.i.d and Y_i are i.i.d and follow the assumptions mentioned above (at the beginning of Subsection 3.2). Assume that D is a deterministic diagonal matrix, whose diagonal entries are positive and denoted by R_j. We assume that A is a positive definite matrix with $A \succeq t\mathrm{Id}_p$, for some $t > 0$.*

Then, for any given vector x, if $f(X) = x'(X'D^2X/n + A)^{-1}x$,

$$|\mathbf{E}\left(f(X) - f(Y)\right)| \leq \sum_{j=1}^{n} U_j(X_j) + U_j(Y_j) \quad where$$

$$U_j(X_j) \leq \frac{R_j^4}{n^{3/2}} \left(\frac{x'A^{-1}x}{t^2} \sqrt{b_L(4; X_j)} \sqrt{b_{Q_2}(2; X_j)/n} \right) \wedge \frac{R_j^2}{n} \frac{x'A^{-1}x}{t} b_L(2; X_j) . \qquad (7)$$

Let us discuss briefly this result. We see that assuming $\max_j |||\Sigma_j|||_2$ is bounded, and making assumptions on b_L and b_{Q_2} that match the Gaussian situation (i.e b_L and b_{Q_2}/n uniformly bounded in n), the upper bound on the error is of the form (up to constants)

$$\sum_{i=1}^{n} \frac{R_i^4}{n^{3/2}} \wedge \frac{R_i^2}{n} .$$

If the R_i's are given by square-integrable i.i.d. random variables (the same for each n), we have

$$\mathbf{E}\left(\frac{R_i^4}{n^{3/2}} \wedge \frac{R_i^2}{n} \right) = o(n^{-1}) .$$

Hence, when this is the case, and the assumptions of our discussion are met, we have

$$\mathbf{E}\left(f(X) - f(Y)\right) \to 0 ,$$

where $\mathbf{E}\left(\cdot\right)$ is here expectations with respect to all sources of random variables (i.e R_i's, X_i's and Y_i's.) Simple computations also show that if R_i's are random and have $2 + \epsilon$ moments, with $\epsilon \leq 2$,

$$\mathbf{E}\left(\frac{R_i^4}{n^{3/2}} \wedge \frac{R_i^2}{n} \right) \leq \frac{K}{n^{1+\epsilon/4}} .$$

Hence, when this is the case, we have

$$\mathbf{E}\left(f(X) - f(Y)\right) \to 0$$

provided that b_L and b_{Q_2} do not grow too fast to infinity. If we are in a situation where $Y_j = \Sigma_j^{1/2} Y_0$ where Y_0 is such that $b_L(k; Y_0) = \mathrm{O}(1)$ and $b_{Q_2}(k; Y_0) = \mathrm{O}(1)$, the theorem can handle the case where $|||\Sigma_j|||_2 \ll n^{\epsilon/8}$ (which allows $|||\Sigma_j|||_2$ go to infinity). Note that because we are interested in covariance matrices, we will always require R_i to have at least 2 moments and so this theorem essentially covers all the cases of interests to us.

The meaning of the theorem is therefore that under these assumptions, i.e when the upper bound goes to 0 for Y_j and say X_j are gaussians, all we have to do is simply to understand $\mathbf{E}\left(f(X)\right)$ when X is Gaussian. For this task, we can use many of the nice and well-known properties of the Gaussian distribution (which include strong concentration properties).

Proof. It is clear that $\mathbf{E}\left(f(X)\right)$ exists since A is positive definite. We employ the Lindeberg approach (Lindeberg (1922), and e.g. Stroock (1993)) to show that the limit does not depend on the distribution of X_i (note that this technique has been used in other random matrix theoretic questions, e.g. Chatterjee (2005), though the results of this paper do not seem directly applicable; note also that here all our expansions are exact whereas often in the Lindeberg method Taylor approximation arguments are used. That is why we choose to present such an approach.). Let us call

$$Z_j = (Y_1, Y_2, \ldots, Y_{j-1}, X_j, \ldots, X_n) ,$$

with the convention that $Z_1 = (X_1, \ldots, X_n)$ and $Z_{n+1} = (Y_1, \ldots, Y_n)$. Clearly,

$$\mathbf{E}\left(f(X) - f(Y)\right) = \sum_{j=1}^{n} \mathbf{E}\left(f(Z_j) - f(Z_{j+1})\right) .$$

Now let us call $M_j = A + Z_j' D^2 Z_j / n - R_j^2 X_j X_j' / n$. Note that

$$f(Z_j) = x'(M_j + R_j^2 X_j X_j')^{-1} x \ , \ f(Z_{j+1}) = x'(M_j + R_j^2 Y_j Y_j')^{-1} x \ ,$$

and M_j is independent of both X_j and Y_j. Therefore, using the fact that $(M + uu')^{-1} = M^{-1} - M^{-1} u u' M^{-1} / (1 + u' M^{-1} u)$ (see Horn and Johnson (1990), Chapter 0), we have

$$f(Z_j) - f(Z_{j+1}) = \frac{R_j^2}{n} \left[\frac{(x' M_j^{-1} Y_j)^2}{1 + \frac{R_j^2}{n} Y_j' M_j^{-1} Y_j} - \frac{(x' M_j^{-1} X_j)^2}{1 + \frac{R_j^2}{n} X_j' M_j^{-1} X_j} \right] \ .$$

Since Y_j and X_j have the same covariance matrix, Σ_j, if we call $d_j = \text{trace}\left(M_j^{-1} \Sigma_j \right)$, and $q_j(Y_j) = Y_j' M_j^{-1} Y_j$, we see that

$$\frac{1}{1 + R_j^2 q_j(Y_j)/n} = \frac{1}{1 + R_j^2 d_j / n} + \frac{1}{n} R_j^2 \delta_j(Y_j) \ , \tag{8}$$

where

$$\delta_j(Y_j) := \frac{(d_j - q_j(Y_j))}{(1 + R_j^2 q_j(Y_j)/n)(1 + R_j^2 d_j / n)} \ . \tag{9}$$

Hence, we see that

$$\frac{(x' M_j^{-1} Y_j)^2}{1 + \frac{R_j^2}{n} q_j(Y_j)} = \frac{(x' M_j^{-1} Y_j)^2}{1 + \frac{R_j^2}{n} d_j} + \frac{1}{n} R_j^2 (x' M_j^{-1} Y_j)^2 \delta_j(Y_j) \ .$$

Therefore,

$$\begin{aligned} f(Z_j) - f(Z_{j+1}) &= \frac{R_j^2}{n} \left[\frac{(x' M_j^{-1} Y_j)^2}{1 + \frac{R_j^2}{n} d_j} - \frac{(x' M_j^{-1} X_j)^2}{1 + \frac{R_j^2}{n} d_j} \right] \\ &+ \frac{R_j^4}{n^2} \left[(x' M_j^{-1} Y_j)^2 \delta_j(Y_j) - (x' M_j^{-1} X_j)^2 \delta_j(X_j) \right] \\ &= \mathcal{R}_j(1) + \mathcal{R}_j(2) \ . \end{aligned}$$

Interestingly, the first term in the above expansion, $\mathcal{R}_j(1)$ has mean 0, since our assumption of independence (on X_j's and Y_j's) guarantees that M_j is independent of both Y_j and X_j. So we have shown that

$$\mathbf{E}\left(f(X) - f(Y) \right) = \sum_{j=1}^{n} \mathbf{E}\left(f(Z_j) - f(Z_{j+1}) - \mathcal{R}_j(1) \right) = \sum_{j=1}^{n} \mathbf{E}\left(\mathcal{R}_j(2) \right) \ .$$

On the one hand, using the Cauchy-Schwarz inequality, we get

$$\mathbf{E}_j \left((Y_j' M_j^{-1} x)^2 |\delta_j(Y_j)| \right) \le \sqrt{\mathbf{E}_j \left((Y_j' M_j^{-1} x)^4 \right)} \sqrt{\mathbf{E}_j \left(\delta_j(Y_j) \right)^2} \ .$$

By our assumptions (3) and (4), we have

$$\mathbf{E}_j \left((Y_j' M_j^{-1} x)^4 \right) \le (x' M_j^{-2} x)^2 b_L(4; Y_j) \le \left(\frac{x' A^{-1} x}{t} \right)^2 b_L(4; Y_j)$$

and

$$\mathbf{E}_j \left(\delta_j(Y_j) \right)^2 \le b_{Q_2}(2; Y_j) \| |M_j^{-1}| \|_2^2 \le b_{Q_2}(2; Y_j) \frac{1}{t^2} \ , \tag{10}$$

since $M_j^{-1} \preceq A^{-1} \preceq t^{-1}\mathrm{Id}$. Putting everything together, and taking expectations over the other variables, we finally obtain

$$\mathbf{E}\left((Y_j'M_j^{-1}x)^2|\delta_j(Y_j)|\right) \leq \frac{x'A^{-1}x}{t^2}\sqrt{b_L(4;Y_j)}\sqrt{b_{Q_2}(2;Y_j)} . \tag{11}$$

On the other hand, by construction, we have

$$\left|\tfrac{1}{n}R_j^2\delta_j(Y_j)\right| = \left|\frac{1}{1+R_j^2d_j/n} - \frac{1}{1+R_j^2q_j(Y_j)/n}\right| \leq 1 , \tag{12}$$

because both d_j and $q_j(Y_j)$ are non-negative. Thus, we see that

$$\tfrac{1}{n}R_j^2\mathbf{E}_j\left((Y_j'M_j^{-1}x)^2|\delta_j(Y_j)|\right) \leq b_L(2;Y_j)x'M_j^{-2}x \leq b_L(2;Y_j)\frac{x'A^{-1}x}{t}$$

and therefore,

$$\tfrac{1}{n}R_j^2\mathbf{E}\left((Y_j'M_j^{-1}x)^2|\delta_j(Y_j)|\right) \leq b_L(2;Y_j)\frac{x'A^{-1}x}{t} . \tag{13}$$

Naturally, the same bounds hold for $\mathbf{E}\left((X_j'M_j^{-1}x)^2\delta_j(X_j)\right)$. We conclude that

$$|\mathbf{E}\left(f(X) - f(Y)\right)| \leq \sum_{j=1}^n \left[\frac{R_j^4}{n^{3/2}}\left(\frac{x'A^{-1}x}{t^2}\sqrt{b_L(4;Y_j)}\sqrt{b_{Q_2}(2;Y_j)/n}\right) \wedge \frac{R_j^2}{n}\frac{x'A^{-1}x}{t}b_L(2;Y_j)\right]$$
$$+ \sum_{j=1}^n \left[\frac{R_j^4}{n^{3/2}}\left(\frac{x'A^{-1}x}{t^2}\sqrt{b_L(4;X_j)}\sqrt{b_{Q_2}(2;X_j)/n}\right) \wedge \frac{R_j^2}{n}\frac{x'A^{-1}x}{t}b_L(2;X_j)\right] ,$$

as announced in the theorem. $\qquad\square$

3.3.2   On quadratic forms involving $DX(X'D^2X/n + A)^{-1}X'D$

We are now interested in quadratic forms of the type

$$\alpha'\frac{DX}{\sqrt{n}}(X'D^2X/n + A)^{-1}\frac{X'D}{\sqrt{n}}\alpha ,$$

which are very useful when working with both sample means and sample covariance matrices. α here will be a vector with norm bounded away from zero and from infinity in most cases. Hence, we will focus without loss of generality on the case $\|\alpha\| = 1$.

Our strategy is once again to use the Lindeberg method in connection with Efron-Stein type variance bounds.

Before we turn to the technical aspects of the questions, let us make a bit more explicit our motivation. Let us call, if $\mathfrak{X}_i = \mu + R_iX_i$, D a diagonal matrix containing the R_i's, and 1 is an n-dimensional vectors having 1 in all its entries,

$$\widehat{\Sigma} = \frac{1}{n}\mathfrak{X}'\mathfrak{X} - \widehat{\mu}_\mathfrak{X}\widehat{\mu}_\mathfrak{X}' = \frac{1}{n}X'D^2X - \frac{1}{n^2}X'D'11'DX .$$

$\widehat{\Sigma}$ is naturally the covariance matrix of our data (we assume that we observe the $\mathfrak{X}_i$'s). Without loss of generality, we can assume that $\mu = 0$ and do so from now on in this discussion. Let us call $\widehat{\mu} = X'D'1/n$, the mean of the vectors R_iX_i's. Suppose we are interested in

$$\widehat{\mu}_\mathfrak{X}'(\widehat{\Sigma} + A)^{-1}\widehat{\mu}_\mathfrak{X} = (\mu + \widehat{\mu})'(\widehat{\Sigma} + A)^{-1}(\mu + \widehat{\mu}) .$$

These quantities occur naturally in various optimization problems, as well as in theoretical investigations of classification problems. Calling as before

$$M = X'D^2X/n + A , \quad \text{we see that } \widehat{\Sigma} + A = M - \widehat{\mu}\widehat{\mu}' ,$$

and hence, using the rank-1 update formula,

$$\widehat{\mu}'(\widehat{\Sigma} + A)^{-1}\widehat{\mu} = 1 - \frac{1}{1 - \widehat{\mu}'M^{-1}\widehat{\mu}} \ .$$

Spelling out M and $\widehat{\mu}$, we see that

$$\widehat{\mu}'M^{-1}\widehat{\mu} = \alpha'\frac{DX}{\sqrt{n}}(X'D^2X/n + A)^{-1}\frac{X'D}{\sqrt{n}}\alpha \ ,$$

with $\alpha = 1/\sqrt{n}$. Hence our motivation for understanding these problems.
Naturally, we will also be interested in

$$\mu'(\widehat{\Sigma} + A)^{-1}\mu = \mu'M^{-1}\mu - \frac{(\mu'M^{-1}\widehat{\mu})^2}{1 - \widehat{\mu}'M^{-1}\widehat{\mu}} \ .$$

and

$$\mu'(\widehat{\Sigma} + A)^{-1}\widehat{\mu} = \frac{\widehat{\mu}'M^{-1}\mu}{1 - \widehat{\mu}'M^{-1}\widehat{\mu}} \ .$$

- **Lindeberg Approach**

We are now interested in

$$g(\alpha; X) = \alpha'\frac{DX}{\sqrt{n}}(X'D^2X/n + A)^{-1}\frac{X'D}{\sqrt{n}}\alpha \ .$$

The entries of D are assumed to be deterministic and non-negative at this point. It is clear that this can be done without loss of generality, since $(D\alpha)_i = d_{i,i}\alpha_i$ (so negative signs in D could be handled by changing the corresponding signs in α, which would not affect $\|\alpha\|$).

Let us observe that

$$|g(\alpha; X)| \le \|\alpha\|^2 \ . \tag{14}$$

Indeed, setting

$$M \triangleq (X'D^2X/n + A) \succcurlyeq 0 \ , \tag{15}$$

we have, since $M \succeq (X'D^2X/n)$,

$$DXM^{-1}X'D \preceq \mathrm{Id}_n \ ,$$

since Id_n is greater in the Loewner order than any projection matrix.

Theorem 3.3. *Suppose X_i are i.i.d and Y_i are i.i.d and follow the assumptions mentioned above (see Subsection 3.2). Assume that D is a deterministic diagonal matrix, whose diagonal entries are positive and denoted by R_j. We assume that A is a positive definite matrix with $A \succeq t\mathrm{Id}_p$, for some $t > 0$. Let us call, for a deterministic vector α with $\|\alpha\| = 1$ (without loss of generality),*

$$g(\alpha; X) = \alpha'\frac{DX}{\sqrt{n}}(X'D^2X/n + A)^{-1}\frac{X'D}{\sqrt{n}}\alpha \ .$$

Then

$$|\mathbf{E}\left(g(\alpha; X) - g(\alpha; Y)\right)| \le \sum_{i=1}^{n} U(X_i; R_i; \alpha_i) + U(Y_i; R_i; \alpha_i) \ , \tag{16}$$

where $U_i(X_i; R_i; \alpha_i)$ are deterministic quantities depending only on the distribution of X_i. We have, for a numerical constant K that does not depend on the distribution of X_i and Y_i, and not on n or p either,

$$\sum_{i=1}^{n} U(X_i; R_i; \alpha_i) \le K \sum_{i=1}^{n} \left(\frac{R_i^2}{\sqrt{nt}}\sqrt{b_{Q_2}(2; X_i)/n} \wedge 1\right)\left(\alpha_i^2 + \frac{R_i^2}{nt}\sqrt{b_L(4; X_i)}\right) \ .$$

Once again when the R_i's are random (but independent of $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, it is clear that under minimal assumptions on the existence of moments for R_i, the right hand side will converge to 0. Suppose for the moment that $b_{Q_2}(2; X_i)/n$ and $b_L(4; X_i)$ are uniformly bounded and that the R_i are random and uniformly square-integrable. Then we have

$$\mathbf{E}\left(\sum_{i=1}^n \left(\frac{R_i^2}{\sqrt{n}} \wedge 1\right) \alpha_i^2\right) \leq \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i^2 \mathbf{E}\left(R_i^2\right) = \mathrm{O}(n^{-1/2})$$

and

$$\mathbf{E}\left(\sum_{i=1}^n \left(\frac{R_i^2}{\sqrt{n}} \wedge 1\right) \frac{R_i^2}{n}\right) = o(1),$$

so that the upper bound converges to zero in R_i-probability (and also in expectation when the expectation is taken over R_i's, X_i's and Y_i's). Let us now prove this theorem.

Proof. Let

$$M := X'D^2X/n + A \quad \text{and} \quad m := X'D\alpha/\sqrt{n}.$$

Also, let M_i and m_i be the corresponding functionals for $X_{(i)}$, where $X_{(i)} := \sum_{j \neq i} e_j X_j'$. In other words, $X_{(i)}$ is obtained from X by setting the ith row to zero. Clearly, we have

$$M = M_i + \frac{1}{n}R_i^2 X_i X_i' \quad \text{and} \quad m = m_i + \frac{1}{\sqrt{n}}\alpha_i R_i X_i.$$

Note that $X_{(i)}$ is independent of X_i and so are M_i and m_i. After computing the rank-1 perturbation for $(X'D^2X/n + A)^{-1}$, we get that

$$g(\alpha; X) = \frac{1}{n}\alpha'(DX_{(i)} + R_i e_i X_i')\left[M_i^{-1} - \frac{R_i^2}{n}\frac{M_i^{-1}X_i X_i' M_i^{-1}}{1 + R_i^2 \frac{X_i' M_i^{-1} X_i}{n}}\right](X_{(i)}'D + R_i X_i e_i')\alpha.$$

A straightforward calculation shows that, if $g_i(\alpha; X) = \frac{1}{n}\alpha'DX_{(i)}M_i^{-1}X_{(i)}'D\alpha$, we have the key estimate

$$g(\alpha; X) = g_i(\alpha; X) + \alpha_i^2 - \frac{1}{1 + \frac{R_i^2}{n}q_i(X_i)}\left(\alpha_i - \frac{R_i}{\sqrt{n}}\zeta_i\right)^2. \tag{17}$$

where

$$\zeta_i(X_i) = X_i' M_i^{-1} m_i \quad \text{and} \quad q_i(X_i) = X_i' M_i^{-1} X_i.$$

We are now interested in $g(\alpha; X) - g(\alpha; Y)$. Calling $Z_j = (Y_1, Y_2, \ldots, Y_{j-1}, X_j, \ldots, X_n)$, we write as before

$$\mathbf{E}\left(g(\alpha; X) - g(\alpha; Y)\right) = \sum_{j=1}^n \mathbf{E}\left(g(\alpha; Z_j) - g(\alpha; Z_{j+1})\right).$$

It should be noted that the expansion we just got for $g(\alpha; X)$ as a function of X_j also holds if we replace X by Z_j.

With our decomposition (17) above, we immediately see that

$$g(\alpha; Z_i) - g(\alpha; Z_{i+1}) = \frac{1}{1 + \frac{R_i^2}{n}q_i(X_i)}\left(\alpha_i - \frac{R_i}{\sqrt{n}}\zeta_i(X_i)\right)^2 - \frac{1}{1 + \frac{R_i^2}{n}q_i(Y_i)}\left(\alpha_i - \frac{R_i}{\sqrt{n}}\zeta_i(Y_i)\right)^2,$$

where now M_i and m_i are computed from Z_i instead of X. Note that $\mathbf{E}_i\left(\zeta_i(X_i)\right) = 0 = \mathbf{E}_i\left(\zeta_i(Y_i)\right)$ and $\mathbf{E}_i\left(\zeta_i^2(X_i)\right) = \mathbf{E}_i\left(\zeta_i^2(Y_i)\right)$ because the two have the same covariance.

Now let us call

$$\psi_i(X_i) = \left(\alpha_i - \frac{R_i}{\sqrt{n}}\zeta_i(X_i)\right)^2,$$

and let us define $q_i(X_i)$, d_i and $\delta_i(X_i)$ as in the proof of Theorem 3.1. Then, using Equation (8), we have

$$\frac{\psi_i(X_i)}{1 + \frac{R_i^2}{n} q_i(X_i)} = \frac{\psi_i(X_i)}{1 + \frac{R_i^2}{n} d_i} + \frac{R_i^2}{n} \psi_i(X_i)\delta_i(X_i)$$

and therefore

$$g(\alpha; Z_i) - g(\alpha; Z_{i+1}) = \frac{\psi_i(X_i) - \psi_i(Y_i)}{1 + \frac{R_i^2}{n} d_i} + \frac{R_i^2}{n} \left(\psi_i(X_i)\delta_i(X_i) - \psi_i(Y_i)\delta_i(Y_i) \right)$$

So we clearly see that

$$\mathbf{E}_i \left(g(\alpha; Z_i) - g(\alpha; Z_{i+1}) \right) = \frac{R_i^2}{n} \mathbf{E}_i \left(\psi_i(X_i)\delta_i(X_i) - \psi_i(Y_i)\delta_i(Y_i) \right) \ .$$

Recall that we have shown earlier that

$$\mathbf{E}_i \left(\delta_i(X_i)^2 \right) \leq \frac{b_{Q_2}(2; X_i)}{t^2} \text{ and } \frac{R_i^2}{n} |\delta_i(X_i)| \leq 1 \ .$$

Recall also that $\zeta_i = X_i' M_i^{-1} m_i$. It is clear from (14) that

$$\| M_i^{-1} m_i \| \leq \|\alpha\|/\sqrt{t} = 1/\sqrt{t}$$

Hence,

$$\mathbf{E}_i \left(|\zeta_i(X_i)|^k \right) \leq t^{-k/2} b_L(k; X_i) \ .$$

Using Hölder's inequality, we therefore see that

$$\mathbf{E}_i \left(|\psi_i(X_i)\delta_i(X_i)| \right) \leq \frac{K}{t} \sqrt{b_{Q_2}(2; X_i)} \sqrt{\alpha_i^4 + \frac{R_i^4}{n^2} b_L(4; X_i)/t^2} \leq \frac{K}{t} \sqrt{b_{Q_2}(2; X_i)} \left(\alpha_i^2 + \frac{R_i^2}{n} \sqrt{b_L(4; X_i)/t^2} \right) \ .$$

By Equation (12), we also have

$$\frac{1}{n} R_i^2 |\psi_i(X_i)\delta_i(X_i)| \leq |\psi_i(X_i)| \ ,$$

whence

$$\mathbf{E}_i \left(\frac{1}{n} R_i^2 |\psi_i(X_i)\delta_i(X_i)| \right) \leq 2 \left(\alpha_i^2 + \frac{R_i^2}{n} \mathbf{E}_i \left(\zeta_i(X_i) \right)^2 \right) \leq 2 \left(\alpha_i^2 + \frac{R_i^2}{n} b_L(2; X_i)/t \right) \leq 2 \left(\alpha_i^2 + \frac{R_i^2}{n} \sqrt{b_L(4; X_i)/t^2} \right) \ ,$$

since $b_L(2; X_i) \leq \sqrt{b_L(4; X_i)}$ by the Cauchy-Schwarz inequality. Since similar estimates hold for $\psi_i(Y_i)\delta_i(Y_i)$, it finally follows that

$$|\mathbf{E} \left(g(\alpha; X) - g(\alpha; Y) \right)| \leq K \sum_{i=1}^{n} \left[\left(\frac{R_i^2}{n^{1/2} t} \sqrt{b_{Q_2}(2; X_i)/n} \wedge 1 \right) \left(\alpha_i^2 + \frac{R_i^2}{nt} \sqrt{b_L(4; X_i)} \right) \right.$$
$$\left. + \left(\frac{R_i^2}{n^{1/2} t} \sqrt{b_{Q_2}(2; Y_i)/n} \wedge 1 \right) \left(\alpha_i^2 + \frac{R_i^2}{nt} \sqrt{b_L(4; Y_i)} \right) \right] \ .$$

$\square$

• **Efron-Stein aspects** We now turn to the Efron-Stein aspects of the problem, namely we show that our statistic has small variance.

Theorem 3.4. *Suppose X_i are i.i.d and Y_i are i.i.d and follow the assumptions mentioned above. Assume that D is a deterministic diagonal matrix, whose diagonal entries are positive and denoted by R_j. We assume that A is a positive definite matrix with $A \succeq t\mathrm{Id}_p$, for some $t > 0$. Let us call, for a deterministic vector α with $\|\alpha\| = 1$ (without loss of generality),*

$$g(\alpha; X) = \alpha' \frac{DX}{\sqrt{n}} (X' D^2 X/n + A)^{-1} \frac{X'D}{\sqrt{n}} \alpha \ .$$

Then we have, for a certain constant K,

$$\mathrm{var} \left(g(\alpha; X) \right) \leq K \sum_{i=1}^{n} \left[\left(\alpha_i^4 \frac{R_i^4}{n} \frac{b_{Q_2}(2; X_i)}{nt^2} + \frac{R_i^4}{n^2} b_L(4; X_i) \frac{1}{t^2} + \alpha_i^2 \frac{R_i^2}{n} \frac{b_L(2; X_i)}{t} \right) \wedge 1 \right] \ .$$

Before we turn to the proof, let us show that when R_i's are independent and have two moments, the upper bound converges to 0 in (R_i-) probability, when $b_L(4; X_i)$ and $b_{Q_2}(2; X_i)/n$ remain bounded as n grows. Using the Marcienkiewicz-Zygmund strong law of large numbers, we know that

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

Now, suppose for the moment that $b_{Q_2}(2; X_i)/n$ and $b_L(4; X_i)$ are uniformly bounded and that the R_i are random. Since

$$\mathbf{E}\left(\tfrac{1}{n}\alpha_i^4 R_i^4 \wedge 1\right) = \mathbf{E}\left((\tfrac{1}{n}\alpha_i^4 R_i^4 \wedge 1)\mathbf{1}_{\{\alpha_i^4 R_i^4/n \geq 1\}}\right) + \mathbf{E}\left((\tfrac{1}{n}\alpha_i^4 R_i^4 \wedge 1)\mathbf{1}_{\{\alpha_i^4 R_i^4/n < 1\}}\right)$$
$$\leq \mathbf{P}\left(\alpha_i^2 R_i^2/\sqrt{n} \geq 1\right) + \mathbf{E}\left(\tfrac{1}{\sqrt{n}}\alpha_i^2 R_i^2\right) \leq 2\alpha_i^2 \mathbf{E}\left(R_i^2\right)/\sqrt{n}$$

and $\sum_{i=1}^{n} \alpha_i^2 = 1$, we see that

$$\sum \alpha_i^4 \frac{R_i^4}{n} \to 0 \text{ in } R_i - \text{probability} .$$

Proof. A little bit of care is needed to handle the situation where $\|\alpha\|_4^4$ is not small - otherwise the result could be obtained in a slightly easier fashion with slightly coarser bounds. Recall that

$$g(\alpha; X) - g_i(\alpha; X) = \alpha_i^2 - \frac{(\alpha_i - R_i \zeta_i(X_i)/\sqrt{n})^2}{1 + R_i^2 q_i(X_i)/n} ,$$

and therefore

$$g(\alpha; X) - g_i(\alpha; X) = \alpha_i^2 \left(1 - \frac{1}{1 + \frac{R_i^2}{n} q_i}\right) + \frac{1}{1 + \frac{R_i^2}{n} q_i}(R_i^2/n\zeta_i^2 - 2\alpha_i R_i/\sqrt{n}\zeta_i) .$$

Thus, if we set $T := g(\alpha; X)$ and $T_i := g_i(\alpha; X) - \alpha_i^2\left(1 - \frac{1}{1 + \frac{R_i^2}{n} d_i}\right)$, which does not depend on X_i, we have

$$T - T_i = \alpha_i^2 \frac{R_i^2}{n} \frac{\delta_i(X_i)}{(1 + \frac{R_i^2}{n} q_i)(1 + \frac{R_i^2}{n} d_i)} + \frac{1}{1 + \frac{R_i^2}{n} q_i}(R_i^2/n\zeta_i^2 - 2\alpha_i R_i/\sqrt{n}\zeta_i) .$$

So, using the bounds used in the proof of the previous theorem,

$$\mathbf{E}_i\left(|T - T_i|^2\right) \leq K\left(\alpha_i^4 \frac{R_i^4}{n^2} b_{Q_2}(2; X_i)\frac{1}{t^2} + \frac{R_i^4}{n^2} b_L(4; X_i)\frac{1}{t^2} + \alpha_i^2 \frac{R_i^2}{n} \frac{b_L(2; X_i)}{t}\right) .$$

Using the fact that $0 \leq T = g(\alpha; X) \leq 1$ and $0 \leq g_i(\alpha; X) \leq 1$, we also have $|T - T_i| \leq 1 + \alpha_i^2$, and hence

$$\mathbf{E}_i\left(|T - T_i|^2\right) \leq 2\mathbf{E}_i\left(|g(\alpha; X) - g_i(\alpha; X)|^2 + \alpha_i^4\right) \leq 4 .$$

Thus, the Efron-Stein inequality gives us

$$\operatorname{var}(g(\alpha; X)) \leq K \sum_{i=1}^{n}\left(\left(\alpha_i^4 \frac{R_i^4}{n^2} b_{Q_2}(2; X_i)\frac{1}{t^2} + \frac{R_i^4}{n^2} b_L(4; X_i)\frac{1}{t^2} + \alpha_i^2 \frac{R_i^2}{n} \frac{b_L(2; X_i)}{t}\right) \wedge 1\right) .$$

$\square$

• **Gaussian computations** To understand the form we care about, it is now sufficient to compute its mean in a simple case. We naturally turn to the Gaussian case for this final task.

We now compute $\mathbf{E}(g(\alpha; X))$ when the X_i's are independent with (mean 0) normal distribution and possibly different covariance. Let us call

$$P_R = \frac{1}{n} DXM^{-1}X'D' ,$$

with $M = X'D^2X/n + A$. P_R is a $n \times n$ matrix. We have the following result.

Lemma 3.2. *Suppose that X_i are independent normally distributed random variables, with mean 0 and covariance Σ_i. Then $\mathbf{E}(P_R)$ is diagonal and*

$$\mathbf{E}(g(\alpha; X)) = \sum_{i=1}^{n} \alpha_i^2 \mathbf{E}(P_R(i, i)) \; ,$$

where

$$P_R(i, i) = 1 - \frac{1}{1 + \frac{R_i^2}{n} X_i' M_i^{-1} X_i} \; ,$$

and $M_i = \frac{1}{n} \sum_{j \neq i} R_j^2 X_j X_j' + A$.

A particularly interesting case is that where X_j are exchangeable (so for instance, we now allow the covariance Σ_j to be random with a certain prior, and conditional on Σ_j, X_j's are $\mathcal{N}(0, \Sigma_j)$ - the resulting random variables being exchangeable), and so are R_i^2 (which are assumed independent of X_i's). Then we have (if $\mathbf{E}(\cdot)$ is expectation with respect to all sources of randomness) $\mathbf{E}(P_R(i, i)) = \mathbf{E}(P_R(j, j))$, for all (i, j). In this case, we also have

$$\mathbf{E}(\alpha' P_R \beta) = \alpha' \beta \left(1 - \mathbf{E}\left(\frac{1}{1 + \frac{R_1^2}{n} X_1' M_1^{-1} X_1} \right) \right) \; .$$

Therefore, if $\alpha' \beta = 0$, we have

$$\mathbf{E}(\alpha' D X (X' D^2 X / n + A)^{-1} X' D \beta) = 0 \; .$$

Another very interesting case is the situation where R_i's are non-random (or random but independent of X_i's) and X_i's are i.i.d. Then,

$$\mathbf{E}(g(\alpha; X)) = \|\alpha\|_2^2 - \sum_{i=1}^{n} \mathbf{E}\left(\frac{\alpha_i^2}{1 + \frac{R_i^2}{n} X_i' M_i(A)^{-1} X_i} \right) \; .$$

Now, when $\|\Sigma_i\|$ is not too large (i.e $o(p^{1/2 - \eta})$, $\eta > 0$, it is easy to see (by concentration of Gaussian random variables, see Ledoux (2001) and El Karoui (2009a) for details of the application) that $X_i' M_i(A)^{-1} X_i / p$ is concentrated around its mean, which is trace $\left(M_i(A)^{-1} \Sigma_i / p \right)$. When $\Sigma_j = \Sigma$, this quantity has a limit as n and p tend to ∞ with $p/n \to \rho$, and this limit is known (see e.g. Marčenko and Pastur (1967); Silverstein and Bai (1995)). As a matter of fact, then

$$\text{trace}\left(M_i(A)^{-1} \Sigma \right) = \text{trace}\left(X_0' D_i^2 X_0 / n + \Sigma^{-1/2} A \Sigma^{-1/2} \right) \; ,$$

where X_0 are i.i.d $\mathcal{N}(0, \mathrm{Id}_p)$ and $D_i = D - R_i e_i e_i'$. Calling L this limit (which naturally depends on the distribution of R's), we have

$$g(\alpha; X) \simeq 1 - \sum_{i=1}^{n} \frac{\alpha_i^2}{1 + \frac{p}{n} R_i^2 L} \; .$$

Proof. Notice that

$$P_R(i, j) = \frac{1}{n} R_i R_j X_i' (M'M/n + A)^{-1} X_j \; .$$

Now changing X_i into $-X_i$ does not affect the term $M'M/n + A = \frac{1}{n} \sum_{i=1}^{n} R_i^2 X_i X_i' + A$, but changes the sign of $P_R(i, j)$. On the other hand, $\{X_1, \ldots, X_{i-1}, X_i, X_{i+1}, \ldots, X_n\} \overset{\mathcal{L}}{=} \{X_1, \ldots, X_{i-1}, -X_i, X_{i+1}, \ldots, X_n\}$. So we conclude that

$$P_R(i, j) \overset{\mathcal{L}}{=} -P_R(i, j) \text{ when } i \neq j \; .$$

Now it is easy to check that in the positive semi-definite ordering, $P_R^2 \preceq P_R$. So $\||P_R\||_2 \leq 1$. So in particular, all of its entries are less than 1 in absolute value and therefore have moments.

So we have shown that when X_i are independent mean 0 Gaussian variables,

$$\mathbf{E}(P_R(i, j)) = 0 \text{ if } i \neq j \; .$$

And we therefore have the proved the lemma. (The description of the diagonal comes from using rank-1 update formulas.) □

3.3.3 On $n^{-1/2}\alpha'DX(X'D^2X/n + A)^{-1}x$

These forms naturally occur in the study of quadratic forms involving both the sample mean and the sample covariance matrix as we explained at the beginning of Subsubsection 3.3.2, hence our interest in them.

Therefore, for our applications, we also need results about the quantity

$$h(\alpha; X) := n^{-1/2}\alpha'DX(X'D^2X/n + A)^{-1}x \, .$$

where α and x are deterministic vectors, whose norm we will generally assume (without loss of generality) to be 1.

Note that if $M = X'D^2X/n + A$, $|||M^{-1/2}(X'D^2X/n)M^{-1/2}|||_2 \leq 1$, and hence,

$$|h(\alpha; X)| \leq \|\alpha\| \sqrt{x'M^{-1}x} \leq \|\alpha\| \sqrt{x'A^{-1}x} \leq 1/\sqrt{t} \, , \tag{18}$$

which follows from the Cauchy-Schwarz inequality, and (14).

Concentration   Our first aim is to show that $h(\alpha; X)$ is also essentially deterministic.

Theorem 3.5. *Under our usual assumptions (stated in Subsection 3.2), we have*

$$\mathbf{E}\left(h(\alpha; X) - \mathbf{E}\left(h(\alpha; X)\right)\right)^2 \leq K \sum_{j=1}^{n} \left[\left(\tfrac{1}{n}\alpha_i^2 R_i^2 b_L(2, X_i)\frac{x'A^{-1}x}{t} + \tfrac{1}{n^2}R_i^4 b_L(4, X_i)\frac{x'A^{-1}x}{t^2}\right) \wedge \left(x'A^{-1}x\right)\right] \, .$$

Proof. This is an application of the Efron-Stein inequality. Let M and M_i be defined as in the proof of Theorem 3.1, and let $m := n^{-1/2}X'D\alpha$ and $m_i := n^{-1/2}X'_{(i)}D\alpha$, where $X_{(i)}$ is defined as in the proof of Theorem 3.3. Using the rank-1 perturbation formula once more, we get

$$h(\alpha; X) := m'M^{-1}x = (m'_i + n^{-1/2}\alpha_i R_i X'_i)\left[M_i^{-1} - \frac{R_i^2}{n}\frac{M_i^{-1}X_iX'_iM_i^{-1}}{1 + R_i^2\frac{X'_iM_i^{-1}X_i}{n}}\right] x \, .$$

A straightforward calculation shows that, if $h_i(\alpha; X) = m'_iM_i^{-1}x$,

$$h(\alpha; X) = h_i(\alpha; X) + \frac{\varphi_i(X_i)}{1 + \tfrac{1}{n}R_i^2 q_i(X_i)} \, , \tag{19}$$

where

$$\varphi_i(X_i) := \left(n^{-1/2}\alpha_i R_i X'_iM_i^{-1}x - \tfrac{1}{n}R_i^2 X'_iM_i^{-1}m_i X'_iM_i^{-1}x\right) \, . \tag{20}$$

Note that $h_i(\alpha; X)$ is independent of X_i here. Thus, the Efron-Stein inequality yields

$$\mathrm{var}\left(h(\alpha; X)\right) \leq \sum_{i=1}^{n} \mathrm{var}\left(h(\alpha; X) - h_i(\alpha; X)\right) \leq \sum_{i=1}^{n} \mathbf{E}\left(\frac{\varphi_i(X_i)}{1 + \tfrac{1}{n}R_i^2 q_i(X_i)}\right)^2 \, .$$

Now, on the one hand, using (3), we have

$$\mathbf{E}\left(X'_iM_i^{-1}x\right)^2 \leq b_L(2, X_i)\, x'A^{-1}x/t \, ,$$

$$\mathbf{E}\left(X'_iM_i^{-1}m_i X'_iM_i^{-1}x\right)^2 \leq \sqrt{b_L(4, X_i)\,(x'A^{-1}x/t)^2}\sqrt{b_L(4, X_i)/t^2} \, ,$$

and therefore

$$\mathbf{E}\left(\frac{\varphi_i(X_i)}{1 + \tfrac{1}{n}R_i^2 q_i(X_i)}\right)^2 \leq \mathbf{E}\left(\varphi_i(X_i)\right)^2 \leq 2\left(\tfrac{1}{n}\alpha_i^2 R_i^2 b_L(2, X_i)\frac{x'A^{-1}x}{t} + \tfrac{1}{n^2}R_i^4 b_L(4, X_i)\frac{x'A^{-1}x}{t^2}\right) \, . \tag{21}$$

On the other hand, it follows from (19) and (18) that

$$\mathbf{E}\left(\frac{\varphi_i(X_i)}{1 + \tfrac{1}{n}R_i^2 q_i(X_i)}\right)^2 \leq 2\left(\mathbf{E}\left(h(\alpha; X)\right)^2 + \mathbf{E}\left(h_i(\alpha; X)\right)^2\right) \leq 4x'A^{-1}x \, .$$

The proof is completed by combining these estimates. $\square$

- **Lindeberg approach**

Our next aim is to show that the limit of $h(\alpha; X)$ does not depend on the distribution of the X_i.

Theorem 3.6. *Under our usual assumptions (stated in Subsection 3.2), we have*

$$|\mathbf{E}\left(h(\alpha; X) - h(\alpha; Y)\right)| \le K \sum_{j=1}^{n} U_j(X_j) + U_j(Y_j), \quad with$$

$$U_j(X_j) \le K \left[\frac{R_i^2}{nt}\sqrt{b_Q(2; X_i)} \wedge 1\right] \cdot \left[\left(\frac{1}{\sqrt{n}}|\alpha_i|R_i\sqrt{b_L(2, X_i)}\sqrt{\frac{x'A^{-1}x}{t}} + \frac{1}{n}R_i^2\sqrt{b_L(4, X_i)}\sqrt{\frac{x'A^{-1}x}{t^2}}\right)\right].$$

Proof. We use the notation from the proof of Theorem 3.2. Using the decomposition (19) with X replaced by Z_j, Z_{j+1} and observing that $h_j(\alpha; Z_j) = h_j(\alpha; Z_{j+1})$, we get

$$\mathbf{E}\left(h(\alpha; X) - h(\alpha; Y)\right) = \sum_{j=1}^{n} \mathbf{E}\left(h(\alpha; Z_j) - h(\alpha; Z_{j+1})\right)$$

$$= \sum_{i=1}^{n} \mathbf{E}\left(\frac{\varphi_i(X_i)}{1 + \frac{1}{n}R_i^2 q_i(X_i)} - \frac{\varphi_i(Y_i)}{1 + \frac{1}{n}R_i^2 q_i(Y_i)}\right),$$

where $\varphi_i(X_i)$ is defined as in (20), but with X replaced by Z_i. Next, using (8), we have

$$\frac{\varphi_i(X_i)}{1 + \frac{1}{n}R_i^2 q_i(X_i)} - \frac{\varphi_i(Y_i)}{1 + \frac{1}{n}R_i^2 q_i(Y_i)} = \left(\frac{\varphi_i(X_i)}{1 + \frac{1}{n}R_i^2 d_i} - \frac{\varphi_i(Y_i)}{1 + \frac{1}{n}R_i^2 d_i}\right) + \frac{R_i^2}{n}\left(\varphi_i(X_i)\delta_i(X_i) - \varphi_i(Y_i)\delta_i(Y_i)\right).$$

Since X_i and Y_i both have mean 0 and covariance Σ_i, it follows that

$$\mathbf{E}_i\left(\frac{\varphi_i(X_i)}{1 + \frac{1}{n}R_i^2 q_i(X_i)} - \frac{\varphi_i(Y_i)}{1 + \frac{1}{n}R_i^2 q_i(Y_i)}\right) = \mathbf{E}_i\left(\frac{R_i^2}{n}\left(\varphi_i(X_i)\delta_i(X_i) - \varphi_i(Y_i)\delta_i(Y_i)\right)\right).$$

Now, on the one hand, using Cauchy-Schwarz inequality as well as (10) and (21), we have

$$\mathbf{E}_i\left(|\varphi_i(X_i)\delta_i(X_i)|\right) \le \left(\mathbf{E}_i\left(\delta_i(X_i)\right)^2 \mathbf{E}_i\left(\varphi_i(X_i)\right)^2\right)^{1/2}$$

$$\le \frac{K}{t}\sqrt{b_Q(2; X_i)}\left(\frac{1}{n}\alpha_i^2 R_i^2 b_L(2, X_i)\frac{x'A^{-1}x}{t} + \frac{1}{n^2}R_i^4 b_L(4, X_i)\frac{x'A^{-1}x}{t^2}\right)^{1/2}$$

$$\le \frac{K}{t}\sqrt{b_Q(2; X_i)}\left(\frac{1}{\sqrt{n}}|\alpha_i|R_i\sqrt{b_L(2, X_i)}\sqrt{\frac{x'A^{-1}x}{t}} + \frac{1}{n}R_i^2\sqrt{b_L(4, X_i)}\sqrt{\frac{x'A^{-1}x}{t^2}}\right).$$

On the other hand, using (12), we get

$$\frac{1}{n}R_i^2\mathbf{E}_i\left(|\varphi_i(X_i)\delta_i(X_i)|\right) \le \mathbf{E}_i\left(|\varphi_i(X_i)|\right) \le \frac{1}{\sqrt{n}}|\alpha_i|R_i b_L(1, X_i)\sqrt{\frac{x'A^{-1}x}{t}} + \frac{1}{n}R_i^2 b_L(2, X_i)\sqrt{\frac{x'A^{-1}x}{t^2}}.$$

We now use that $b_L(k, X_i) \le \sqrt{b_L(2k, X_i)}$. Combining these estimates, we get

$$\frac{1}{n}R_i^2\mathbf{E}\left(|\varphi_i(X_i)\delta_i(X_i)|\right) \le K\left(\frac{R_i^2}{nt}\sqrt{b_Q(2; X_i)} \wedge 1\right)\left(\frac{1}{\sqrt{n}}|\alpha_i|R_i\sqrt{b_L(2, X_i)}\sqrt{\frac{x'A^{-1}x}{t}} + \frac{1}{n}R_i^2\sqrt{b_L(4, X_i)}\sqrt{\frac{x'A^{-1}x}{t^2}}\right).$$

Since similar estimates hold for $\varphi_i(Y_i)\delta_i(Y_i)$, this completes the proof. $\qquad\square$

- **Gaussian computations**

Consider the case where the X_i are independent normal random vectors with mean zero and covariance Σ_i. Then we clearly have $X \overset{\mathcal{L}}{=} -X$ and therefore

$$h(\alpha, X) \overset{\mathcal{L}}{=} h(\alpha, -X) = -h(\alpha, X).$$

But this means that we must have $\mathbf{E}\left(h(\alpha, X)\right) = 0$. (Recall that $|h(\alpha; X)| \le 1/\sqrt{t}$ by Equation (18), so the existence of the equation is not a problem.)

3.4 Forms in $M^{-1}\Sigma_\epsilon M^{-1}$, $\Sigma_\epsilon \succeq 0$

In a variety of situations, we will need to work with quantities of the type

$$x'M^{-1}\Sigma_\epsilon M^{-1}x .$$

These quantities will occur when we study $\mathrm{var}\left(x'M^{-1}\epsilon\right)$ where ϵ has mean 0 and covariance Σ_ϵ. So these quantities will appear when we investigate the risk of various estimators (or asset allocations). This is why we restrict ourselves to $\Sigma_\epsilon \succeq 0$, though our proofs would go through with minor adjustments if Σ_ϵ was allowed to be more general.

We will also need to understand

$$\widehat{\mu}'M^{-1}\Sigma_\epsilon M^{-1}\widehat{\mu} ,$$

if we want to understand the risk properties of certain portfolio allocations.

Hence our problem is the following: in all the forms where before M^{-1} was involved, we now want to work with $M^{-1}\Sigma_\epsilon M^{-1}$ instead. Our idea - somewhat similar to the one developed in El Karoui (2009c) - is the following: consider

$$M_u = X'D^2X/n + A + u\Sigma_\epsilon , \text{ and } M_0 = M .$$

We remark that

$$\left.\frac{\partial}{\partial u}M_u^{-1}\right|_{u=0} = -M^{-1}\Sigma_\epsilon M^{-1} .$$

Hence, at least formally, our previous proofs will go through; the only thing we have to do is replace A by $A + u\Sigma_\epsilon$ and take a derivative with respect to u so we can get the decompositions that will help us make our methods work.

3.4.1 Forms in $x'M^{-1}\Sigma_\epsilon M^{-1}x$

It is natural to study these forms in a variety of contexts, for instance when $x = \mu$. We have the following theorem, which holds under what we now call our "usual assumptions", namely $A \succeq t\mathrm{Id}$, X_i are i.i.d with mean 0 and covariance Σ_i, and so are Y_i's, though X_i and Y_i have different distributions.

Theorem 3.7. *Let $M = \frac{1}{n}\sum_{i=1}^n R_i^2 X_i X_i' + A$ and*

$$F(X) = x'M^{-1}\Sigma_\epsilon M^{-1}x .$$

Let us call

$$b(A,\Sigma_\epsilon) = |||A^{-1/2}\Sigma_\epsilon A^{-1/2}|||_2 .$$

Then

$$\mathrm{var}\left(F(X)\right) \leq K(x'A^{-1}x)^2 b(A,\Sigma_\epsilon)^2 \sum_{i=1}^n \left[\left(\frac{R_i^4}{n^2}\frac{1}{t^2}b_L(4;X_i) \wedge 1\right)\right] .$$

Also,

$$\left|\mathbf{E}\left(F(X) - F(Y)\right)\right| \leq \sum_{i=1}^n U_i(X_i) + U_i(Y_i)$$

$$U_i(X_i) \leq K\, b(A;\Sigma_\epsilon)\frac{x'A^{-1}x}{t}\left\{\left[\frac{R_i^4}{n^{3/2}}\sqrt{\frac{b_{Q_2}(2;X_i)}{n}}\sqrt{b_L(4;X_i)}\frac{1}{t}\right] \wedge \frac{R_i^2}{n}b_L(2;X_i)\right\}$$

As is explained in the proof of the theorem, when R_i's are i.i.d and uniformly square integrable, the upper bound goes to zero, provided $\sqrt{\frac{b_{Q_2}(2;X_i)}{n}}$ and $b_L(4;X_i)$ remain bounded.

It should be noted that when X_i are i.i.d with covariance Σ, we have found in Heuristic 2.2 and its proof a deterministic equivalent for $F(X)$. Naturally, our theorem shows that doing computations in the Gaussian case is enough to understand $\mathbf{E}\left(F(Y)\right)$ for Y with a variety of distributions - that is the essence of Lindeberg-style results.

Proof. Let us call
$$f(X) = f(X; A) = x'M^{-1}x = x'(X'D^2X/n + A)^{-1}x .$$

Call $f_i(X)$ the same quantity where $D_{i,i} = R_i$ is replaced by $D_{i,i} = 0$ (or equivalently X_i is replaced by 0).

Our key estimate was
$$f(X) - f_i(X) = -\frac{R_i^2}{n}\frac{(x'M_i^{-1}X_i)^2}{1 + \frac{R_i^2}{n}X_i'M_i^{-1}X_i} .$$

This equality is true if A is replaced by $A + u\Sigma_\epsilon$. Now we can take the derivative of this expression with respect to u. Call
$$F(X) = x'M^{-1}\Sigma_\epsilon M^{-1}x ,$$

and F_i the same quantity when X_i is replaced by 0. We have

$$F(X) - F_i(X) = -\left.\frac{\partial}{\partial u}\right|_{u=0}[f(X; A + u\Sigma_\epsilon) - f_i(X; A + u\Sigma_\epsilon)] .$$

Recall the notations $q_i(X_i) = X_i'M_i^{-1}X_i$, $d_i = \mathrm{trace}\left(M_i^{-1}\Sigma_i\right)$. After taking the derivative, we get

$$-(F(X) - F_i(X)) = 2\frac{R_i^2}{n}\frac{x'M_i^{-1}X_i\,x'M_i^{-1}\Sigma_\epsilon M_i^{-1}X_i}{1 + \frac{R_i^2}{n}q_i} - \frac{R_i^4}{n^2}\frac{(x'M_i^{-1}X_i)^2\,X_i'M_i^{-1}\Sigma_\epsilon M_i^{-1}X_i}{(1 + \frac{R_i^2}{n}q_i)^2} .$$

Let us call

$$\tilde{q}_i(X_i) = X_i'M_i^{-1}\Sigma_\epsilon M_i^{-1}X_i ,$$
$$\tilde{d}_i(X_i) = \mathrm{trace}\left(\Sigma_i M_i^{-1}\Sigma_\epsilon M_i^{-1}\right) ,$$
$$E_1 = 2\frac{R_i^2}{n}\frac{x'M_i^{-1}X_i\,x'M_i^{-1}\Sigma_\epsilon M_i^{-1}X_i}{1 + \frac{R_i^2}{n}q_i} ,$$
$$E_2 = \frac{R_i^4}{n^2}\frac{(x'M_i^{-1}X_i)^2\,\tilde{q}_i(X_i)}{(1 + \frac{R_i^2}{n}q_i)^2}$$

- **Control of E_1** By using the fact that $|v'M_i^{-1}X_i| \le \sqrt{v'M_i^{-1}v}\sqrt{X_i'M_i^{-1}X_i}$, we see that

$$|E_1| \le 2\sqrt{x'M_i^{-1}x}\sqrt{x'M_i^{-1/2}(M_i^{-1/2}\Sigma_\epsilon M_i^{-1/2})^2 M_i^{-1/2}x}$$

On $\mathbf{M_i^{-1}\Sigma_\epsilon M_i^{-1}}$ We first note that $M_i^{-1/2}M_i^{-1/2}\Sigma_\epsilon M_i^{-1/2}M_i^{-1/2} \preceq M_i^{-1}|||M_i^{-1/2}\Sigma_\epsilon M_i^{-1/2}|||_2$. Now $|||M_i^{-1/2}\Sigma_\epsilon M_i^{-1/2}|||_2 = \lambda_{\max}(M_i^{-1/2}\Sigma_\epsilon M_i^{-1/2}) = \lambda_{\max}(\Sigma_\epsilon^{1/2}M_i^{-1}\Sigma_\epsilon^{1/2})$ by e.g. similarity. Now $\Sigma_\epsilon^{1/2}M_i^{-1}\Sigma_\epsilon^{1/2} \preceq \Sigma_\epsilon^{1/2}A^{-1}\Sigma_\epsilon^{1/2}$, so

$$|||M_i^{-1/2}\Sigma_\epsilon M_i^{-1/2}|||_2 \le \lambda_{\max}(\Sigma_\epsilon^{1/2}A^{-1}\Sigma_\epsilon^{1/2}) = |||\Sigma_\epsilon^{1/2}A^{-1}\Sigma_\epsilon^{1/2}|||_2 = b(A; \Sigma_\epsilon) .$$

We therefore also have
$$|||M_i^{-1}\Sigma_\epsilon M_i^{-1}|||_2 \le \frac{b(A; \Sigma_\epsilon)}{t} .$$

We will also repeatedly need to control $\|M_i^{-1}\Sigma_\epsilon M_i^{-1}v\|$ for a fixed vector v. Call $u = M_i^{-1}\Sigma_\epsilon M_i^{-1}v$. Clearly
$$u'u = v'M_i^{-1/2}(M_i^{-1/2}\Sigma_\epsilon M_i^{-1/2})M_i^{-1}(M_i^{-1/2}\Sigma_\epsilon M_i^{-1/2})M_i^{-1/2}v .$$

Now, using our bounds on $|||(M_i^{-1/2}\Sigma_\epsilon M_i^{-1/2})|||_2$, $|||M_i^{-1}|||_2 \le t$ and the fact that $||| \cdot |||_2$ is submultiplicative, we have
$$|||(M_i^{-1/2}\Sigma_\epsilon M_i^{-1/2})M_i^{-1}(M_i^{-1/2}\Sigma_\epsilon M_i^{-1/2})|||_2 \le \frac{b^2(A; \Sigma_\epsilon)}{t} .$$

So
$$\|M_i^{-1}\Sigma_\epsilon M_i^{-1}v\|^2 = u'u \le \frac{b^2(A;\Sigma_\epsilon)}{t}v'M_i^{-1}v \le \frac{b^2(A;\Sigma_\epsilon)}{t}v'A^{-1}v$$

Finally, we conclude that
$$\|M_i^{-1}\Sigma_\epsilon M_i^{-1}v\| \le \frac{b(A;\Sigma_\epsilon)}{\sqrt{t}}\|M_i^{-1}v\| \le \frac{b(A;\Sigma_\epsilon)}{\sqrt{t}}\sqrt{v'A^{-1}v} .$$

Using the previous bounds, we clearly then have
$$|E_1| \le 2x'A^{-1}x \; b(A,\Sigma_\epsilon) .$$

On the other hand, using the Cauchy-Schwarz inequality, we see that
$$\mathbf{E}_i\left(|E_1|^k\right) \le K\frac{R_i^{2k}}{n^k}(x'A^{-1}x)^k\frac{b(A,\Sigma_\epsilon)^k}{t^k}b_L(2k;X_i) .$$

Hence,
$$\mathbf{E}\left(|E_1|^k\right) \le K(x'A^{-1}x)^kb(A,\Sigma_\epsilon)^k\left[\frac{R_i^{2k}}{n^k}\frac{1}{t^k}b_L(2k;X_i)\wedge 1\right]$$

- **Control of** E_2 Writing
$$-E_2 = \frac{R_i^2}{n}\frac{(x'M_i^{-1}X_i)^2}{(1+\frac{R_i^2}{n}q_i)}\frac{R_i^2}{n}\frac{\tilde{q}_i(X_i)}{(1+\frac{R_i^2}{n}q_i)}$$

we remark that
$$0 \le \frac{R_i^2}{n}\frac{\tilde{q}_i(X_i)}{(1+\frac{R_i^2}{n}q_i)} \le b(A;\Sigma_\epsilon) \text{ and } |E_2| \le \frac{R_i^2}{n}\frac{(x'M_i^{-1}X_i)^2}{(1+\frac{R_i^2}{n}q_i)}b(A;\Sigma_\epsilon) .$$

Hence, we can conclude that
$$|E_2| \le x'A^{-1}x \; b(A,\Sigma_\epsilon) .$$

We also have the inequalities
$$\mathbf{E}_i\left(|E_2|^k\right) \le \frac{R_i^{2k}}{n^k}b_L(2k;X_i)(x'M_i^{-2}x)^kb(A;\Sigma_\epsilon)^k \le \frac{R_i^{2k}}{n^k}b_L(2k;X_i)(x'A^{-1}x)^k\frac{b(A;\Sigma_\epsilon)^k}{t^k} .$$

Therefore,
$$\mathbf{E}\left(|E_2|^k\right) \le K(x'A^{-1}x)^k \; b(A,\Sigma_\epsilon)^k\left(\frac{R_i^{2k}}{t^kn^k}b_L(2k;X_i)\wedge 1\right) .$$

- **Efron-Stein aspects**

Using the Efron-Stein inequality, we have
$$\text{var}\left(F(X)\right) \le \sum_{i=1}^n\text{var}\left(F(X)-F_i(X)\right) \le K(x'A^{-1}x)^2b(A,\Sigma_\epsilon)^2\sum_{i=1}^n\left[\left(\frac{R_i^4}{n^2}\frac{1}{t^2}b_L(4;X_i)\wedge 1\right)\right] .$$

Hence, when R_i's have 2 moments, var $(F(X)) \to 0$ in R_i-probability.

- **Lindeberg aspects**

We go a bit fast here. M_i is now computed from data $X_1,\ldots,X_{i-1},0,Y_{i+1},\ldots,Y_n$. Recall that
$$E_1(X_i) = 2\frac{R_i^2}{n}\frac{x'M_i^{-1}X_i \; x'M_i^{-1}\Sigma_\epsilon M_i^{-1}X_i}{1+\frac{R_i^2}{n}q_i(X_i)}$$

Let us show that, when Y_i and X_i have the same covariance Σ_i and mean 0, we can control
$$\sum_{i=1}^n\mathbf{E}\left(E_1(X_i)-E_1(Y_i)\right) .$$

We call $N_i = x' M_i^{-1} X_i \, x' M_i^{-1} \Sigma_\epsilon M_i^{-1} X_i$ and (as before) $d_i = \text{trace} \left(\Sigma_i M_i^{-1} \right)$. We have

$$E_1(X_i) = 2 \frac{R_i^2}{n} \frac{N_i(X_i)}{1 + \frac{R_i^2}{n} d_i} + 2 \frac{R_i^2}{n} N_i(X_i) \frac{R_i^2/n(d_i - q_i(X_i))}{(1 + \frac{R_i^2}{n} d_i)(1 + \frac{R_i^2}{n} q_i(X_i))} .$$

Note that $\mathbf{E}_i \left(N_i(X_i) \right) = \mathbf{E}_i \left(N_i(Y_i) \right)$, so to control $\mathbf{E} \left(E_1(X_i) - E_1(Y_i) \right)$, we just need to understand the second term, namely

$$\mathcal{R}_1(X_i) = 2 \frac{R_i^2}{n} N_i(X_i) \frac{R_i^2/n(d_i - q_i(X_i))}{(1 + \frac{R_i^2}{n} d_i)(1 + \frac{R_i^2}{n} q_i(X_i))} .$$

Our studies in Subsubsection 3.3 show that

$$\delta_i(X_i) = \frac{(d_i - q_i(X_i))}{(1 + \frac{R_i^2}{n} d_i)(1 + \frac{R_i^2}{n} q_i(X_i))}$$

is such that

$$|R_i^2/n \delta_i(X_i)| \leq 1 .$$

On the other hand, we have essentially given bounds earlier for $\mathbf{E}_i \left(|N_i|^k \right)$ (see the work on $\mathbf{E}_i \left(|E_i|^k \right)$), so we have

$$\mathbf{E}_i \left(|\mathcal{R}_1(X_i)| \right) \leq K \frac{R_i^2}{n} b_L(2; X_i) \frac{x' A^{-1} x}{t} b(A; \Sigma_\epsilon) .$$

Furthermore,

$$|\mathcal{R}_1(X_i)| \leq K \frac{R_i^4}{n^2} |N_i| |d_i - q_i(X_i)| .$$

So

$$\mathbf{E}_i \left(|\mathcal{R}_1(X_i)| \right) \leq K \frac{R_i^4}{n^2} \sqrt{\mathbf{E}_i \left(|d_i - q_i(X_i)|^2 \right)} \sqrt{\mathbf{E}_i \left(N_i^2 \right)} .$$

Using our bounds on $\mathbf{E}_i \left(N_i^k \right)$ and those on $\mathbf{E}_i \left(|d_i - q_i(X_i)|^2 \right)$, we get

$$\mathbf{E}_i \left(|\mathcal{R}_1(X_i)| \right) \leq K \frac{R_i^4}{n^2} \frac{x' A^{-1} x}{t} b(A; \Sigma_\epsilon) \frac{\sqrt{b_{Q_2}(2; X_i)}}{t} \sqrt{b_L(4; X_i)} .$$

We conclude that

$$|\mathbf{E} \left(\mathcal{R}_1(X_i) \right)| \leq K \frac{x' A^{-1} x}{t} b(A; \Sigma_\epsilon) \left[\frac{R_i^4}{n^{3/2}} \sqrt{\frac{b_{Q_2}(2; X_i)}{n}} \frac{\sqrt{b_L(4; X_i)}}{t} \wedge \frac{R_i^2}{n} b_L(2; X_i) \right] ,$$

and similarly for Y_i. We have shown that

$$|\mathbf{E} \left(E_1(X_i) - E_1(Y_i) \right)| \leq K \frac{x' A^{-1} x}{t} b(A; \Sigma_\epsilon) \left[\frac{R_i^4}{n^{3/2}} \sqrt{\frac{b_{Q_2}(2; X_i)}{n}} \frac{\sqrt{b_L(4; X_i)}}{t} \wedge \frac{R_i^2}{n} b_L(2; X_i) \right] ,$$

and we can therefore control

$$\left| \sum_{i=1}^n \mathbf{E} \left(E_1(X_i) - E_1(Y_i) \right) \right| .$$

• **About E_2**

We now turn to the E_2 part of the problem. The strategy is to replace

$$\tilde{q}_i(X_i) = X_i' M_i^{-1} \Sigma_\epsilon M_i^{-1} X_i \text{ by the "equivalent" (and independent of } X_i)$$
$$\tilde{d}_i = \text{trace} \left(\Sigma_i M_i^{-1} \Sigma_\epsilon M_i^{-1} \right) ,$$

and similarly to replace $q_i(X_i) = X_i' M_i^{-1} X_i$ by $d_i(X_i) = \text{trace} \left(\Sigma_i M_i^{-1} \right)$. Hence the first term is going to have the same mean for both X_i and Y_i and we just have to work on the remainders. Let us call

$$\Delta_i(X_i) = \frac{1}{(1 + \frac{R_i^2}{n} q_i(X_i))^2} - \frac{1}{(1 + \frac{R_i^2}{n} d_i(X_i))^2}$$

31

and let us remark that, with the $\delta_i(X_i)$ notation we just recalled, we have

$$\Delta_i(X_i) = \frac{R_i^2}{n}\delta_i(X_i)\left[\frac{1}{1+\frac{R_i^2}{n}q_i(X_i)} + \frac{1}{1+\frac{R_i^2}{n}d_i(X_i)}\right] .$$

With this notation, we have

$$E_2(X_i) = \frac{R_i^4}{n^2}\frac{(x'M_i^{-1}X_i)^2\tilde{d}_i}{(1+\frac{R_i^2}{n}d_i)^2} + \frac{R_i^4}{n^2}(x'M_i^{-1}X_i)^2\frac{\tilde{q}_i(X_i)-\tilde{d}_i}{(1+\frac{R_i^2}{n}d_i)^2} + \Delta_i\frac{R_i^4}{n^2}(x'M_i^{-1}X_i)^2\tilde{q}_i(X_i)$$

$$\triangleq \mathcal{M}_i(X_i) + \mathcal{R}_{2,1}(X_i) + \mathcal{R}_{2,2}(X_i) .$$

Note that by construction $\mathbf{E}_i\left(\mathcal{M}_i(X_i)\right) = \mathbf{E}_i\left(\mathcal{M}_i(Y_i)\right)$, so to bound $\mathbf{E}\left(E_2(X_i) - E_2(Y_i)\right)$, all we will have to do is bound $\mathbf{E}\left(|\mathcal{R}_{2,1}(X_i)|\right)$ and $\mathbf{E}\left(|\mathcal{R}_{2,2}(X_i)|\right)$. Before we turn to this task, let us recall that

$$|E_2(X_i)| \leq \frac{R_i^2}{n}(x'M_i^{-1}X_i)^2\,b(A,\Sigma_\epsilon) .$$

In other respects, if A and B are positive semi-definite (psd) matrices and $|||B|||_2 \leq C$, then $\text{trace}(AB) \leq C\text{trace}(A)$ (because when A and B are psd, $A^{1/2}BA^{1/2} \preceq |||B|||_2A$). Therefore,

$$\tilde{d}_i \leq b(A;\Sigma_\epsilon)d_i \text{ and } \mathcal{M}_i(X_i) \leq \frac{R_i^2}{n}(x'M_i^{-1}X_i)^2\,b(A,\Sigma_\epsilon) .$$

Hence,

$$|\mathcal{R}_{2,1}(X_i) + \mathcal{R}_{2,2}(X_i)| \leq K\frac{R_i^2}{n}(x'M_i^{-1}X_i)^2\,b(A,\Sigma_\epsilon) .$$

Let us now work more precisely on $\mathcal{R}_{2,1}(X_i)$ and $\mathcal{R}_{2,2}(X_i)$.

• **On $\mathcal{R}_{2,1}(X_i)$.**
Note that, using $|||M_i^{-1}\Sigma_\epsilon M_i^{-1}|||_2 \leq b(A;\Sigma_\epsilon)/t$, we have

$$\mathbf{E}_i\left(|\tilde{q}_i(X_i) - \tilde{d}_i(X_i)|^2\right) \leq \frac{b^2(A;\Sigma_\epsilon)}{t^2}b_{Q_2}(2;X_i) .$$

Using the Cauchy-Schwarz inequality in connection with the previous remark, we get

$$\mathbf{E}_i\left(|\mathcal{R}_{2,1}(X_i)|\right) \leq \frac{R_i^4}{n^{3/2}}(x'M_i^{-2}x)\frac{b(A;\Sigma_\epsilon)}{t}\sqrt{\frac{b_{Q_2}(2;X_i)}{n}}\sqrt{b_L(4;X_i)}$$

$$\leq \frac{R_i^4}{n^{3/2}}\frac{(x'A^{-1}x)}{t^2}b(A;\Sigma_\epsilon)\sqrt{\frac{b_{Q_2}(2;X_i)}{n}}\sqrt{b_L(4;X_i)} .$$

• **On $\mathcal{R}_{2,2}(X_i)$.**
Let us first note that

$$\frac{\tilde{q}_i(X_i)}{q_i(X_i)} \leq b(A;\Sigma_\epsilon) \text{ and } |\Delta_i| \leq K\frac{R_i^2}{n}\frac{|d_i - q_i|}{(1+\frac{R_i^2}{n}d_i)(1+\frac{R_i^2}{n}q_i(X_i))} .$$

Hence,

$$|\Delta_i\tilde{q}_i(X_i)| \leq Kb(A;\Sigma_\epsilon)\frac{|d_i - q_i|}{(1+\frac{R_i^2}{n}d_i)} .$$

We can therefore conclude that

$$|\mathcal{R}_{2,2}(X_i)| \leq K\frac{R_i^4}{n^2}b(A;\Sigma_\epsilon)\frac{|d_i - q_i|}{(1+\frac{R_i^2}{n}d_i)}(x'M_i^{-1}X_i)^2 .$$

Using the Cauchy-Schwarz inequality we also get

$$\mathbf{E}_i\left(|\mathcal{R}_{2,2}(X_i)|\right) \le K \frac{R_i^4}{n^{3/2}}\frac{b(A;\Sigma_\epsilon)}{t}(x'M_i^{-2}x)\sqrt{\frac{b_{Q_2}(2;X_i)}{n}}\sqrt{b_L(4;X_i)}$$

$$\le K\frac{R_i^4}{n^{3/2}}b(A;\Sigma_\epsilon)\frac{x'A^{-1}x}{t^2}\sqrt{\frac{b_{Q_2}(2;X_i)}{n}}\sqrt{b_L(4;X_i)}\ .$$

We conclude that if $U_i = \mathbf{E}\left(|\mathcal{R}_{2,1}(X_i) + \mathcal{R}_{2,2}(X_i)|\right)$,

$$|\mathbf{E}\left(E_2(X_i) - E_2(Y_i)\right)| \le U_i(X_i) + U_i(Y_i)\ ,$$

where

$$U_i(X_i) \le K\,b(A;\Sigma_\epsilon)\frac{x'A^{-1}x}{t}\left[\frac{R_i^4}{n^{3/2}}\sqrt{\frac{b_{Q_2}(2;X_i)}{n}}\sqrt{b_L(4;X_i)}\frac{1}{t}\right] \wedge \frac{R_i^2}{n}b_L(2;X_i)\ .$$

Finally, putting everything together we have shown that

$$|\mathbf{E}\left(F(X) - F(Y)\right)| \le K\sum_{i=1}^{n} U_i(X_i) + U_i(Y_i)\ .$$

We conclude that when R_i are independent and have 2 moments, the upper bound goes to zero in R_i-probability, provided $b_L(4;X_i)$ and $b_{Q_2}(2;X_i)/n$ remain uniformly bounded (we have already analyzed similar series previously). $\qquad\square$

3.4.2 Forms in $\alpha'\frac{DX'}{\sqrt{n}}M^{-1}\Sigma_\epsilon M^{-1}\frac{X'D}{\sqrt{n}}\alpha$

In the analysis of quantities of the type

$$\widehat{\mu}'(\widehat{\Sigma} + A)^{-1}\Sigma_\epsilon(\widehat{\Sigma} + A)^{-1}\widehat{\mu}$$

we will naturally have to understand quantities of the type, if $M = X'D^2X/n + A$,

$$G(\alpha;X) = \alpha'\frac{DX}{\sqrt{n}}M^{-1}\Sigma_\epsilon M^{-1}\frac{X'D}{\sqrt{n}}\alpha\ .$$

We work under our usual assumptions, and in particular $A \succeq t\mathrm{Id}$.
We have the following theorem.

Theorem 3.8. *Under the usual assumptions of this paper, when $\|\alpha\| = 1$, we have, for K a constant,*

$$\mathrm{var}\left(G(\alpha;X)\right) \le \sum_{i=1}^{n} V_i,\ \ with$$

$$V_i \le Kb^2(A;\Sigma_\epsilon)\left\{\alpha_i^4\left[\frac{R_i^4}{n^2}\frac{b_{Q_2}(2;X_i)}{t^2} \wedge 1\right]\right.$$

$$+ \left[\left(\alpha_i^4\frac{R_i^4}{n^2}b_{Q_2}(2;X_i)\frac{1}{t^2} + \frac{R_i^4}{n^2}b_L(4;X_i)\frac{1}{t^2} + \alpha_i^2\frac{R_i^2}{n}\frac{b_L(2;X_i)}{t}\right) \wedge 1\right.$$

$$\left.\left.+ \frac{R_i^4}{n^2}b_L(4;X_i)\frac{1}{t^2} + \alpha_i^2\frac{R_i^2}{n}b_L(2;X_i)\frac{1}{t}\right]\right\}\ .$$

Furthermore,

$$|\mathbf{E}\left(G(\alpha;X) - G(\alpha;Y)\right)| \le \sum_{i=1}^{n} U_{i,1}(X_i) + U_{i,2}(X_i) + U_{i,1}(Y_i) + U_{i,2}(Y_i)\ ,\ \ where$$

$$U_{i,1}(X_i) \le K\frac{b(A;\Sigma_\epsilon)}{\sqrt{t}}\left(\frac{R_i^2}{\sqrt{n}}\sqrt{\frac{b_{Q_2}(2;X_i)}{nt}} \wedge 1\right)\left[\frac{|\alpha_i|R_i}{\sqrt{n}}\sqrt{b_L(2;X_i)} + \frac{R_i^2}{n}\sqrt{b_L(4;X_i)}\frac{1}{\sqrt{t}}\right]\ .$$

$$U_{i,2}(X_i) \le Kb(A;\Sigma_\epsilon)\left[(\alpha_i^2 + \frac{R_i^2}{n}\frac{b_L(2;X_i)}{t})\right.$$

$$\wedge\sqrt{\frac{b_{Q_2}(2;X_i)}{n}}\left[\frac{R_i^2}{\sqrt{n}}\left(\frac{\alpha_i^2}{t} + \frac{R_i^2}{n}\frac{1}{t^2}\sqrt{b_L(4;X_i)}\right) + \frac{1}{t^{3/2}}\left(\frac{R_i^4}{n^{3/2}}\frac{1}{t^{1/2}}\sqrt{b_L(4;X_i)} + \frac{|\alpha_i|R_i^3}{n}\sqrt{b_L(2;X_i)}\right)\right]\ .$$

It is shown in the course of the proof that the upper bounds go to zero in probability when R_i's are i.i.d and uniformly square integrable and $b_L(4; X_i)$ as well as $\sqrt{b_{Q_2}(2; X_i)/n}$ remain uniformly bounded.

We note that in the Gaussian case (i.e X_i are $\mathcal{N}(0, \Sigma_i)$), by the symmetry trick we have now used several times, it is clear that the off-diagonal elements of the matrix

$$\frac{DX}{\sqrt{n}} M^{-1} \Sigma_\epsilon M^{-1} \frac{X'D}{\sqrt{n}}$$

have mean 0. Hence, to understand $\mathbf{E}(G(\alpha; X))$, all that is needed is to understand the diagonal entries of

$$\frac{DX}{\sqrt{n}} M^{-1} \Sigma_\epsilon M^{-1} \frac{X'D}{\sqrt{n}} .$$

If we further assume that X_i have the same Σ, computations similar to the ones done in Subsubsection 3.3.2 (also using our derivative trick) and fairly standard random matrix results yield a reasonably simple expression. In the interest of space, and since this is a very simple problem, we do not state in more details the deterministic equivalent.

Proof. We use the same trick as in the previous subsection, namely calling

$$g(\alpha; X; A + u\Sigma_\epsilon) = \alpha' \frac{DX}{\sqrt{n}} (X'D^2 X/n + A + u\Sigma_\epsilon)^{-1} \frac{X'D}{\sqrt{n}} \alpha ,$$

we see that

$$G(\alpha; X) = - \left. \frac{\partial}{\partial u} \right|_{u=0} g(\alpha; X; A + u\Sigma_\epsilon) .$$

Hence we can use the refined understanding of g we have developed earlier to study G.

In particular, the key equation in the study of g was

$$g(\alpha; X; A + u\Sigma_\epsilon) - g_i(\alpha; X; A + u\Sigma_\epsilon) = \alpha_i^2 - \frac{(\alpha_i - R_i \zeta_i(X_i)/\sqrt{n})^2}{1 + \frac{R_i^2}{n} q_i(X_i)} .$$

with, if $D_i = D - R_i e_i e_i'$, (i.e D_i is D where we replace the (i, i) entry by a 0)

$$q_i(X_i; u) = X_i' [M_i(A + u\Sigma_\epsilon)]^{-1} X_i$$

$$\zeta_i(X_i; u) = X_i' [M_i(A + u\Sigma_\epsilon)]^{-1} m_i, \quad m_i = \frac{X'D_i \alpha}{\sqrt{n}} .$$

Hence, if $M_i = X'D_i^2 X/n + A$,

$$\left. \frac{\partial}{\partial u} \right|_{u=0} q_i(X_i) = -X_i' M_i^{-1} \Sigma_\epsilon M_i^{-1} X_i \triangleq -\tilde{q}_i(X_i) ,$$

$$\left. \frac{\partial}{\partial u} \right|_{u=0} \zeta_i(X_i) = -X_i' M_i^{-1} \Sigma_\epsilon M_i^{-1} m_i \triangleq -\tilde{\zeta}_i(X_i) .$$

Hence, if $G_i(\alpha; X)$ is the same statistic as $G(\alpha; X)$ where X_i is replaced by 0 (and hence it does not depend on X_i), we have

$$G(\alpha; X) - G_i(\alpha; X) = 2 \frac{R_i}{\sqrt{n}} \frac{\tilde{\zeta}_i(R_i \zeta_i/\sqrt{n} - \alpha_i)}{1 + \frac{R_i^2}{n} q_i(X_i)} - \frac{R_i^2}{n} \tilde{q}_i \left(\frac{(\alpha_i - R_i \zeta_i(X_i; u)/\sqrt{n})}{1 + \frac{R_i^2}{n} q_i(X_i; u)} \right)^2 .$$

In preparation for Lindeberg-style work below, we note that if Y_i and X_i have mean 0 and the same covariance Σ_i,

$$\mathbf{E}_i(\zeta_i(X_i)) = \mathbf{E}_i(\zeta_i(Y_i)) \qquad\qquad \mathbf{E}_i(\zeta_i^2(X_i)) = \mathbf{E}_i(\zeta_i^2(Y_i))$$

$$\mathbf{E}_i\left(\zeta_i(X_i)\tilde{\zeta}_i(X_i)\right) = \mathbf{E}_i\left(\zeta_i(Y_i)\tilde{\zeta}_i(Y_i)\right) \qquad\qquad \mathbf{E}_i\left(\tilde{\zeta}_i^2(X_i)\right) = \mathbf{E}_i\left(\tilde{\zeta}_i^2(Y_i)\right) .$$

Recall also that $\|M_i^{-1/2}m_i\| \leq 1$, $\|M_i^{-1}m_i\| \leq t^{-1/2}$, so $\|M_i^{-1}\Sigma_\epsilon M_i^{-1}m_i\| \leq b(A;\Sigma_\epsilon)/\sqrt{t}$. We therefore have the estimates

$$\mathbf{E}\left(|\tilde{\zeta}_i(X_i)|^k\right) \leq \left(\frac{b(A;\Sigma_\epsilon)}{\sqrt{t}}\right)^k b_L(k;X_i) \text{ and } \mathbf{E}\left(|\zeta_i(X_i)|^k\right) \leq b_L(k;X_i)t^{-k/2} .$$

Let us call, if $\psi_i(X_i) = (\alpha_i - \frac{R_i}{\sqrt{n}}\zeta_i(X_i))^2$,

$$E_1(X_i) = \frac{R_i}{\sqrt{n}}\frac{\tilde{\zeta}_i(X_i)(R_i/\sqrt{n}\zeta_i - \alpha_i)}{1 + \frac{R_i^2}{n}q_i(X_i)}$$

$$E_2(X_i) = \frac{R_i^2/n\,\tilde{q}_i(X_i)}{1 + \frac{R_i^2}{n}q_i(X_i)}\frac{\psi_i(X_i)}{1 + \frac{R_i^2}{n}q_i(X_i)} .$$

Clearly,

$$G(\alpha;X) - G_i(\alpha;X) = 2E_1(X_i) - E_2(X_i) .$$

● **Efron-Stein aspects**

The aim here is to find $Z_{i,1}$, independent of X_i such that we can control $\mathbf{E}\left(|E_1(X_i) - Z_{i,1}|^2\right)$ and similarly for $E_2(X_i)$, we will try to find a $Z_{i,2}$ such that we control $\mathbf{E}\left(|E_2(X_i) - Z_{i,2}|^2\right)$. This will give us control of $\mathrm{var}\left(G(\alpha;X)\right)$.

1) Controlling $\mathbf{E_1(X_i)}$ Let us call

$$T_{1,i} = \frac{R_i}{\sqrt{n}}\frac{\alpha_i\tilde{\zeta}_i(X_i)}{1 + \frac{R_i^2}{n}q_i(X_i)} .$$

Clearly, $T_{1,i}^2 \leq \alpha_i^2 R_i^2/n\tilde{\zeta}_i^2(X_i)$ and therefore

$$\mathbf{E}\left(T_{1,i}^2\right) \leq \alpha_i^2\frac{R_i^2}{n}b_L(2;X_i)\frac{b^2(A;\Sigma_\epsilon)}{t} .$$

This term will not cause problem in our analysis as $\sum_{i=1}^n \mathbf{E}\left(T_{1,i}^2\right)$ will clearly go to zero when R_i's have two moments and $b(A;\Sigma_\epsilon)$ as well as $b_L(k;X_i)$ remain bounded. (Recall that $\|\alpha\| = 1$.)

Let us call

$$T_{2,i} = \frac{R_i^2}{n}\frac{\zeta_i\tilde{\zeta}_i}{1 + \frac{R_i^2}{n}q_i(X_i)} .$$

Clearly,

$$\mathbf{E}_i\left(|T_{2,i}|^k\right) \leq \frac{R_i^{2k}}{n^k}\sqrt{\mathbf{E}_i\left(\zeta_i^{2k}\right)\mathbf{E}_i\left(\tilde{\zeta}_i^{2k}\right)} \leq \frac{R_i^{2k}}{n^k}b_L(2k;X_i)\frac{b^k(A;\Sigma_\epsilon)}{t^k} .$$

In particular,

$$\mathbf{E}\left(|T_{2,i}|^2\right) \leq \frac{R_i^4}{n^2}b_L(4;X_i)\frac{b^2(A;\Sigma_\epsilon)}{t^2} .$$

(Since, when R_i's are i.i.d and have two moments, $\sum_i R_i^4/n^2 \to 0$ a.s, this terms is again not going to create any problems when we try to control the variance of G.)

So we have shown that

$$\mathbf{E}_i\left(E_1(X_i)^2\right) \leq 2\left[\frac{R_i^4}{n^2}b_L(4;X_i)\frac{1}{t} + \alpha_i^2\frac{R_i^2}{n}b_L(2;X_i)\right]\frac{b^2(A;\Sigma_\epsilon)}{t} .$$

2) Controlling $\mathbf{E_2(X_i)}$ Recall the decomposition from the proof of Theorem 3.4

$$\frac{\psi_i(X_i)}{1 + \frac{R_i^2}{n}q_i(X_i)} - \frac{\alpha_i^2}{1 + \frac{R_i^2}{n}d_i} = \alpha_i^2\frac{R_i^2}{n}\frac{q_i(X_i) - d_i}{(1 + \frac{R_i^2}{n}q_i)(1 + \frac{R_i^2}{n}d_i)} + \frac{1}{1 + \frac{R_i^2}{n}q_i}(R_i^2/n\zeta_i^2 - 2\alpha_iR_i/\sqrt{n}\zeta_i) .$$

Let us call

$$\Delta_{2,i}(X_i) \triangleq \alpha_i^2\frac{R_i^2}{n}\frac{q_i(X_i) - d_i}{(1 + \frac{R_i^2}{n}q_i)(1 + \frac{R_i^2}{n}d_i)} + \frac{1}{1 + \frac{R_i^2}{n}q_i}(R_i^2/n\zeta_i^2 - 2\alpha_iR_i/\sqrt{n}\zeta_i) .$$

35

Let us note that since $|\frac{R_i^2/n\,\tilde{q}_i(X_i)}{1+\frac{R_i^2}{n}q_i(X_i)}| \le b(A;\Sigma_\epsilon)$, we will have, using the work we did in the proof of Theorem 3.4,

$$\mathbf{E}_i\left(\left[\frac{R_i^2/n\,\tilde{q}_i(X_i)}{1+\frac{R_i^2}{n}q_i(X_i)}\Delta_{2,i}(X_i)\right]^2\right) \le Kb^2(A;\Sigma_\epsilon)\left[\left(\alpha_i^4\frac{R_i^4}{n^2}b_{Q_2}(2;X_i)\frac{1}{t^2} + \frac{R_i^4}{n^2}b_L(4;X_i)\frac{1}{t^2} + \alpha_i^2\frac{R_i^2}{n}\frac{b_L(2;X_i)}{t}\right)\wedge 1\right].$$

As we saw then, these terms will not cause any problem in our eventual control of the variance.

So we just need to focus on understanding

$$\mathcal{R}_{2,i}(X_i) = \frac{R_i^2/n\,\tilde{q}_i(X_i)}{1+\frac{R_i^2}{n}q_i(X_i)}\frac{\alpha_i^2}{1+\frac{R_i^2}{n}d_i}\,.$$

Now recall that we called $\frac{R_i^2}{n}\delta_i(X_i) = 1/(1+\frac{R_i^2}{n}q_i(X_i)) - 1/(1+\frac{R_i^2}{n}d_i)$; with this notation, we have

$$\frac{R_i^2/n\,\tilde{q}(X_i)}{1+\frac{R_i^2}{n}q_i(X_i)} = \frac{R_i^2/n\,\tilde{d}_i}{1+\frac{R_i^2}{n}d_i} + \frac{R_i^2}{n}\frac{\tilde{q}_i(X_i)-\tilde{d}_i}{1+\frac{R_i^2}{n}q_i(X_i)} + \frac{R_i^2}{n}\tilde{d}_i\frac{R_i^2}{n}\delta_i(X_i)\,.$$

Hence,

$$\mathcal{R}_{2,i}(X_i) - \frac{R_i^2/n\,\tilde{d}_i}{1+\frac{R_i^2}{n}d_i}\frac{\alpha_i^2}{1+\frac{R_i^2}{n}d_i} = \frac{\alpha_i^2}{1+\frac{R_i^2}{n}d_i}\left[\frac{R_i^2}{n}\frac{\tilde{q}_i(X_i)-\tilde{d}_i}{1+\frac{R_i^2}{n}q_i(X_i)} + \frac{R_i^2}{n}\tilde{d}_i\frac{R_i^2}{n}\delta_i(X_i)\right]\,.$$

Since $\tilde{d}_i(X_i) \le b(A;\Sigma_\epsilon)d_i$, we have

$$\frac{R_i^2}{n}\frac{\tilde{d}_i}{1+\frac{R_i^2}{n}d_i} \le b(A;\Sigma_\epsilon)\,.$$

We have seen that

$$\mathbf{E}_i\left(|\tilde{q}_i(X_i)-\tilde{d}_i(X_i)|^2\right) \le |||M_i^{-1}\Sigma_\epsilon M_i^{-1}|||_2^2 b_{Q_2}(2;X_i) \le \frac{b^2(A;\Sigma_\epsilon)}{t^2}b_{Q_2}(2;X_i)\,.$$

Furthermore, we saw previously that

$$\mathbf{E}_i\left(\delta_i^2(X_i)\right) \le \frac{1}{t^2}b_{Q_2}(2;X_i)\,.$$

So we have

$$\mathbf{E}_i\left(\left[\mathcal{R}_{2,i}(X_i) - \frac{R_i^2\tilde{d}_i}{1+\frac{R_i^2}{n}d_i}\frac{\alpha_i^2}{1+\frac{R_i^2}{n}d_i}\right]^2\right) \le K\alpha_i^4\frac{R_i^4}{n^2}\frac{b_{Q_2}(2;X_i)}{t^2}b^2(A;\Sigma_\epsilon)\,.$$

On the other hand,

$$\left|\mathcal{R}_{2,i}(X_i) - \frac{R_i^2/n\,\tilde{d}_i}{1+\frac{R_i^2}{n}d_i}\frac{\alpha_i^2}{1+\frac{R_i^2}{n}d_i}\right| \le Kb(A;\Sigma_\epsilon)\alpha_i^2\,.$$

So we conclude that

$$\mathbf{E}_i\left(\left[\mathcal{R}_{2,i}(X_i) - \frac{R_i^2\tilde{d}_i}{1+\frac{R_i^2}{n}d_i}\frac{\alpha_i^2}{1+\frac{R_i^2}{n}d_i}\right]^2\right) \le K\alpha_i^4 b^2(A;\Sigma_\epsilon)\left[\frac{R_i^4}{n^2}\frac{b_{Q_2}(2;X_i)}{t^2}\wedge 1\right]\,.$$

Hence,

$$\mathbf{E}_i\left(\left[E_2(X_i) - \frac{R_i^2\tilde{d}_i}{1+\frac{R_i^2}{n}d_i}\frac{\alpha_i^2}{1+\frac{R_i^2}{n}d_i}\right]^2\right) \le Kb^2(A;\Sigma_\epsilon)\left\{\alpha_i^4\left[\frac{R_i^4}{n^2}\frac{b_{Q_2}(2;X_i)}{t^2}\wedge 1\right]\right.$$
$$\left. + \left[\left(\alpha_i^4\frac{R_i^4}{n^2}b_{Q_2}(2;X_i)\frac{1}{t^2} + \frac{R_i^4}{n^2}b_L(4;X_i)\frac{1}{t^2} + \alpha_i^2\frac{R_i^2}{n}\frac{b_L(2;X_i)}{t}\right)\wedge 1\right]\right\}\,.$$

So if
$$Z_i = G_i(\alpha; X) - \frac{R_i^2 \tilde{d}_i}{1 + \frac{R_i^2}{n} d_i} \frac{\alpha_i^2}{1 + \frac{R_i^2}{n} d_i} ,$$

and $Z = G(\alpha; X)$, we have shown that

$$\mathbf{E}_i \left(|Z - Z_i|^2 \right) \leq K b^2(A; \Sigma_\epsilon) \left\{ \alpha_i^4 \left[\frac{R_i^4}{n^2} \frac{b_{Q_2}(2; X_i)}{t^2} \wedge 1 \right] \right.$$
$$+ \left[\left(\alpha_i^4 \frac{R_i^4}{n^2} b_{Q_2}(2; X_i) \frac{1}{t^2} + \frac{R_i^4}{n^2} b_L(4; X_i) \frac{1}{t^2} + \alpha_i^2 \frac{R_i^2}{n} \frac{b_L(2; X_i)}{t} \right) \wedge 1 \right.$$
$$\left. \left. + \frac{R_i^4}{n^2} b_L(4; X_i) \frac{1}{t^2} + \alpha_i^2 \frac{R_i^2}{n} b_L(2; X_i) \frac{1}{t} \right] \right\} .$$

This bound is sufficient to allow us to apply the Efron-Stein inequality, as we saw earlier: as soon as the R_i's have two moments and are i.i.d, the (sum over i of the) upper bound goes to zero.

• **Lindeberg aspects**

1) Controlling $\mathbf{E}\left(\mathbf{E_1}(\mathbf{X_i}) - \mathbf{E_1}(\mathbf{Y_i})\right)$

To alleviate the notation, we make a slight abuse of notation and change the meaning of M_i compared to what was used in the previous part of the proof: because we are now in the Lindeberg setting, the matrix M_i is (as usual) computed by using $(X_1, \ldots, X_{i-1}, 0, Y_{i+1}, \ldots, Y_n)$, but it is still independent of X_i and Y_i.

Recall that
$$E_1(X_i) = \frac{R_i}{\sqrt{n}} \frac{\tilde{\zeta}_i(X_i)(R_i/\sqrt{n}\zeta_i - \alpha_i)}{1 + \frac{R_i^2}{n} q_i(X_i)} \triangleq \frac{N_i(X_i)}{1 + \frac{R_i^2}{n} q_i(X_i)} .$$

Calling as usual
$$\frac{R_i^2}{n} \delta_i(X_i) = \frac{1}{1 + \frac{R_i^2}{n} q_i(X_i)} - \frac{1}{1 + \frac{R_i^2}{n} d_i} ,$$

we have
$$E_1(X_i) = \frac{N_i(X_i)}{1 + \frac{R_i^2}{n} d_i} + \frac{R_i^2}{n} N_i(X_i) \delta_i(X_i) .$$

Of course $\mathbf{E}\left(N_i(X_i)\right) = \mathbf{E}\left(N_i(Y_i)\right)$ when X_i and Y_i have mean 0 and the same covariance, Σ_i. Therefore, since $d_i = \mathrm{trace}\left(\Sigma_i M_i^{-1}\right) = \mathbf{E}_i\left(X_i' M_i^{-1} X_i\right) = \mathbf{E}_i\left(Y_i' M_i^{-1} Y_i\right)$, we have

$$\mathbf{E}\left(\frac{N_i(X_i)}{1 + \frac{R_i^2}{n} d_i} \right) = \mathbf{E}\left(\frac{N_i(Y_i)}{1 + \frac{R_i^2}{n} d_i} \right) .$$

The only question now is to try to control the remainder term

$$\mathcal{R}_1(X_i) = \frac{R_i^2}{n} N_i(X_i) \delta_i(X_i) = E_1(X_i) \frac{R_i^2}{n} \frac{(d_i - q_i(X_i))}{1 + \frac{R_i^2}{n} d_i(X_i)} .$$

We have, after using Cauchy-Schwarz and our usual bounds,

$$\mathbf{E}_i \left(|\mathcal{R}_1(X_i)| \right) \leq \frac{R_i^2}{\sqrt{n}} \sqrt{\frac{b_{Q_2}(2; X_i)}{nt}} \sqrt{\mathbf{E}_i\left(E_1^2(X_i)\right)} .$$

Using the results we got earlier on $\sqrt{\mathbf{E}_i\left(E_1^2(X_i)\right)}$, we have

$$\mathbf{E}_i \left(|\mathcal{R}_1(X_i)| \right) \leq K \frac{R_i^2}{\sqrt{n}} \sqrt{\frac{b_{Q_2}(2; X_i)}{nt}} \left(\frac{R_i^2}{n} \frac{1}{\sqrt{t}} \sqrt{b_L(4; X_i)} + |\alpha_i| \frac{R_i}{\sqrt{n}} \sqrt{b_L(2; X_i)} \right) \frac{b(A; \Sigma_\epsilon)}{\sqrt{t}} .$$

On the other hand,
$$|\mathcal{R}_1(X_i)| \leq |N_i(X_i)| .$$

From its definition, we see that

$$\mathbf{E}_i\left(|N_i(X_i)|\right) \leq \frac{|\alpha_i|R_i}{\sqrt{n}}\frac{b(A;\Sigma_\epsilon)}{\sqrt{t}}b_L(1;X_i) + \frac{R_i^2}{n}b_L(2;X_i)\frac{b(A;\Sigma_\epsilon)}{t} \ .$$

Hence,

$$\mathbf{E}_i\left(|\mathcal{R}_1(X_i)|\right) \leq K\frac{R_i^2}{\sqrt{n}}\frac{b(A;\Sigma_\epsilon)}{\sqrt{t}}\sqrt{\frac{b_{Q_2}(2;X_i)}{nt}}\left(\frac{R_i^2}{n}\frac{\sqrt{b_L(4;X_i)}}{\sqrt{t}} + |\alpha_i|\frac{R_i}{\sqrt{n}}\sqrt{b_L(2;X_i)}\right)$$
$$\wedge\ \frac{b(A;\Sigma_\epsilon)}{\sqrt{t}}\left(\frac{|\alpha_i|R_i}{\sqrt{n}}b_L(1;X_i) + \frac{R_i^2}{n}b_L(2;X_i)\frac{1}{\sqrt{t}}\right) \ ,$$

and

$$\mathbf{E}_i\left(|\mathcal{R}_1(X_i)|\right) \leq K\frac{b(A;\Sigma_\epsilon)}{\sqrt{t}}\left(\frac{R_i^2}{\sqrt{n}}\sqrt{\frac{b_{Q_2}(2;X_i)}{nt}}\wedge 1\right)\left[\frac{|\alpha_i|R_i}{\sqrt{n}}\sqrt{b_L(2;X_i)} + \frac{R_i^2}{n}\sqrt{b_L(4;X_i)}\frac{1}{\sqrt{t}}\right] \ .$$

At this point, we would like to show that the control we have is sufficient for the Lindeberg method to work when R_i's are i.i.d and have two moments. For this, it is sufficient to show that

$$\mathbf{E}\left(\sum_{i=1}^n |\alpha_i|(R_i/\sqrt{n}\wedge R_i^3/n)\right) \to 0 \ .$$

We will simply show that $\mathbf{E}\left((R_i/\sqrt{n}\wedge R_i^3/n)\right) = \mathrm{o}(n^{-1/2})$. We note that, since $R_i \geq 0$,

$$\mathbf{E}\left(R_i \wedge n^{-1/2}R_i^3\right) = \mathbf{E}\left(R_i 1_{R_i\geq n^{1/4}}\right) + \mathbf{E}\left(R_i^3 n^{-1/2}1_{R_i\leq n^{1/4}}\right) \leq \mathbf{E}\left(R_i 1_{R_i\geq n^{1/4}}\right) + n^{-1/4}\mathbf{E}\left(R_i^2\right) \ .$$

Since R_i has two moments (and hence one), the monotone convergence theorem guarantees that

$$\mathbf{E}\left(R_i \wedge n^{-1/2}R_i^3\right) = \mathrm{o}(1) \ .$$

We now remark that since $\|\alpha\| = 1$, $\|\alpha\|_1 \leq \sqrt{n}$. Therefore,

$$\mathbf{E}\left(\sum_{i=1}^n |\alpha_i|(R_i/\sqrt{n}\wedge R_i^3/n)\right) = \|\alpha\|_1\mathbf{E}\left((R_i/\sqrt{n}\wedge R_i^3/n)\right) = \mathrm{o}(\|\alpha\|_1/n^{1/2}) = \mathrm{o}(1) \ .$$

2) Controlling $\mathbf{E}\left(\mathbf{E_2(X_i)} - \mathbf{E_2(Y_i)}\right)$
Recall the notation
$$\Delta_i = \frac{1}{(1 + \frac{R_i^2}{n}q_i(X_i))^2} - \frac{1}{(1 + \frac{R_i^2}{n}d_i)^2} \ .$$

Let us write
$$\psi_i(X_i) = \alpha_i^2 - 2\alpha_i\frac{R_i}{\sqrt{n}}\zeta_i + \frac{R_i^2}{n}\zeta_i^2 = \alpha_i^2 - \Gamma_i \ .$$

By definition,
$$E_2(X_i) = \frac{R_i^2\tilde{q}_i(X_i)/n}{(1 + \frac{R_i^2}{n}q_i(X_i))^2}\psi_i(X_i) \ .$$

Therefore,

$$E_2(X_i) = \frac{\alpha_i^2\tilde{q}_i(X_i) - \Gamma_i\tilde{d}_i}{(1 + \frac{R_i^2}{n}d_i)^2}\frac{R_i^2}{n} + \frac{R_i^2}{n}\frac{\Gamma_i(\tilde{d}_i - \tilde{q}_i)}{(1 + \frac{R_i^2}{n}d_i)^2} + \Delta_i(X_i)\psi_i(X_i)\tilde{q}_i(X_i)\frac{R_i^2}{n} \ ,$$
$$= \mathcal{M}_2(X_i) + \mathcal{R}_{2,1}(X_i) + \mathcal{R}_{2,2}(X_i) \ .$$

It is clear that when X_i and Y_i have the same covariance Σ_i and mean 0, $\mathbf{E}_i\left(\mathcal{M}_2(X_i)\right) = \mathbf{E}_i\left(\mathcal{M}_2(Y_i)\right)$. Hence, in controlling $\mathbf{E}\left(E_2(X_i) - E_2(Y_i)\right)$, all we will have to do is control

$$\mathbf{E}_i\left(|\mathcal{R}_{2,1}(X_i) + \mathcal{R}_{2,2}(X_i)|\right) .$$

a) Controlling $\mathbf{E}\left(|\mathcal{R}_{2,1}(X_i)|\right)$

We have, using the Cauchy-Schwarz inequality:

$$\mathbf{E}_i\left(\zeta_i^2|\tilde{d}_i - \tilde{q}_i(X_i)|\right) \leq \frac{b(A;\Sigma_\epsilon)}{t^2}\sqrt{b_L(4;X_i)}\sqrt{b_{Q_2}(2;X_i)} ,$$

$$\mathbf{E}_i\left(|\zeta_i||\tilde{d}_i - \tilde{q}_i(X_i)|\right) \leq \frac{b(A;\Sigma_\epsilon)}{t^{3/2}}\sqrt{b_L(2;X_i)}\sqrt{b_{Q_2}(2;X_i)} .$$

Therefore,

$$\frac{R_i^4}{n^2}\mathbf{E}_i\left(\frac{\zeta_i^2|\tilde{d}_i - \tilde{q}_i(X_i)|}{(1 + \frac{R_i^2}{n}d_i)^2}\right) \leq \frac{R_i^4}{n^{3/2}}\frac{b(A;\Sigma_\epsilon)}{t^2}\sqrt{b_L(4;X_i)}\sqrt{b_{Q_2}(2;X_i)/n} .$$

and

$$|\alpha_i|\frac{R_i^3}{n^{3/2}}\mathbf{E}_i\left(\frac{|\zeta_i||\tilde{d}_i - \tilde{q}_i(X_i)|}{(1 + \frac{R_i^2}{n}d_i)^2}\right) \leq \frac{|\alpha_i|R_i^3}{n}\frac{b(A;\Sigma_\epsilon)}{t^{3/2}}\sqrt{b_L(2;X_i)}\sqrt{b_{Q_2}(2;X_i)/n}$$

We conclude that

$$\mathbf{E}_i\left(|\mathcal{R}_{2,1}(X_i)|\right) \leq K\frac{b(A;\Sigma_\epsilon)}{t^{3/2}}\sqrt{b_{Q_2}(2;X_i)/n}\left(\frac{R_i^4}{n^{3/2}}\frac{1}{t^{1/2}}\sqrt{b_L(4;X_i)} + \frac{|\alpha_i|R_i^3}{n}\sqrt{b_L(2;X_i)}\right) .$$

b) Controlling $\mathbf{E}\left(|\mathcal{R}_{2,2}(X_i)|\right)$

Since

$$\left|\frac{R_i^2}{n}\tilde{q}_i(X_i)\Delta_i(X_i)\right| \leq Kb(A;\Sigma_\epsilon)\frac{R_i^2}{n}\frac{|q_i - d_i|}{1 + \frac{R_i^2}{n}d_i} ,$$

we have

$$\mathbf{E}_i\left(\left|\Delta_i(X_i)\psi_i(X_i)\frac{R_i^2}{n}\tilde{q}_i(X_i)\right|\right) \leq Kb(A;\Sigma_\epsilon)\mathbf{E}_i\left((\alpha_i^2 + \frac{R_i^2}{n}\zeta_i^2)\frac{R_i^2}{n}\frac{|q_i - d_i|}{1 + \frac{R_i^2}{n}d_i}\right) ,$$

Hence,

$$\mathbf{E}_i\left(\left|\Delta_i(X_i)\psi_i(X_i)\frac{R_i^2}{n}\tilde{q}_i(X_i)\right|\right) \leq Kb(A;\Sigma_\epsilon)\frac{R_i^2}{\sqrt{n}}\sqrt{b_{Q_2}(2;X_i)/n}\left(\frac{\alpha_i^2}{t} + \frac{R_i^2}{n}\frac{1}{t^2}\sqrt{b_L(4;X_i)}\right) .$$

c) Controlling $|\mathbf{E}_i\left(E_2(X_i) - E_2(Y_i)\right)|$

We note that $|E_2(X_i)| \leq \psi_i(X_i)b(A;\Sigma_\epsilon)$ and therefore

$$\mathbf{E}_i\left(|E_2(X_i)|\right) \leq Kb(A;\Sigma_\epsilon)(\alpha_i^2 + \frac{R_i^2}{n}\frac{b_L(2;X_i)}{t}) .$$

We can finally conclude that

$$|\mathbf{E}_i\left(E_2(X_i) - E_2(Y_i)\right)| \leq \Phi_1(X_i) + \Phi_1(Y_i) ,$$

where

$$\Phi_1(X_i) \leq Kb(A;\Sigma_\epsilon)\left[(\alpha_i^2 + \frac{R_i^2}{n}\frac{b_L(2;X_i)}{t})\right.$$

$$\wedge\sqrt{\frac{b_{Q_2}(2;X_i)}{n}}\left[\frac{R_i^2}{\sqrt{n}}\left(\frac{\alpha_i^2}{t} + \frac{R_i^2}{n}\frac{1}{t^2}\sqrt{b_L(4;X_i)}\right)\right.$$

$$\left.\left. + \frac{1}{t^{3/2}}\left(\frac{R_i^4}{n^{3/2}}\frac{1}{t^{1/2}}\sqrt{b_L(4;X_i)} + \frac{|\alpha_i|R_i^3}{n}\sqrt{b_L(2;X_i)}\right)\right]\right] .$$

This expression is somewhat unseemly, however, assuming that b_L, b_{Q_2}/n and $b(A; \Sigma_\epsilon)$ stay bounded we see that it is of the form

$$\Phi_1(X_i) \leq K \left(\frac{R_i^2}{\sqrt{n}} (\frac{R_i^2}{n} + \alpha_i^2 + \frac{|\alpha_i| R_i}{\sqrt{n}}) \right) \wedge (\alpha_i^2 + \frac{R_i^2}{n}) \leq K \left(\alpha_i^2 + \frac{R_i^2}{n} \right) \left(\frac{R_i^2}{\sqrt{n}} \wedge 1 \right) .$$

We have already seen how to control this expression when R_i are i.i.d and uniformly square integrable in the proof of Theorem 3.5. So we conclude that when this is the case $\sum_{i=1}^n \Phi_i(X_i)$ will tend to 0 (for instance in R_i-probability).

$\square$

3.4.3 Forms in $\frac{1}{\sqrt{n}} \alpha' D X' M^{-1} \Sigma_\epsilon M^{-1} x$

The third and last situation we need to consider are forms of the type

$$H(\alpha; X) = \frac{1}{\sqrt{n}} \alpha' D X' M^{-1} \Sigma_\epsilon M^{-1} x ,$$

where as usual

$$M = \frac{1}{n} X' D^2 X + A .$$

We work under our usual assumptions, and in particular $A \succeq t\mathrm{Id}$.

We have the following theorem.

Theorem 3.9. *Under the usual assumptions of this paper (see Subsection 3.2), we have, for K a constant,*

$$\mathrm{var}\left(H(\alpha; X) \right) \leq \sum_{i=1}^n V_i, \ \ with$$

$$V_i \leq K b^2(A; \Sigma_\epsilon) \frac{x' A^{-1} x}{t} \left[\frac{R_i^2}{n} \alpha_i^2 b_L(2; X_i) + \frac{R_i^4}{n^2} b_L(4; X_i) \frac{1}{t} \right] .$$

$$\left| \mathbf{E} \left(H(\alpha; X) - H(\alpha; Y) \right) \right| \leq \sum_{i=1}^n U_i(X_i) + U_i(Y_i), \ \ where$$

$$U_i(X_i) \leq K b(A; \Sigma_\epsilon) \sqrt{\frac{x' A^{-1} x}{t}} \left[\frac{1}{\sqrt{n}} |\alpha_i| R_i \sqrt{b_L(2; X_i)} + \frac{R_i^2}{n} \sqrt{\frac{b_L(4; X_i)}{t}} \right] \left[\frac{R_i^2}{n} \frac{\sqrt{b_{Q_2}(2; X_i)}}{t} \wedge 1 \right]$$

The proof of the theorem uses the same ideas as before and will rely on the work of Subsubsection 3.3.3.

We also note that by the same symmetry arguments as before, in the Gaussian case, we trivially have $\mathbf{E}\left(H(\alpha; X) \right) = 0$.

Proof. Naturally, $H(\alpha; X)$ is closely related to

$$h(\alpha; X) = \frac{1}{\sqrt{n}} \alpha' D X' M^{-1} x$$

which we studied earlier. Recall that we got the key decomposition

$$h(\alpha; X; A) = h(\alpha; X) = h_i(\alpha; X) + \frac{\varphi_i(X_i)}{1 + \frac{R_i^2}{n} q_i(X_i)} ,$$

where h_i did not involve X_i and

$$\varphi_i(X_i) = \frac{1}{\sqrt{n}} \alpha_i R_i X_i' M_i^{-1} x - \frac{R_i^2}{n} X_i' M_i^{-1} m_i X_i' M_i^{-1} x ,$$
$$q_i(X_i) = X_i' M_i^{-1} X_i .$$

As before, we can deduce H from $h(\alpha; X; A + u\Sigma_\epsilon)$ by taking the derivative of the latter with respect to u and appropriately modifying the sign.

We call

$$H_i = -\frac{\partial h_i(\alpha; X; A + u\Sigma_\epsilon)}{\partial u}$$

$$\Upsilon_i(X_i) \triangleq -\frac{\partial \varphi_i(X_i; A + u\Sigma_\epsilon)}{\partial u}$$

$$= \frac{R_i}{\sqrt{n}}\alpha_i X_i' M_i^{-1}\Sigma_\epsilon M_i^{-1}x - \frac{R_i^2}{n}\left[X_i' M_i^{-1}\Sigma_\epsilon M_i^{-1}m_i X_i' M_i^{-1}x + X_i' M_i^{-1}m_i X_i' M_i^{-1}\Sigma_\epsilon M_i^{-1}x\right]$$

$$\tilde{q}_i(X_i) = X_i' M_i^{-1}\Sigma_\epsilon M_i^{-1}X_i \, .$$

The new "key equality" is

$$H(\alpha; X) = H_i(\alpha; X) - \frac{\Upsilon_i}{1 + \frac{R_i^2}{n}q_i(X_i)} + \frac{R_i^2}{n}\tilde{q}_i(X_i)\frac{\varphi_i(X_i)}{(1 + \frac{R_i^2}{n}q_i(X_i))^2} \, .$$

We are now in a position to do our usual analysis with the Efron-Stein inequality and the Lindeberg approach.

• **Efron-Stein aspects** Because $H_i(\alpha; X)$ does not involve X_i, we clearly have

$$\mathrm{var}\left(H(\alpha; X)\right) \leq \sum_{i=1}^{n} \mathrm{var}\left(H(\alpha; X) - H_i(\alpha; X_i)\right) \, .$$

Now, clearly,

$$V_i = \mathrm{var}\left(H(\alpha; X) - H_i(\alpha; X_i)\right) \leq K\left[\mathbf{E}\left(\left(\frac{\Upsilon_i(X_i)}{1 + \frac{R_i^2}{n}q_i(X_i)}\right)^2\right) + \frac{R_i^4}{n^2}\mathbf{E}\left(\frac{\tilde{q}_i(X_i)^2\varphi_i^2(X_i)}{(1 + \frac{R_i^2}{n}q_i(X_i))^2}\right)\right] \, .$$

If $v = M_i^{-1}\Sigma_\epsilon M_i^{-1}x$, we have seen that $\|v\| \leq b(A; \Sigma_\epsilon)\sqrt{x'A^{-1}x/t}$. So we conclude that

$$\mathbf{E}\left(\left(\frac{R_i}{\sqrt{n}}\alpha_i X_i' M_i^{-1}\Sigma_\epsilon M_i^{-1}x\right)^2\right) \leq \frac{R_i^2}{n}\alpha_i^2 b_L(2; X_i)b^2(A; \Sigma_\epsilon)\frac{x'A^{-1}x}{t} \, .$$

Recall now that $\|M_i^{-1}m_i\| \leq \frac{\|\alpha\|}{\sqrt{t}} \leq \frac{1}{\sqrt{t}}$ and $\|M_i^{-1}\Sigma_\epsilon M_i^{-1}m_i\| \leq \frac{b(A;\Sigma_\epsilon)}{\sqrt{t}}$. Therefore,

$$\mathbf{E}\left(\left[X_i' M_i^{-1}\Sigma_\epsilon M_i^{-1}m_i X_i' M_i^{-1}x + X_i' M_i^{-1}m_i X_i' M_i^{-1}\Sigma_\epsilon M_i^{-1}x\right]^2\right) \leq Kb_L(4; X_i)\frac{b^2(A; \Sigma_\epsilon)}{t}\frac{x'A^{-1}x}{t} \, .$$

We finally have

$$\mathbf{E}\left(\left(\frac{\Upsilon_i(X_i)}{1 + \frac{R_i^2}{n}q_i(X_i)}\right)^2\right) \leq K\frac{b^2(A; \Sigma_\epsilon)x'A^{-1}x}{t}\left[\frac{R_i^2}{n}\alpha_i^2 b_L(2; X_i) + \frac{R_i^4}{n^2}b_L(4; X_i)\frac{1}{t}\right] \, .$$

For the second part of this simple variance bounding exercise, we first remind the reader that

$$\left|\frac{\frac{R_i^2}{n}\tilde{q}_i(X_i)}{1 + \frac{R_i^2}{n}q_i(X_i)}\right| \leq b(A; \Sigma_\epsilon) \, .$$

Hence, we simply need to bound

$$\mathbf{E}\left(\left(\frac{\varphi_i(X_i)}{1 + \frac{R_i^2}{n}q_i(X_i)}\right)^2\right) \, ,$$

something we have essentially already done, and we get easily

$$\mathbf{E}\left(\left(\frac{\varphi_i(X_i)}{1+\frac{R_i^2}{n}q_i(X_i)}\right)^2\right) \leq K\left[\alpha_i^2\frac{R_i^2}{n}b_L(2;X_i)\frac{x'A^{-1}x}{t} + \frac{R_i^4}{n^2}\frac{x'A^{-1}x}{t^2}b_L(4;X_i)\right] .$$

We have therefore shown that

$$V_i \leq Kb^2(A;\Sigma_\epsilon)\frac{x'A^{-1}x}{t}\left[\frac{R_i^2}{n}\alpha_i^2 b_L(2;X_i) + \frac{R_i^4}{n^2}b_L(4;X_i)\frac{1}{t}\right] .$$

We note that when R_i are i.i.d uniformly square integrable, the Marcienkiewicz-Zygmund law of large numbers guarantees that $\sum_{i=1}^n V_i \to 0$ for instance in probability.

We now turn to Lindeberg-type questions.

• **Lindeberg aspects** As usual, we will go a bit fast here. Essentially the previous decomposition can still be used, but it should now be understood that the M_i matrix we are dealing with involves both $\{X_m\}_{m<i}$ and $\{Y_k\}_{k>i}$, instead of just $\{X_j\}_{j=1}^n$ or $\{Y_j\}_{j=1}^n$. However, the key fact is that M_i is independent of both X_i and Y_i. Hence, we have for instance

$$\mathbf{E}\left(\varphi_i(X_i)\right) = \mathbf{E}\left(\varphi_i(Y_i)\right)$$

and

$$\mathbf{E}\left(\Upsilon_i(X_i)\right) = \mathbf{E}\left(\Upsilon_i(Y_i)\right) .$$

Let us call

$$T_1(X_i) = \frac{\Upsilon_i(X_i)}{1+\frac{R_i^2}{n}q_i(X_i)}$$

and

$$T_2(X_i) = \frac{R_i^2}{n}\frac{\tilde{q}_i(X_i)\varphi_i(X_i)}{(1+\frac{R_i^2}{n}q_i(X_i))^2} .$$

It is clear that if we can control

$$\sum_{i=1}^n |\mathbf{E}\left(T_1(X_i) - T_1(Y_i)\right) - \mathbf{E}\left(T_2(X_i) - T_2(Y_i)\right)|$$

we will have control over $|\mathbf{E}\left(H(\alpha;X) - H(\alpha;Y)\right)|$. We recall that we have already showed that

$$|H(\alpha;X)| \leq K\frac{\|x\|b(A;\Sigma_\epsilon)}{\sqrt{t}} .$$

• **Control of $\mathbf{E}\left(T_1(X_i) - T_1(Y_i)\right)$**

As usual, we use the fact that

$$T_1(X_i) = \frac{\Upsilon_i(X_i)}{1+\frac{R_i^2}{n}d_i} + \Upsilon_i(X_i)\frac{R_i^2}{n}\frac{d_i - q_i(X_i)}{(1+\frac{R_i^2}{n}q_i(X_i))(1+\frac{R_i^2}{n}d_i)}$$

$$\triangleq T_{1,1}(X_i) + T_{1,2}(X_i) .$$

Naturally, since X_i and Y_i have mean 0 and the same covariance,

$$\mathbf{E}\left(\frac{\Upsilon_i(X_i)}{1+\frac{R_i^2}{n}d_i}\right) = \mathbf{E}\left(\frac{\Upsilon_i(Y_i)}{1+\frac{R_i^2}{n}d_i}\right) ,$$

so all that is left to do is control $\mathbf{E}\left(|T_{1,2}(X_i)|\right)$. To do so, we can use Cauchy-Schwarz and recall that

$$\mathbf{E}\left(\left(\frac{\Upsilon_i(X_i)}{1+\frac{R_i^2}{n}q_i(X_i)}\right)^2\right) \leq K\frac{b^2(A;\Sigma_\epsilon)x'A^{-1}x}{t}\left[\frac{R_i^2}{n}\alpha_i^2 b_L(2;X_i) + \frac{R_i^4}{n^2}b_L(4;X_i)\frac{1}{t}\right] .$$

and
$$\mathbf{E}\left((d_i - q_i)^2\right) \leq \frac{b_{Q_2}(2; X_i)}{t^2} \ .$$

Hence,

$$\mathbf{E}\left(|T_{1,2}(X_i)|\right) \leq K\frac{R_i^2}{n}\sqrt{\frac{b^2(A; \Sigma_\epsilon)x'A^{-1}x}{t}\left[\frac{R_i^2}{n}\alpha_i^2 b_L(2; X_i) + \frac{R_i^4}{n^2}b_L(4; X_i)\frac{1}{t}\right]}\frac{\sqrt{b_{Q_2}(2; X_i)}}{t}$$

$$\leq K\frac{R_i^2}{n}\frac{b(A; \Sigma_\epsilon)\sqrt{x'A^{-1}x}}{t^{3/2}}\left[\frac{R_i}{\sqrt{n}}|\alpha_i|\sqrt{b_L(2; X_i)} + \frac{R_i^2}{n}\sqrt{\frac{b_L(4; X_i)}{t}}\right]\sqrt{b_{Q_2}(2; X_i)} \ .$$

In other respects, let us note that

$$\mathbf{E}\left(|T_1(X_i)|\right) \leq \mathbf{E}\left(|\Upsilon_i(X_i)|\right) \ .$$

We have

$$\mathbf{E}\left(|\Upsilon_i(X_i)|\right) \leq \frac{R_i}{\sqrt{n}}b(A; \Sigma_\epsilon)\sqrt{\frac{x'A^{-1}x}{t}}\left[|\alpha_i|b_L(1; X_i) + \frac{R_i}{\sqrt{n}}\frac{2b_L(2; X_i)}{\sqrt{t}}\right] \ ,$$

$$\leq K\frac{R_i}{\sqrt{n}}b(A; \Sigma_\epsilon)\sqrt{\frac{x'A^{-1}x}{t}}\left[|\alpha_i|\sqrt{b_L(2; X_i)} + \frac{R_i}{\sqrt{n}}\frac{\sqrt{b_L(4; X_i)}}{\sqrt{t}}\right]$$

Hence,

$$|\mathbf{E}\left(T_1(X_i) - T_1(Y_i)\right)| \leq \Psi_i(X_i) + \Psi_i(Y_i) \ ,$$

where

$$\Psi_i(X_i) = K\frac{R_i}{\sqrt{n}}b(A; \Sigma_\epsilon)\sqrt{\frac{x'A^{-1}x}{t}}\left[|\alpha_i|\sqrt{b_L(2; X_i)} + \frac{R_i}{\sqrt{n}}\frac{\sqrt{b_L(4; X_i)}}{\sqrt{t}}\right]\left(1 \wedge \frac{R_i^2}{\sqrt{n}t}\sqrt{\frac{b_{Q_2}(2; X_i)}{n}}\right) \ .$$

• **Control of** $\mathbf{E}\left(\mathbf{T_2(X_i)} - \mathbf{T_2(Y_i)}\right)$
Recall that

$$T_2(X_i) = \frac{R_i^2}{n}\frac{\tilde{q}_i(X_i)\varphi_i(X_i)}{(1 + \frac{R_i^2}{n}q_i(X_i))^2}$$

Clearly, using the notation $\Delta_i(X_i) = 1/(1 + R_i^2/nq_i(X_i))^2 - 1/(1 + R_i^2/nd_i)^2$, we have

$$\frac{\tilde{q}_i(X_i)}{(1 + \frac{R_i^2}{n}q_i(X_i))^2} = \frac{\tilde{d}_i}{(1 + \frac{R_i^2}{n}d_i)^2} + \frac{\tilde{q}_i(X_i) - \tilde{d}_i}{(1 + \frac{R_i^2}{n}q_i(X_i))^2} + \tilde{d}_i\Delta_i(X_i) \ .$$

Now $\mathbf{E}\left(\varphi_i(X_i)\right) = \mathbf{E}\left(\varphi_i(Y_i)\right)$, so to control $\mathbf{E}\left(T_2(X_i) - T_2(Y_i)\right)$, all we need to do is control

$$T_{2,1}(X_i) = \frac{R_i^2}{n}\frac{\tilde{q}_i(X_i) - \tilde{d}_i}{(1 + \frac{R_i^2}{n}q_i(X_i))^2}\varphi_i(X_i)$$

$$T_{2,2}(X_i) = \frac{R_i^2}{n}\tilde{d}_i\Delta_i(X_i)\varphi_i(X_i) \ .$$

Recall that

$$\frac{R_i^2}{n}\left|\tilde{d}_i\Delta_i(X_i)\right| \leq Kb(A; \Sigma_\epsilon)\frac{R_i^2}{n}\frac{|q_i(X_i) - d_i|}{1 + \frac{R_i^2}{n}q_i(X_i)}$$

Hence,

$$\mathbf{E}\left(|T_{2,2}(X_i)|\right) \leq Kb(A; \Sigma_\epsilon)\frac{R_i^2}{n}\mathbf{E}\left(|q_i(X_i) - d_i||\varphi_i(X_i)|\right) \ ,$$

and we have already gotten a bound on $\mathbf{E}\left(|q_i(X_i) - d_i||\varphi_i(X_i)|\right)$, so we get

$$\mathbf{E}\left(|T_{2,2}(X_i)|\right) \leq K\frac{R_i^2}{n}b(A; \Sigma_\epsilon)\sqrt{\frac{x'A^{-1}x}{t}}\frac{\sqrt{b_{Q_2}(2; X_i)}}{t}\left[\frac{1}{\sqrt{n}}|\alpha_i|R_i\sqrt{b_L(2; X_i)} + \frac{R_i^2}{n}\sqrt{\frac{b_L(4; X_i)}{t}}\right] \ .$$

Similarly, using the fact that $\sqrt{\mathbf{E}\left(|\tilde{q}_i(X_i) - \tilde{d}_i|^2\right)} \le \sqrt{b_{Q_2}(2; X_i)}b(A; \Sigma_\epsilon)/t$, we see that

$$\mathbf{E}\left(|T_{2,1}(X_i)|\right) \le K \frac{R_i^2}{n} \frac{\sqrt{b_{Q_2}(2; X_i)}b(A; \Sigma_\epsilon)}{t} \left[\frac{1}{\sqrt{n}}|\alpha_i|R_i\sqrt{b_L(2; X_i)} + \frac{R_i^2}{n}\sqrt{\frac{b_L(4; X_i)}{t}}\right]\sqrt{\frac{x'A^{-1}x}{t}} \ .$$

On the other hand,
$$|\mathbf{E}\left(T_2(X_i)\right)| \le b(A; \Sigma_\epsilon)\mathbf{E}\left(|\varphi_i(X_i)|\right)$$

and we have already seen that

$$\mathbf{E}\left(|\varphi_i(X_i)|\right) \le K\sqrt{\frac{x'A^{-1}x}{t}}\frac{R_i}{\sqrt{n}}\left(|\alpha_i|\sqrt{b_L(2; X_i)} + \frac{R_i}{\sqrt{n}}\frac{\sqrt{b_L(4; X_i)}}{\sqrt{t}}\right) \ .$$

So we conclude that
$$|\mathbf{E}\left(T_2(X_i) - T_2(Y_i)\right)| \le U_i(X_i) + U_i(Y_i)$$

where

$$U_i(X_i) = Kb(A; \Sigma_\epsilon)\sqrt{\frac{x'A^{-1}x}{t}}\left[\frac{1}{\sqrt{n}}|\alpha_i|R_i\sqrt{b_L(2; X_i)} + \frac{R_i^2}{n}\sqrt{\frac{b_L(4; X_i)}{t}}\right]\left[\frac{R_i^2}{n}\frac{\sqrt{b_{Q_2}(2; X_i)}}{t} \wedge 1\right] \ .$$

•**Putting everything together** Since K can be chosen so that $\Psi_i(X_i) = U_i(X_i)$, we conclude that

$$|\mathbf{E}\left(H(\alpha; X) - H(\alpha; Y)\right)| \le 2\sum_{i=1}^{n} U_i(X_i) + U_i(Y_i) \ .$$

$\square$

3.5   Checking the heuristics

In Subsection 2.3, we gave some heuristics to compute an asymptotically deterministic equivalent of forms like $x'(X'D^2X/n + A)^{-1}x$ and $x'(X'D^2X/n + A)^{-1}\Sigma_\epsilon(X'D^2X/n + A)^{-1}x$ in the case where all the X's have the same covariance. We now prove them rigorously.

Of course, the centerpiece of our analysis is the fact that this only need to be done in the Gaussian case. The proof is somewhat involved, since at the level of generality at which we operate, we cannot seem to rely on invariance properties of the Gaussian distribution which were recently systematically exploited in El Karoui (2009b), El Karoui (2009c) and have been a mainstay of multivariate statistics (Anderson (2003), Eaton (2007), Chikuse (2003)). As is often the case, computing the limit (or a deterministic equivalent) of the quantities we are interested in is in fact at least as difficult as showing that the limit does not depend on the particulars of the distributions we consider, or bounding the variance (or higher central moments).

It should be noted that our Lindeberg style results are valid for families when each X_i has a different Σ_i. The limits we are investigating here are for the case (mostly encountered or assumed in practice) where all the X_i's have the same Σ.

Let us begin by clarifying our assumptions and by introducing some notation: We assume throughout this subsection that the rows X_j' of the matrix X are independent Gaussian random vectors with mean 0 and (identical) covariance Σ. Due to the concentration properties of the Gaussian distribution (see e.g Ledoux (2001)), or using the properties of normal and weighted-χ^2 random variables, this implies that for any $r \ge 1$,

$$\mathbf{E}\left(|v'X_j|^r\right) \le K_r\|v\|_2^r\||\Sigma\||_2^{r/2} \tag{22}$$

for any deterministic vector v and

$$\mathbf{E}\left(|X_j'BX_j - \text{trace}\left(\Sigma B\right)|^r\right) \le K_r p^{r/2}\||B\||_2^r\||\Sigma\||_2^r \tag{23}$$

for any deterministic matrix B, where K_r is a numerical constant.

Given a matrix $C \succcurlyeq 0$, put

$$M_C := (A + C),$$

where $A \succcurlyeq t\text{Id}$ is our regularizing matrix as above. Note that $|||M_C^{-1}|||_2 \leq t$. In the special case where $C = \mathcal{S}(j) := \mathcal{S} - \frac{1}{n}R_j^2 X_j X_j'$, we simply write M_j instead of $M_{\mathcal{S}(j)}$. We now recall the classic rank-1 update formula which will again be used repeatedly in this part of the paper.

$$M_{\mathcal{S}}^{-1} = M_j^{-1} - \frac{\frac{1}{n}R_j^2\, M_j^{-1} X_j X_j' M_j^{-1}}{1 + \frac{1}{n}R_j^2\, X_j' M_j^{-1} X_j}\,. \tag{24}$$

Unless otherwise mentioned, B is always a deterministic positive semidefinite matrix in the sequel. For $j = 1, \ldots, n$, let

$$q_j := X_j' M_j^{-1} X_j, \quad d_j := \text{trace}\left(\Sigma M_j^{-1}\right), \quad \tilde{q}_j := X_j' M_j^{-1} B M_j^{-1} X_j, \quad \tilde{d}_j := \text{trace}\left(\Sigma M_j^{-1} B M_j^{-1}\right).$$

In this subsection, we will usually replace q_j and $\tilde{q}_j$ with the fully deterministic quantities $\mathbf{E}(d_j)$ and $\mathbf{E}(\tilde{d}_j)$ (instead of d_j and $\tilde{d}_j$). Using the fact that B, Σ and M_j^{-1} are positive definite, it is easy to see that

$$0 \leq \frac{1}{1 + \frac{1}{n}R_j^2\, q_j} \leq 1, \quad 0 \leq \frac{1}{1 + \frac{1}{n}R_j^2\, \mathbf{E}(d_j)} \leq 1 \tag{25}$$

and

$$0 \leq \frac{\frac{1}{n}R_j^2\, \tilde{q}_j}{1 + \frac{1}{n}R_j^2\, q_j} \leq \frac{|||B|||_2}{t}, \quad 0 \leq \frac{\frac{1}{n}R_j^2\, \mathbf{E}(\tilde{d}_j)}{1 + \frac{1}{n}R_j^2\, \mathbf{E}(d_j)} \leq \frac{|||B|||_2}{t}\,. \tag{26}$$

The following lemma provides some additional estimates which will be used in this subsection:

Lemma 3.3. *Suppose that the above-mentioned assumptions are satisfied.*

(a) We have

$$|trace\left(\Sigma M_{\mathcal{S}}^{-1} - \Sigma M_j^{-1}\right)| \leq |||\Sigma|||_2 t^{-1}$$

and

$$|trace\left(\Sigma M_{\mathcal{S}}^{-1} B M_{\mathcal{S}}^{-1} - \Sigma M_j^{-1} B M_j^{-1}\right)| \leq 2|||B|||_2 |||\Sigma|||_2 t^{-2}\,.$$

(b) For fixed $r \geq 1$, we have

$$\mathbf{E}\left(\left|trace\left(\Sigma M_{\mathcal{S}}^{-1}\right) - \mathbf{E}\left(trace\left(\Sigma M_{\mathcal{S}}^{-1}\right)\right)\right|^r\right) \leq K_r' n^{r/2} |||\Sigma|||_2^r t^{-r}$$

and

$$\mathbf{E}\left(\left|trace\left(\Sigma M_{\mathcal{S}}^{-1} B M_{\mathcal{S}}^{-1}\right) - \mathbf{E}\left(trace\left(\Sigma M_{\mathcal{S}}^{-1} B M_{\mathcal{S}}^{-1}\right)\right)\right|^r\right) \leq K_r'' n^{r/2} |||B|||_2^r |||\Sigma|||_2^r t^{-2r}\,,$$

where K_r' and K_r'' are constants depending only on r.

(c) We have

$$\mathbf{E}\left(\left|\frac{1}{1 + \frac{1}{n}R_j^2\, q_j} - \frac{1}{1 + \frac{1}{n}R_j^2\, \mathbf{E}(d_j)}\right|\right) \leq K'\left(1 \wedge \frac{1}{n}R_j^2(\sqrt{p} + \sqrt{n})|||\Sigma|||_2 t^{-1}\right)$$

and

$$\mathbf{E}\left(\left|\frac{\frac{1}{n}R_j^2\, \tilde{q}_j}{(1 + \frac{1}{n}R_j^2\, q_j)^2} - \frac{\frac{1}{n}R_j^2\, \mathbf{E}(\tilde{d}_j)}{(1 + \frac{1}{n}R_j^2\, \mathbf{E}(d_j))^2}\right|\right) \leq K''|||B|||_2 t^{-1}\left(1 \wedge \frac{1}{n}R_j^2(\sqrt{p} + \sqrt{n})|||\Sigma|||_2 t^{-1}\right)\,,$$

where K' and K'' are numerical constants.

(d) For any square-integrable random variables Z_j such that $\mathbf{E}\,(Z_j)^2 \le L^2$, we have

$$\sum_{j=1}^{n} \tfrac{1}{n} R_j^2\, \mathbf{E}\left(\left|\left(\frac{1}{1+\frac{1}{n}R_j^2\,q_j} - \frac{1}{1+\frac{1}{n}R_j^2\,\mathbf{E}(d_j)}\right) Z_j\right|\right) = \mathrm{O}(LU)$$

and

$$\sum_{j=1}^{n} \tfrac{1}{n} R_j^2\, \mathbf{E}\left(\left|\left(\frac{\frac{1}{n}R_j^2\,\tilde{q}_j}{(1+\frac{1}{n}R_j^2\,q_j)^2} - \frac{\frac{1}{n}R_j^2\,\mathbf{E}(\tilde{d}_j)}{(1+\frac{1}{n}R_j^2\,\mathbf{E}(\tilde{d}_j))^2}\right) Z_j\right|\right) = \mathrm{O}(LU|||B|||_2 t^{-1}),$$

where

$$U := \sum_{j=1}^{n} \left(\tfrac{1}{n}R_j^2 \wedge \tfrac{1}{n^2}R_j^4(\sqrt{p}+\sqrt{n})\right)|||\Sigma|||_2 t^{-1}\ .$$

(e) For any bounded random vectors V_j and W_j independent of X_j such that $\|V_j\|_2 \le L_1$ and $\|W_j\|_2 \le L_2$, we have

$$\sum_{j=1}^{n} \tfrac{1}{n} R_j^2 \mathbf{E}\left(\left|V_j' M_S^{-1} W_j - V_j' M_j^{-1} W_j\right|\right) = \mathrm{O}(L_1 L_2 \widetilde{U})$$

and

$$\sum_{j=1}^{n} \tfrac{1}{n} R_j^2 \mathbf{E}\left(\left|V_j' M_S^{-1} B M_S^{-1} W_j - V_j' M_j^{-1} B M_j^{-1} W_j\right|\right) = \mathrm{O}(L_1 L_2 \widetilde{U}|||B|||_2 t^{-1}),$$

where

$$\widetilde{U} := \sum_{j=1}^{n} \left(\tfrac{1}{n}R_j^2 t^{-1} \wedge \tfrac{1}{n^2}R_j^4 |||\Sigma|||_2 t^{-2}\right)\ .$$

Proof. Throughout this proof, K denotes a numerical constant which may change from step to step.

(a) At least the first inequality is well known in random matrix theory (see e.g. Silverstein and Bai (1995)). We include a proof for the sake of completeness. Using (24), we get

$$\left|\mathrm{trace}\left(\Sigma M_S^{-1} - \Sigma M_j^{-1}\right)\right| = \left|\frac{\frac{1}{n}R_j^2 X_j' M_j^{-1} \Sigma M_j^{-1} X_j}{1+\frac{1}{n}R_j^2 X_j' M_j^{-1} X_j}\right| \le |||M_j^{-1/2}\Sigma M_j^{-1/2}|||_2 \le |||\Sigma|||_2 t^{-1}\ .$$

In fact, this continues to hold for a general square matrix Σ. It therefore follows that

$$\left|\mathrm{trace}\left(\Sigma M_S^{-1} B M_S^{-1} - \Sigma M_j^{-1} B M_j^{-1}\right)\right|$$
$$\le \left|\mathrm{trace}\left(\Sigma(M_S^{-1} - M_j^{-1})B M_S^{-1}\right)\right| + \left|\mathrm{trace}\left(\Sigma M_j^{-1} B(M_S^{-1} - M_j^{-1})\right)\right| \le 2|||B|||_2|||\Sigma|||_2 t^{-2}\ .$$

(b) This is a simple consequence of Azuma's inequality (see e.g. Lemma 4.1 in Ledoux). We follow the proof of Lemma 6 in El Karoui (2009a) : For $j = 0, \ldots, n$, let $\mathcal{F}_j$ denote the σ-field generated by $X_1, \ldots, X_j$. Then, using part (a), we have

$$\left|\mathbf{E}\left(\mathrm{trace}\left(\Sigma M_S^{-1}\right)|\mathcal{F}_j\right) - \mathbf{E}\left(\mathrm{trace}\left(\Sigma M_S^{-1}\right)|\mathcal{F}_{j-1}\right)\right|$$
$$= \left|\mathbf{E}\left(\mathrm{trace}\left(\Sigma(M_S^{-1} - M_j^{-1})\right)|\mathcal{F}_j\right) - \mathbf{E}\left(\mathrm{trace}\left(\Sigma(M_S^{-1} - M_j^{-1})\right)|\mathcal{F}_{j-1}\right)\right| \le 2|||\Sigma|||_2 t^{-1},$$

so, by Azuma's inequality, we get

$$\Pr(|\mathrm{trace}\left(\Sigma M_S^{-1}\right) - \mathbf{E}\left(\mathrm{trace}\left(\Sigma M_S^{-1}\right)\right)| > u) \le 2\exp(-u^2/8n|||\Sigma|||_2^2 t^{-2})$$

for all $u > 0$. Since $\mathbf{E}\,(|Z|^r) = \int_0^\infty r u^{r-1} \Pr(|Z| > u)\,du$ for any real random variable Z, the first inequality follows easily. The second inequality is derived similarly.

(c) Recall that from (25), we have the simple estimate

$$\left|\frac{1}{1+\frac{1}{n}R_j^2\,q_j} - \frac{1}{1+\frac{1}{n}R_j^2\,\mathbf{E}(d_j)}\right| \le 1\ .$$

Using (23) and part (b), we also have the estimate

$$\mathbf{E}\left(\left|\frac{1}{1+\frac{1}{n}R_j^2\,q_j}-\frac{1}{1+\frac{1}{n}R_j^2\,\mathbf{E}(d_j)}\right|\right)\le \tfrac{1}{n}R_j^2\,\mathbf{E}\left(|q_j-\mathbf{E}\,(d_j)|\right)$$

$$\le \tfrac{1}{n}R_j^2\,\mathbf{E}\left(|q_j-d_j|\right)+\tfrac{1}{n}R_j^2\,\mathbf{E}\left(|d_j-\mathbf{E}\,(d_j)|\right)\le K\,\tfrac{1}{n}R_j^2(\sqrt{p}+\sqrt{n})|||\Sigma|||_2 t^{-1}\,.$$

It follows that

$$\mathbf{E}\left(\left|\frac{1}{1+\frac{1}{n}R_j^2\,q_j}-\frac{1}{1+\frac{1}{n}R_j^2\,\mathbf{E}(d_j)}\right|\right)\le K\left(1\wedge \tfrac{1}{n}R_j^2(\sqrt{p}+\sqrt{n})|||\Sigma|||_2 t^{-1}\right)\,,\qquad(27)$$

and the first inequality is proved. For the second inequality, first observe that from (25) and (26), we have the simple estimate

$$\left|\frac{\frac{1}{n}R_j^2\,\tilde{q}_j}{(1+\frac{1}{n}R_j^2\,q_j)^2}-\frac{\frac{1}{n}R_j^2\,\mathbf{E}(\tilde{d}_j)}{(1+\frac{1}{n}R_j^2\,\mathbf{E}(d_j))^2}\right|\le |||B|||_2 t^{-1}\,.$$

Moreover, writing

$$\frac{\frac{1}{n}R_j^2\,\tilde{q}_j}{(1+\frac{1}{n}R_j^2\,q_j)^2}-\frac{\frac{1}{n}R_j^2\,\mathbf{E}(\tilde{d}_j)}{(1+\frac{1}{n}R_j^2\,\mathbf{E}(d_j))^2}=\frac{\frac{1}{n}R_j^2\,(\tilde{q}_j-\mathbf{E}(\tilde{d}_j))}{(1+\frac{1}{n}R_j^2\,q_j)^2\,(1+\frac{1}{n}R_j^2\,\mathbf{E}(d_j))^2}$$

$$+\frac{\frac{1}{n}R_j^2\,\tilde{q}_j}{1+\frac{1}{n}R_j^2\,q_j}\left(\frac{1}{1+\frac{1}{n}R_j^2\,q_j}-\frac{1}{1+\frac{1}{n}R_j^2\,\mathbf{E}(d_j)}\right)+\frac{\frac{1}{n}R_j^2\,\mathbf{E}(\tilde{d}_j)}{1+\frac{1}{n}R_j^2\,\mathbf{E}(d_j)}\left(\frac{1}{1+\frac{1}{n}R_j^2\,q_j}-\frac{1}{1+\frac{1}{n}R_j^2\,\mathbf{E}(d_j)}\right)$$

and using (23) and part (b), (25), (26) as well as (27), we get the estimate

$$\mathbf{E}\left(\left|\frac{\frac{1}{n}R_j^2\,\tilde{q}_j}{(1+\frac{1}{n}R_j^2\,q_j)^2}-\frac{\frac{1}{n}R_j^2\,\mathbf{E}(\tilde{d}_j)}{(1+\frac{1}{n}R_j^2\,\mathbf{E}(d_j))^2}\right|\right)\le \tfrac{1}{n}R_j^2\mathbf{E}\left(\left|\tilde{q}_j-\mathbf{E}(\tilde{d}_j)\right|\right)$$

$$+2|||B|||_2 t^{-1}\mathbf{E}\left(\left|\frac{1}{1+\frac{1}{n}R_j^2\,q_j}-\frac{1}{1+\frac{1}{n}R_j^2\,\mathbf{E}(d_j)}\right|\right)\le K\,\tfrac{1}{n}R_j^2(\sqrt{p}+\sqrt{n})|||B|||_2|||\Sigma|||_2 t^{-2}\,.$$

It follows that

$$\mathbf{E}\left(\left|\frac{\frac{1}{n}R_j^2\,\tilde{q}_j}{(1+\frac{1}{n}R_j^2\,q_j)^2}-\frac{\frac{1}{n}R_j^2\,\mathbf{E}(\tilde{d}_j)}{(1+\frac{1}{n}R_j^2\,\mathbf{E}(d_j))^2}\right|\right)\le K|||B|||_2 t^{-1}\left(1\wedge \tfrac{1}{n}R_j^2(\sqrt{p}+\sqrt{n})|||\Sigma|||_2 t^{-1}\right)\,.\qquad(28)$$

(d) Similar arguments as in part (c) show that

$$\left(\mathbf{E}\left(\left|\frac{1}{1+\frac{1}{n}R_j^2\,q_j}-\frac{1}{1+\frac{1}{n}R_j^2\,\mathbf{E}(d_j)}\right|^2\right)\right)^{1/2}\le K\left(1\wedge \tfrac{1}{n}R_j^2(\sqrt{p}+\sqrt{n})|||\Sigma|||_2 t^{-1}\right)$$

and

$$\left(\mathbf{E}\left(\left|\frac{\frac{1}{n}R_j^2\,\tilde{q}_j}{(1+\frac{1}{n}R_j^2\,q_j)^2}-\frac{\frac{1}{n}R_j^2\,\mathbf{E}(\tilde{d}_j)}{(1+\frac{1}{n}R_j^2\,\mathbf{E}(d_j))^2}\right|^2\right)\right)^{1/2}\le K|||B|||_2 t^{-1}\left(1\wedge \tfrac{1}{n}R_j^2(\sqrt{p}+\sqrt{n})|||\Sigma|||_2 t^{-1}\right)\,.$$

Thus, the claim follows from Cauchy-Schwarz inequality.

(e) On the one hand, we have the simple estimate

$$\left|V_j'(M_S^{-1}-M_j^{-1})W_j\right|\le 2L_1 L_2 t^{-1}\,.$$

On the other hand, using (24), Cauchy-Schwarz inequality and (22), we have the estimate

$$\mathbf{E}\left(\left|V_j'(M_S^{-1} - M_j^{-1})W_j\right|\right) \leq \tfrac{1}{n}R_j^2 \mathbf{E}\left(\left|V_j'M_j^{-1}X_jX_j'M_j^{-1}W_j\right|\right) \leq K\tfrac{1}{n}R_j^2 L_1 L_2 |||\Sigma|||_2 t^{-2}.$$

Combining these estimates, it follows that

$$\sum_{j=1}^{n} \tfrac{1}{n}R_j^2 \mathbf{E}\left(\left|V_j'M_S^{-1}W_j - V_j'M_j^{-1}W_j\right|\right) = \mathrm{O}(L_1 L_2 \widetilde{U}),$$

which establishes the first part of (e). For the second part of (e), write

$$\mathbf{E}\left(\left|V_j'M_S^{-1}BM_S^{-1}W_j - V_j'M_j^{-1}BM_j^{-1}W_j\right|\right)$$
$$\leq \mathbf{E}\left(\left|V_j'(M_S^{-1} - M_j^{-1})BM_j^{-1}W_j\right|\right) + \mathbf{E}\left(\left|V_j'M_j^{-1}B(M_S^{-1} - M_j^{-1})W_j\right|\right)$$
$$+ \mathbf{E}\left(\left|V_j'(M_S^{-1} - M_j^{-1})B(M_S^{-1} - M_j^{-1})W_j\right|\right).$$

By the preceding estimates, the first two expectations are bounded by $K\tfrac{1}{n}R_j^2 L_1 L_2 |||B|||_2 |||\Sigma|||_2 t^{-3}$ here. For the third expectation, we can use (24) and (26) to get

$$|V_j'(M_S^{-1} - M_j^{-1})B(M_S^{-1} - M_j^{-1})W_j| = \frac{\tfrac{1}{n^2}R_j^4}{(1 + \tfrac{1}{n}R_j^2 q_j)^2} |V_j M_j^{-1}X_j X_j' M_j^{-1} B M_j^{-1} X_j X_j' M_j^{-1} W_j|$$
$$\leq \tfrac{1}{n}R_j^2 |V_j M_j^{-1}X_j X_j' M_j^{-1} W_j| |||B|||_2 t^{-1}$$

and therefore, by Cauchy-Schwarz inequality and (22),

$$\mathbf{E}\left(|V_j'(M_S^{-1} - M_j^{-1})B(M_S^{-1} - M_j^{-1})W_j|\right) \leq K\tfrac{1}{n}R_j^2 L_1 L_2 |||B|||_2 |||\Sigma|||_2 t^{-3}.$$

Combining this with the simple estimate

$$\left|V_j'M_S^{-1}BM_S^{-1}W_j - V_j'M_j^{-1}BM_j^{-1}W_j\right| \leq 2L_1 L_2 |||B|||_2 t^{-2},$$

it follows that

$$\sum_{j=1}^{n} \tfrac{1}{n}R_j^2 \mathbf{E}\left(\left|V_j'M_S^{-1}BM_S^{-1}W_j - V_j'M_j^{-1}BM_j^{-1}W_j\right|\right) = \mathrm{O}(L_1 L_2 \widetilde{U}|||B|||_2 t^{-1}).$$

This completes the proof of the lemma. $\qquad\square$

To verify Heuristic 2.1, we will prove the following result:

Proposition 3.1. *Suppose that the assumptions from the beginning of this subsection hold, the ratio p/n stays bounded, $\|v\| = 1$ and*

$$\sum_{j=1}^{n} \frac{R_j^2}{n}|||\Sigma|||_2 = \mathrm{O}(1) \qquad and \qquad \sum_{j=1}^{n} \left(\frac{R_j^2}{n} \wedge \frac{R_j^4}{n^{3/2}}|||\Sigma|||_2\right)|||\Sigma|||_2 = \mathrm{o}(1) \qquad (29)$$

as $n \to \infty$. Then we have

$$\mathbf{E}\left(v'(S+A)^{-1}v\right) - \mathbf{E}\left(v'(\gamma(A)\Sigma + A)^{-1}v\right) \to 0,$$

where

$$\gamma(A) := \frac{1}{n}\sum_{i=1}^{n} \frac{R_i^2}{1 + \tfrac{1}{n}R_i^2 \, \mathrm{trace}\left(\Sigma M_S^{-1}\right)}.$$

48

Proof. We first show that we may replace $\gamma(A)$ with the deterministic quantity

$$\overline{\gamma}(A) := \sum_{i=1}^{n} \frac{\frac{1}{n}R_i^2}{1 + \frac{1}{n}R_i^2 \mathbf{E}\left(\text{trace}\left(\Sigma M_i^{-1}\right)\right)}.$$

To this end, since $|v'(A + \gamma(A)\Sigma)^{-1}v - v'(A + \overline{\gamma}(A)\Sigma)^{-1}v| \leq t^{-2}|\gamma(A) - \overline{\gamma}(A)|\||\Sigma\||_2$, it suffices to show that

$$\mathbf{E}\left(|\gamma(A) - \overline{\gamma}(A)|\right)\||\Sigma\||_2 = o(1). \tag{30}$$

But now,

$$\mathbf{E}\left(|\gamma(A) - \overline{\gamma}(A)|\right) \leq \sum_{i=1}^{n} \mathbf{E}\left(\left|\frac{\frac{1}{n}R_i^2}{1 + \frac{1}{n}R_i^2 \text{trace}\left(\Sigma M_S^{-1}\right)} - \frac{\frac{1}{n}R_i^2}{1 + \frac{1}{n}R_i^2 \mathbf{E}\left(\text{trace}\left(\Sigma M_S^{-1}\right)\right)}\right|\right)$$

$$+ \sum_{i=1}^{n} \mathbf{E}\left(\left|\frac{\frac{1}{n}R_i^2}{1 + \frac{1}{n}R_i^2 \mathbf{E}\left(\text{trace}\left(\Sigma M_S^{-1}\right)\right)} - \frac{\frac{1}{n}R_i^2}{1 + \frac{1}{n}R_i^2 \mathbf{E}\left(\text{trace}\left(\Sigma M_i^{-1}\right)\right)}\right|\right)$$

$$\leq K\left(\sum_{j=1}^{n}\left(\frac{1}{n}R_j^2 \wedge \frac{1}{n^{3/2}}R_j^4\||\Sigma\||_2 t^{-1}\right) + \sum_{j=1}^{n}\left(\frac{1}{n}R_j^2 \wedge \frac{1}{n^2}R_j^4\||\Sigma\||_2 t^{-1}\right)\right),$$

where the second step follows from similar arguments as in the proof of Lemma 3.3 (c) (using Lemma 3.3 (b) and (a)). Thus, (30) follows from Assumption (29).

We now proceed similarly as in Silverstein (1995) and El Karoui (2009a). Using (24), it is easy to check that $M_S^{-1}X_j = (1 + \frac{1}{n}R_j^2 q_j)^{-1}M_j^{-1}X_j$. Thus, setting $T := \overline{\gamma}(A)\Sigma$, so that $M_T = A + \overline{\gamma}(A)\Sigma$, we get

$$M_S^{-1} - M_T^{-1} = -M_S^{-1}(S - T)M_T^{-1} = -\sum_{i=1}^{n}\frac{1}{n}R_i^2 \, M_S^{-1}X_iX_i'M_T^{-1} + M_S^{-1}TM_T^{-1}$$

$$= -\sum_{i=1}^{n}\frac{\frac{1}{n}R_i^2 \, M_i^{-1}X_iX_i'M_T^{-1}}{1 + \frac{1}{n}R_i^2 \, q_i} + M_S^{-1}TM_T^{-1}$$

$$= -\sum_{i=1}^{n}\left(\frac{\frac{1}{n}R_i^2 \, M_i^{-1}X_iX_i'M_T^{-1}}{1 + \frac{1}{n}R_i^2 \, q_i} - \frac{\frac{1}{n}R_i^2 \, M_S^{-1}\Sigma M_T^{-1}}{1 + \frac{1}{n}R_i^2 \, \mathbf{E}(d_i)}\right)$$

and therefore

$$\mathbf{E}\left(v'M_S^{-1}v\right) - \mathbf{E}\left(v'M_T^{-1}v\right) = -\sum_{i=1}^{n}\mathbf{E}\left(\frac{\frac{1}{n}R_i^2 \, M_i^{-1}X_iX_i'M_T^{-1}}{1 + \frac{1}{n}R_i^2 \, q_i} - \frac{\frac{1}{n}R_i^2 \, M_S^{-1}\Sigma M_T^{-1}}{1 + \frac{1}{n}R_i^2 \, \mathbf{E}(d_i)}\right). \tag{31}$$

Now, using Lemma 3.3 (d), the independence of X_j and $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n$, Lemma 3.3 (e), and Assumption (29), it follows that

$$\sum_{i=1}^{n}\mathbf{E}\left(\frac{\frac{1}{n}R_i^2 \, v'M_i^{-1}X_iX_i'M_T^{-1}v}{1 + \frac{1}{n}R_i^2 \, q_i}\right) = \sum_{i=1}^{n}\mathbf{E}\left(\frac{\frac{1}{n}R_i^2 \, v'M_i^{-1}X_iX_i'M_T^{-1}v}{1 + \frac{1}{n}R_i^2 \, \mathbf{E}(d_i)}\right) + o(1)$$

$$= \sum_{i=1}^{n}\mathbf{E}\left(\frac{\frac{1}{n}R_i^2 \, v'M_i^{-1}\Sigma M_T^{-1}v}{1 + \frac{1}{n}R_i^2 \, \mathbf{E}(d_i)}\right) + o(1)$$

$$= \sum_{i=1}^{n}\mathbf{E}\left(\frac{\frac{1}{n}R_i^2 \, v'M_S^{-1}\Sigma M_T^{-1}v}{1 + \frac{1}{n}R_i^2 \, \mathbf{E}(d_i)}\right) + o(1).$$

This completes the proof of Proposition 3.1. $\qquad\square$

To verify Heuristic 2.2, we will prove the following result:

Proposition 3.2. *Suppose that the assumptions from the beginning of this subsection hold, the ratio p/n stays bounded, $\|v\| = 1$ and*

$$\sum_{j=1}^{n} \frac{R_j^2}{n} |||\Sigma|||_2 = O(1) \qquad and \qquad \sum_{j=1}^{n} \left(\frac{R_j^2}{n} \wedge \frac{R_j^4}{n^{3/2}} |||\Sigma|||_2 \right) |||B|||_2 |||\Sigma|||_2 = o(1) \qquad (32)$$

as $n \to \infty$. Then we have

$$\mathbf{E}\left(v'(S+A)^{-1}B(S+A)^{-1}v \right) - \mathbf{E}\left(v'(A+\gamma(A)\Sigma)^{-1}(B+\xi(A,B)\Sigma)(A+\gamma(A)\Sigma)^{-1}v \right) \to 0\,,$$

where $\gamma(A)$ is defined in Proposition 3.1 and

$$\xi(A,B) := \left[\frac{1}{n} \sum_{i=1}^{n} \frac{R_i^4}{\left(1 + \frac{1}{n}R_i^2 \operatorname{trace}\left(\Sigma M_S^{-1}\right)\right)^2} \right] \frac{1}{n} \operatorname{trace}\left(\Sigma(S+A)^{-1}B(S+A)^{-1} \right)\,.$$

Proof. Similarly as in the proof of Proposition 3.1, we first show that we may replace $\gamma(A)$ and $\xi(A,B)$ with the deterministic quantities $\overline{\gamma}(A)$ and $\overline{\xi}(A,B)$, where $\overline{\gamma}(A)$ is defined in the proof of Proposition 3.1 and

$$\overline{\xi}(A,B) := \sum_{i=1}^{n} \frac{\frac{1}{n^2} R_i^4}{\left(1 + \frac{1}{n}R_i^2 \mathbf{E}\left(\operatorname{trace}\left(\Sigma M_i^{-1}\right)\right)\right)^2} \mathbf{E}\left(\operatorname{trace}\left(\Sigma M_i^{-1} B M_i^{-1}\right) \right)\,.$$

To begin with, similarly as in (26), $\xi(A,B)$ and $\overline{\xi}(A,B)$ are bounded by $\sum_{j=1}^{n} \frac{1}{n} R_j^2 |||B|||_2 t^{-1}$. It therefore follows from Assumption (32) that $\xi(A,B)\Sigma$ and $\overline{\xi}(A,B)\Sigma$ are bounded (in operator norm) by $K|||B|||_2$, where K is a constant. Thus, using the decomposition

$$(A_1 A_2 A_3 - B_1 B_2 B_3) = (A_1 - B_1)B_2 B_3 + A_1(A_2 - B_2)B_3 + A_1 A_2(A_3 - B_3)\,,$$

we see that it suffices to check that

$$\mathbf{E}\left(|\gamma(A) - \overline{\gamma}(A)| \right) |||B|||_2 |||\Sigma|||_2 = o(1) \quad and \quad \mathbf{E}\left(|\xi(A,B) - \overline{\xi}(A,B)| \right) |||\Sigma|||_2 = o(1)\,. \qquad (33)$$

The former bound is clear from (30), since we are assuming (32) instead of (29) now. For the latter bound, let us note that

$$\begin{aligned}
&\mathbf{E}\left(|\xi(A,B) - \overline{\xi}(A,B)| \right) \\
&\leq \sum_{i=1}^{n} \mathbf{E}\left(\left| \frac{\frac{1}{n^2} R_i^4 \operatorname{trace}\left(\Sigma M_S^{-1} B M_S^{-1}\right)}{\left(1 + \frac{1}{n}R_i^2 \operatorname{trace}\left(\Sigma M_S^{-1}\right)\right)^2} - \frac{\frac{1}{n^2} R_i^4 \mathbf{E}\left(\operatorname{trace}\left(\Sigma M_S^{-1} B M_S^{-1}\right)\right)}{\left(1 + \frac{1}{n}R_i^2 \mathbf{E}\left(\operatorname{trace}\left(\Sigma M_S^{-1}\right)\right)\right)^2} \right| \right) \\
&\quad + \sum_{i=1}^{n} \mathbf{E}\left(\left| \frac{\frac{1}{n^2} R_i^4 \mathbf{E}\left(\operatorname{trace}\left(\Sigma M_S^{-1} B M_S^{-1}\right)\right)}{\left(1 + \frac{1}{n}R_i^2 \mathbf{E}\left(\operatorname{trace}\left(\Sigma M_S^{-1}\right)\right)\right)^2} - \frac{\frac{1}{n^2} R_i^4 \mathbf{E}\left(\operatorname{trace}\left(\Sigma M_i^{-1} B M_i^{-1}\right)\right)}{\left(1 + \frac{1}{n}R_i^2 \mathbf{E}\left(\operatorname{trace}\left(\Sigma M_i^{-1}\right)\right)\right)^2} \right| \right) \\
&\leq K |||B|||_2 t^{-1} \left(\sum_{j=1}^{n} \left(\frac{1}{n}R_j^2 \wedge \frac{1}{n^{3/2}} R_j^4 |||\Sigma|||_2 t^{-1} \right) + \sum_{j=1}^{n} \left(\frac{1}{n}R_j^2 \wedge \frac{1}{n^2} R_j^4 |||\Sigma|||_2 t^{-1} \right) \right)\,,
\end{aligned}$$

where the second step follows from similar arguments as in the proof of Lemma 3.3 (c) (using Lemma 3.3 (b) and (a)). In view of Assumption (32), this establishes (33).

Put $T := \overline{\gamma}(A)\Sigma$, $T(u) := \overline{\gamma}(A+uB)\Sigma$ $(u > 0)$ and observe that

$$\frac{d}{du}(S+A+uB)^{-1}\bigg|_{u=0} = -(S+A)^{-1}B(S+A)^{-1}$$

and

$$\frac{d}{du}(T(u)+A+uB)^{-1}\bigg|_{u=0} = -(T+A)^{-1}\left(B + \overline{\xi}(A,B)\Sigma \right)(T+A)^{-1}\,.$$

Thus, replacing A with $A + uB$ in (31) and calculating the derivative with respect to u at $u = 0$, we get

$$-\mathbf{E}(v'M_S^{-1}BM_S^{-1}v) + \mathbf{E}(v'M_T^{-1}(B + \bar{\xi}(A,B)\Sigma)M_T^{-1}v) = D_1 + D_2 + D_3\,, \qquad (34)$$

where

$$D_1 = -\sum_{i=1}^n \mathbf{E}\left(\frac{\frac{1}{n^2}R_i^4\,\tilde{q}_i}{(1 + \frac{1}{n}R_i^2\,q_i)^2}\,v'M_i^{-1}X_iX_i'M_T^{-1}v - \frac{\frac{1}{n}R_i^4\,\mathbf{E}(\tilde{d}_i)}{(1 + \frac{1}{n}R_i^2\,\mathbf{E}(d_i))^2}\,v'M_S^{-1}\Sigma M_T^{-1}v\right),$$

$$D_2 = +\sum_{i=1}^n \mathbf{E}\left(\frac{\frac{1}{n}R_i^2}{1 + \frac{1}{n}R_i^2\,q_i}\,v'M_i^{-1}BM_i^{-1}X_iX_i'M_T^{-1}v - \frac{\frac{1}{n}R_i^2}{1 + \frac{1}{n}R_i^2\,\mathbf{E}(d_i)}\,v'M_S^{-1}BM_S^{-1}\Sigma M_T^{-1}v\right),$$

$$D_3 = +\sum_{i=1}^n \mathbf{E}\left(\frac{\frac{1}{n}R_i^2}{1 + \frac{1}{n}R_i^2\,q_i}\,v'M_i^{-1}X_iX_i'M_T^{-1}(B + \bar{\xi}(A,B)\Sigma)M_T^{-1}v\right.$$

$$\left. - \frac{\frac{1}{n}R_i^2}{1 + \frac{1}{n}R_i^2\,\mathbf{E}(d_i)}\,v'M_S^{-1}\Sigma M_T^{-1}(B + \bar{\xi}(A,B)\Sigma)M_T^{-1}v\right).$$

Similarly as in the proof of Proposition 3.1, the idea is to show that for $i = 1, 2, 3$, $D_i \to 0$ as $n \to \infty$. Making appropriate use of Lemma 3.3, this follows by essentially the same calculation as in the proof of Proposition 3.1, so that we go fast over the details.

For the first difference, we use Lemma 3.3 (d), the independence of X_j and $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n$, and Lemma 3.3 (e) to obtain

$$\sum_{i=1}^n \mathbf{E}\left(\frac{\frac{1}{n^2}R_i^4\,\tilde{q}_i}{(1 + \frac{1}{n}R_i^2\,q_i)^2}\,v'M_i^{-1}X_iX_i'M_T^{-1}v\right)$$

$$= \sum_{i=1}^n \mathbf{E}\left(\frac{\frac{1}{n^2}R_i^4\,\mathbf{E}(\tilde{d}_i)}{(1 + \frac{1}{n}R_i^2\,\mathbf{E}(d_i))^2}\,v'M_i^{-1}X_iX_i'M_T^{-1}v\right) + o(1)$$

$$= \sum_{i=1}^n \mathbf{E}\left(\frac{\frac{1}{n^2}R_i^4\,\mathbf{E}(\tilde{d}_i)}{(1 + \frac{1}{n}R_i^2\,\mathbf{E}(d_i))^2}\,v'M_i^{-1}\Sigma M_T^{-1}v\right) + o(1)$$

$$= \sum_{i=1}^n \mathbf{E}\left(\frac{\frac{1}{n^2}R_i^4\,\mathbf{E}(\tilde{d}_i)}{(1 + \frac{1}{n}R_i^2\,\mathbf{E}(d_i))^2}\,v'M_S^{-1}\Sigma M_T^{-1}v\right) + o(1)\,.$$

(In the final step, we have also used (25) and (26) to see that the fraction is bounded by $\frac{1}{n}R_i^2\|\|B\|\|_2 t^{-1}$.) For the second and third difference, it follows by similar arguments that

$$\sum_{i=1}^n \mathbf{E}\left(\frac{\frac{1}{n}R_i^2}{1 + \frac{1}{n}R_i^2\,q_i}\,v'M_i^{-1}BM_i^{-1}X_iX_i'M_T^{-1}v\right)$$

$$= \sum_{i=1}^n \mathbf{E}\left(\frac{\frac{1}{n}R_i^2}{1 + \frac{1}{n}R_i^2\,\mathbf{E}(d_i)}\,v'M_i^{-1}BM_i^{-1}X_iX_i'M_T^{-1}v\right) + o(1)$$

$$= \sum_{i=1}^n \mathbf{E}\left(\frac{\frac{1}{n}R_i^2}{1 + \frac{1}{n}R_i^2\,\mathbf{E}(d_i)}\,v'M_i^{-1}BM_i^{-1}\Sigma M_T^{-1}v\right) + o(1)$$

$$= \sum_{i=1}^n \mathbf{E}\left(\frac{\frac{1}{n}R_i^2}{1 + \frac{1}{n}R_i^2\,\mathbf{E}(d_i)}\,v'M_S^{-1}BM_S^{-1}\Sigma M_T^{-1}v\right) + o(1)$$

and

$$\sum_{i=1}^{n} \mathbf{E} \left(\frac{\frac{1}{n}R_i^2}{1 + \frac{1}{n}R_i^2 \, q_i} \, v' M_i^{-1} X_i X_i' M_T^{-1} (B + \bar{\xi}(A,B)\Sigma) M_T^{-1} v \right)$$

$$= \sum_{i=1}^{n} \mathbf{E} \left(\frac{\frac{1}{n}R_i^2}{1 + \frac{1}{n}R_i^2 \, \mathbf{E}(d_i)} \, v' M_i^{-1} X_i X_i' M_T^{-1} (B + \bar{\xi}(A,B)\Sigma) M_T^{-1} v \right) + \mathrm{o}(1)$$

$$= \sum_{i=1}^{n} \mathbf{E} \left(\frac{\frac{1}{n}R_i^2}{1 + \frac{1}{n}R_i^2 \, \mathbf{E}(d_i)} \, v' M_i^{-1} \Sigma M_T^{-1} (B + \bar{\xi}(A,B)\Sigma) M_T^{-1} v \right) + \mathrm{o}(1)$$

$$= \sum_{i=1}^{n} \mathbf{E} \left(\frac{\frac{1}{n}R_i^2}{1 + \frac{1}{n}R_i^2 \, \mathbf{E}(d_i)} \, v' M_S^{-1} \Sigma M_T^{-1} (B + \bar{\xi}(A,B)\Sigma) M_T^{-1} v \right) + \mathrm{o}(1).$$

This concludes the proof. $\qquad\square$

3.6 On the Rate of Convergence

Suppose that the constants $b_L(4; X_i)$ and $b_{Q_2}(2; X_i)/n$ from (3) and (4) are uniformly bounded. Then our results show that if the R_i are also uniformly bounded, we have, for instance,

$$\mathbf{E} \left(|g(\alpha; X) - \mathbf{E}(g(\alpha; X))|^2 \right) = \mathrm{O}(n^{-1})$$

and

$$|\mathbf{E}(g(\alpha; X) - g(\alpha; Y))| = \mathrm{O}(n^{-1/2}).$$

More generally, this still holds if the R_i are given by i.i.d. random variables with finite 4th moments. For some applications (e.g. to the field of finance), it may be helpful to have the same results under the weaker assumption that the R_i are given by i.i.d. random variables with finite second moments only. Recall that this is the minimal reasonable assumption, for if the R_i do not have second moments, the covariance matrix of the vectors $R_i X_i$ is not defined. In this section we sketch how to derive results under this minimal assumption.

However, to derive our results, we need somewhat stronger conditions on the covariance matrices Σ_i, the regularizing matrix A and the distributions of the random variables X_i. More precisely, we will work under the following *additional* assumptions:

- We have $p/n \geq c_\varrho$ for some $c_\varrho > 0$.
- We have $\frac{1}{n} \sum_{i=1}^{n} R_i^2 \leq C_R$ for some $C_R < \infty$.
- We have $\frac{1}{p} \mathrm{trace}(A) \leq C_A$ for some $C_A < \infty$.
- We have $\frac{1}{p} \mathrm{trace}(\Sigma)_i \leq C_\Sigma$ for some $C_\Sigma < \infty$.
- There exist $c_\Sigma > 0$ and $\varepsilon \in (0,1)$ such that the number of eigenvalues of Σ_i which are less than $\leq c_\Sigma$ is less than $p\varepsilon$.
- We have $|||\Sigma_i|||_2 \leq C_\Sigma$, and the constants $b_L(8; X_i)$ and $b_{Q_2}(4; X_i)/n^2$ are uniformly bounded.

Let us mention that the very last assumption could be weakened in exchange for a worse rate of convergence in the following results. For instance, we could easily allow for a bound of the order $\mathrm{O}(\log n)$ or $\mathrm{O}(n^\delta)$ (with $\delta > 0$ sufficiently small).

As it is our main intention here to give an idea of what is possible, we concentrate on one particular case and present results for the quadratic form $g(\alpha) := n^{-1} \alpha' D X (X' D^2 X/n + A)^{-1} X' D \alpha$ (from Section 3.3.3) only.

Our results rely on the observation that (i) normalized traces of regularized inverses of random matrices are typically strongly concentrated and (ii) under the assumptions stated above, $\mathbf{E}(\mathrm{trace}(\Sigma_i M_i^{-1}))$ is of the order n. Let us provide precise formulations:

The first observation has already been used several times in this paper (see also El Karoui (2009a)), for instance in the proof of Lemma 3.3 (b), where it is stated that

$$\mathbf{P}\left(\left|\tfrac{1}{p}\mathrm{trace}\left(\Sigma M^{-1}\right) - \mathbf{E}\left(\tfrac{1}{p}\mathrm{trace}\left(\Sigma M^{-1}\right)\right)\right| \geq u\right) \leq 2\exp(-u^2 p^2 t^2/8n|||\Sigma|||_2^2)\,.$$

for any $u > 0$. For the second observation, we show the following lemma.

Lemma 3.4. *Under the afore-mentioned assumptions, we have*

$$\mathbf{E}\left(\tfrac{1}{n}\mathit{trace}\left(\Sigma_i M_i^{-1}\right)\right) \geq c\,,$$

where $c = c(c_\varrho, C_R, C_A, C_\Sigma, c_\Sigma, \varepsilon)$.

In the following proof, if A, B are any matrices and $x \in \mathbb{R}_+$, we call B (by slight abuse of terminology) a rank x modification of A if $\mathrm{rank}(A - B) \leq x$, and if M is any symmetric matrix, we let $\lambda_1(M), \ldots, \lambda_p(M)$ be the eigenvalues M.

Proof. We have

$$\mathbf{E}\left(\mathrm{trace}\left(M_i\right)\right) = \mathbf{E}\left(\mathrm{trace}\left(A + \sum_{j \neq i}\tfrac{1}{n}R_j^2 X_j X_j'\right)\right) = \mathrm{trace}\left(A\right) + \sum_{j \neq i}\tfrac{1}{n}R_j^2\mathrm{trace}\left(\Sigma_j\right) \leq p(C_A + C_R C_\Sigma)\,.$$

Thus, by Markov's inequality, it follows that with $C := 4(C_A + C_R C_\Sigma)/(1 - \varepsilon)$, we have

$$\mathbf{P}\left(\tfrac{1}{p}\sum_{j=1}^{p}\mathbf{1}_{\{\lambda_j(M_i) \geq C\}} \geq \tfrac{1-\varepsilon}{2}\right) \leq \mathbf{P}\left(\tfrac{1}{p}\mathrm{trace}\left(M_i\right) \geq 2(C_A + C_R C_\Sigma)\right) \leq \frac{\mathbf{E}\left(\mathrm{trace}\left(M_i\right)\right)}{2p(C_A + C_R C_\Sigma)} \leq \tfrac{1}{2}\,.$$

Consider the set G where $\tfrac{1}{p}\sum_{j=1}^{p}\mathbf{1}_{\{\lambda_j(M_i) \geq C\}} \leq \tfrac{1-\varepsilon}{2}$, so $\mathbf{P}\left(G\right) \geq \tfrac{1}{2}$. Then, by spectral calculus, there exists a positive-definite rank $p(\tfrac{1-\varepsilon}{2})$ modification $\tilde{M}_i$ of M_i such that $\lambda_j(\tilde{M}_i) \leq C$ for all $j = 1, \ldots, p$. Similarly, there exists a positive-definite rank $p\varepsilon$ modification $\tilde{\Sigma}_i$ of Σ_i such that $\lambda_j(\tilde{\Sigma}_i) \geq c_\Sigma$ for all $j = 1, \ldots, p$. It follows that $\lambda_j(\tilde{\Sigma}_i^{1/2}\tilde{M}_i^{-1}\tilde{\Sigma}_i^{1/2}) \geq c_\Sigma/C$ for all $j = 1, \ldots, p$. Indeed, for any vector x of norm 1,

$$x'\tilde{\Sigma}_i^{1/2}\tilde{M}_i^{-1}\tilde{\Sigma}_i^{1/2}x \geq x'\tilde{\Sigma}_i x/C \geq c_\Sigma/C\,,$$

and thus $\tilde{\Sigma}_i^{1/2}\tilde{M}_i^{-1}\tilde{\Sigma}_i^{1/2} \succcurlyeq c_\Sigma/C$. By Theorem A.43 in Bai and Silverstein (2010), it further follows that at least $p(\tfrac{1-\varepsilon}{2})$ eigenvalues of $\Sigma_i M_i^{-1}$ are $\geq c_\Sigma/C$. Thus, as $\Sigma_i^{1/2}M_i^{-1}\Sigma_i^{1/2} \succcurlyeq 0$, we have shown that on the set G,

$$\tfrac{1}{p}\mathrm{trace}\left(\Sigma_i M_i^{-1}\right) \geq \tfrac{1-\varepsilon}{2}c_\Sigma/C\,,$$

Since $\mathbf{P}\left(G\right) \geq \tfrac{1}{2}$ and $\Sigma_i^{1/2}M_i^{-1}\Sigma_i^{1/2} \succcurlyeq 0$, we may conclude that

$$\tfrac{1}{n}\mathbf{E}\left(\mathrm{trace}\left(\Sigma_i M_i^{-1}\right)\right) \geq \tfrac{1-\varepsilon}{4}c_\varrho c_\Sigma/C =: c\,.$$

$\square$

Corollary 3.1. *We have* $\mathbf{P}\left(\tfrac{1}{n}\mathit{trace}\left(\Sigma_i M_i^{-1}\right) \leq \tfrac{1}{2}c\right) \leq C_0\exp(-c_0 n)$.

Let us now investigate the implications of these observations for our results concerning $g(\alpha, X)$: Recall from the proof of Theorem 3.4 that (with the notation there)

$$\mathbf{E}\left(|g(\alpha; X) - \mathbf{E}\left(g(\alpha; X)\right)|^2\right) \leq \sum_{j=1}^{n}\mathbf{E}\left(|T - T_i|^2\right)\,,$$

where

$$T - T_i = \alpha_i^2\frac{R_i^2}{n}\frac{q_i(X_i) - d_i}{(1 + \frac{R_i^2}{n}q_i)(1 + \frac{R_i^2}{n}d_i)} + \frac{1}{1 + \frac{R_i^2}{n}d_i}(R_i^2/n\zeta_i^2 - 2\alpha_i R_i/\sqrt{n}\zeta_i)$$

$$+ \frac{\frac{1}{n}R_i^2(d_i - q_i(X_i))}{(1 + \frac{R_i^2}{n}q_i(X_i))(1 + \frac{R_i^2}{n}d_i)}(R_i^2/n\zeta_i^2 - 2\alpha_i R_i/\sqrt{n}\zeta_i)\,.$$

Now, on the set $G_i := \{\frac{1}{n}d_i \geq \frac{1}{2}c\}$ (which has probability $1 - o(1)$ by the preceding corollary),

$$|T - T_i| \leq \frac{2}{cn}\alpha_i^2|q_i(X_i) - d_i| + \left(\frac{2}{c} + 1\right)\left(1/n\zeta_i^2 + 2\alpha_i/\sqrt{n}|\zeta_i|\right) + \frac{2}{cn}|d_i - q_i(X_i)|\left(R_i^2/n\zeta_i^2 + 2\alpha_i R_i/\sqrt{n}|\zeta_i|\right).$$

Using (3) and (4) as well as Cauchy-Schwarz inequality, it follows that

$$\mathbf{E}_i\left(|T - T_i|^2 \mathbf{1}_{G_i}\right) \leq K(c)\left(\alpha_i^4 \frac{1}{n^2}b_{Q_2}(2; X_i)\frac{1}{t^2} + \frac{1}{n^2}\frac{b_L(4; X_i)}{t^2} + \alpha_i^2 \frac{1}{n}\frac{b_L(2; X_i)}{t}\right.$$
$$\left. + \frac{1}{n^2}\frac{R_i^4}{n^2}\frac{\sqrt{b_{Q_2}(4; X_i)}}{t^2}\frac{\sqrt{b_L(8; X_i)}}{t^2} + \frac{1}{n^2}\alpha_i^2\frac{R_i^2}{n}\frac{\sqrt{b_{Q_2}(4; X_i)}}{t^2}\frac{\sqrt{b_L(4; X_i)}}{t}\right),$$

where $K(c)$ denotes a numerical constant which depends on c. Since the right-hand side is deterministic, the same bound holds for the unconditional expectation.

On the complementary set G_i^C, we can use the fact that $|T - T_i| \leq 1 + \alpha_i^2$ to obtain

$$\mathbf{E}\left(|T - T_i|^2 \mathbf{1}_{G_i^C}\right) \leq K\mathbf{P}\left(G_i^C\right) \leq KC_0\exp(-c_0 n).$$

Summing over $i = 1, \ldots, n$ and recalling our assumptions, we conclude that

$$\sum_{i=1}^{n}\mathbf{E}\left(g(\alpha; X) - \mathbf{E}\left(g(\alpha; X)\right)\right)^2 = \mathrm{O}(n^{-1}).$$

Similar considerations can be made for the Lindeberg approach, with the result that

$$|\mathbf{E}\left(g(\alpha; X) - g(\alpha; Y)\right)| = \mathrm{O}(n^{-1/2}).$$

4 Relevance to statistical problems

As discussed in the introduction, many quantities of statistical interest can be analyzed using our results. We will find deterministic equivalents for them. To keep the presentation readable for readers interested more in the applications than in the theory, we do not repeat the assumptions of our theorems. So all our statements should be understood as being prefaced: "assuming that the technical conditions led our earlier in the paper are satisfied, we have...".

What the reader should essentially know is that shrinking the sample covariance matrix to a deterministic matrix A has the effect of essentially shrinking a scaled version of the population covariance to the same matrix A. The damping factor depends on A and Σ and is estimable. When the mean is also estimated, the results of Subsection 3.3.2 need to be applied.

Our results show the remarkable robustness of random matrix results - we need very little control over the particulars of the data distributions - though they highlight their sensitivity to geometric assumptions. We now give a few examples where these computations are relevant and shed light on statistical matters.

4.1 Estimation issues

4.1.1 Estimation of $v'(\Sigma + A)^{-1}v$ when Σ is not observed directly

The motivation for this kind of question comes from understanding the population behavior of certain statistical procedures from observed data and hence deriving benchmarks as to how well a procedure could do. This could be used in evaluating a kind of regret, directly from the data.

Recall our general setting, namely we observe

$$\mathfrak{X}_i = \mu + R_i X_i,$$

where R_i are possibly random and X_i are random with distributions satisfying "our usual assumptions" (see Subsection 3.2). In particular, X_i's have mean 0. Recall also the notation $\mathcal{S} = \frac{1}{n}\sum_{i=1}^{n}R_i^2 X_i X_i'$ and $\widetilde{\mu} = \frac{1}{n}\sum_{i=1}^{n}R_i X_i$.

We have shown that we can find a deterministic equivalent to $v'(\mathcal{S} + A)^{-1}v$, namely,

$$v'(\mathcal{S} + A)^{-1}v \simeq v'(\gamma(A)\Sigma + A)^{-1}v \ .$$

We first note that since $\widehat{\Sigma} = \mathcal{S} - \widetilde{\mu}\widetilde{\mu}'$,

$$v'(\widehat{\Sigma} + A)^{-1}v = v'(\mathcal{S} + A)^{-1}v + \frac{(v'(\mathcal{S} + A)^{-1}\widetilde{\mu})^2}{1 - \widetilde{\mu}'(\mathcal{S} + A)^{-1}\widetilde{\mu}} \simeq v'(\mathcal{S} + A)^{-1}v \simeq v'(\gamma(A)\Sigma + A)^{-1}v \ ,$$

as we have seen that $v'(\mathcal{S} + A)^{-1}\widetilde{\mu} \simeq 0$.

Now in certain situation, for instance to when we want to estimate the optimal risk of certain statistical procedures, we will need to estimate $v'(\Sigma + A)^{-1}v$. We now sketch how to come up with an estimator of this quantity.

Let t be a real in $\mathbb{R}_+$. We clearly have

$$v'(\widehat{\Sigma} + tA)^{-1}v \simeq v'(\gamma(tA)\Sigma + tA)^{-1}v = \frac{v'(\Sigma + tA/\gamma(tA))^{-1}v}{\gamma(tA)} \ .$$

Now recall that

$$\gamma(A) = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i^2}{1 + \frac{R_i^2}{n}\operatorname{trace}\left(\Sigma(\mathcal{S} + A)^{-1}\right)} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i^2}{1 + R_i^2 \alpha(A)} \ ,$$

and under our assumptions, $\gamma(A)$ has an asymptotically deterministic equivalent. Note that under concentration assumptions on X_i's,

$$\frac{X_i'(\mathcal{S}_i + A)^{-1}X_i}{n} \simeq \frac{\operatorname{trace}\left(\Sigma(\mathcal{S} + A)^{-1}\right)}{n} \ ,$$

and using rank-1 update,

$$\frac{R_i^2}{n} X_i'(\mathcal{S}_i + A)^{-1}X_i = \frac{R_i^2}{n} \frac{X_i'(\mathcal{S} + A)^{-1}X_i}{1 - \frac{R_i^2}{n}X_i'(\mathcal{S} + A)^{-1}X_i} \ ,$$

so we need only invert $(\mathcal{S} + A)^{-1}$ once to compute efficiently all the terms we are interested in. (Of course in practice, we do not have access to R_iX_i, so we will use $Y_i - \widehat{\mu} = R_iX_i - \widetilde{\mu}$. Because $\widetilde{\mu}'(\mathcal{S} + A)^{-1}\widetilde{\mu}$ is of order 1 and we will be dividing everything by n, we can neglect this term in this discussion. The same applies to terms of the form $\widetilde{\mu}'(\mathcal{S} + A - \widetilde{\mu}\widetilde{\mu}')^{-1}X_i)$.

So we can now estimate $R_i^2 \alpha(A)$, and using the fact that

$$\gamma(A) = \frac{1}{\alpha(A)}\left(1 - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + R_i^2\alpha(A)}\right) \ ,$$

we can also estimate $\gamma(A)$.

So to estimate $v'(\Sigma + A)^{-1}v$, all we need to do is find t such that

$$\frac{\gamma(t_0 A)}{t_0} = 1 \ .$$

We will now show that $\gamma(tA)/t$ is decreasing; hence a simple dichotomous search will yield a fast algorithm for finding this t_0.

We note that

$$\frac{\gamma(tA)}{t} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i^2}{t + \frac{R_i^2}{n}\operatorname{trace}\left(\Sigma(\mathcal{S}/t + A)^{-1}\right)} \ .$$

Now $(\mathcal{S}/t + A)^{-1}$ is clearly increasing in the Loewner order, and hence so is $\operatorname{trace}\left(\Sigma(\mathcal{S}/t + A)^{-1}\right)$ since we are dealing with positive semi-definite matrices. Therefore,

$$\frac{\gamma(tA)}{t} \text{ is decreasing } \ .$$

We note that its limit is 0 at infinity and infinity at 0. Hence the equation

$$\frac{\gamma(tA)}{t} = 1 \text{ has a unique solution, } t_0 .$$

We now have found an estimator of $v'(\Sigma + A)^{-1}v$, since

$$v'(\Sigma + A)^{-1}v \simeq t_0 v'(\widehat{\Sigma} + t_0 A)^{-1}v .$$

4.2   Classification

Random matrix techniques offer us insights into the behavior of standard methods in high-dimension. Our work could be helpful in tuning regularization parameters, picking penalties etc... because we are able to predict performance of the methods, provided our assumptions are met. It is nonetheless clear that sometimes (actually many times), some of the quantities we are considering could be evaluated by leave-one out methods, which can be implemented efficiently because of rank-1 updates. In that case, our analysis has the merit of explaining the behavior of the techniques considered, something that alternative estimators (such as cross-validation) do not do.

A standard technique in classification is linear discriminant analysis. Some analysis in the high-dimensional context has already been done (Bickel and Levina (2003)), in a somewhat different direction. Here our aim is to explain what creates problems with LDA in high-dimension, even in the Gaussian case, and discuss briefly the behavior of Regularized discriminant analysis (RDA) proposed in Friedman (1989).

4.2.1   A preliminary remark

In the classification context (see details below), we will often be faced with a situation where a (regularized) covariance matrix is a pooled estimator of covariance computed from two groups, i.e

$$\widehat{\Sigma} = p_1 \widehat{\Sigma}_1 + p_2 \widehat{\Sigma}_2 .$$

In our context we will assume that the observations in each group have the same mean μ_i, where μ_i may depend on $i = 1, 2$. Assuming that the data is of the form

$$\mathfrak{X}_k = \mu_i + R_k X_k ,$$

where X_k has mean 0, we have for instance

$$\widehat{\Sigma}_1 = \frac{1}{N_1} \sum_{k=1}^{N_1} R_k^2 X_k X_k' - \widetilde{\mu}_1 \widetilde{\mu}_1' ,$$

where X_k have mean 0, and

$$\widetilde{\mu}_1 = \frac{1}{N_1} \sum_{i=1}^{N_1} R_k X_k .$$

We will naturally encounter forms of the type

$$(\widehat{\mu}_2 - \widehat{\mu}_1)'(\widehat{\Sigma} + A)^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1)$$

and we now explain how to find deterministic equivalents for the limiting behavior of these forms. We note that $\widehat{\mu}_i = \mu_i + \widetilde{\mu}_i$, so we will have to work out three quantities:

$$(\mu_2 - \mu_1)'(\widehat{\Sigma} + A)^{-1}(\mu_2 - \mu_1) , (\mu_2 - \mu_1)'(\widehat{\Sigma} + A)^{-1}(\widetilde{\mu}_2 - \widetilde{\mu}_1) \text{ and } (\widetilde{\mu}_2 - \widetilde{\mu}_1)'(\widehat{\Sigma} + A)^{-1}(\widetilde{\mu}_2 - \widetilde{\mu}_1) .$$

The first one is simple as it involves a shrunken matrix and deterministic vectors. The other two are a bit more subtle, since $\widehat{\Sigma}$ and $\widetilde{\mu}_i$'s interact (the Gaussian case being an exception for obvious reasons).

We call

$$\mathcal{S}_1 = \frac{1}{N_1 - 1} \sum_{k=1}^{N_1} R_k^2 X_k X_k' \text{ and } \mathcal{S}_2 = \frac{1}{N_2 - 1} \sum_{k=N_1+1}^{N_1+N_2} R_k^2 X_k X_k' ,$$

and, if $p_i = (N_i - 1)/(N_1 + N_2 - 2)$,

$$\mathcal{S} = p_1 \mathcal{S}_1 + p_2 \mathcal{S}_2 .$$

We note that, more generally we will have, for some $\tilde{p}_i$ (for instance $\tilde{p}_i = p_i N_i/(N_i - 1)$ if we wish to preserve unbiasedness),

$$\widehat{\Sigma} + A = \mathcal{S} + A - \tilde{p}_1 \widetilde{\mu}_1 \widetilde{\mu}_1' - \tilde{p}_2 \widetilde{\mu}_2 \widetilde{\mu}_2' ,$$

where

$$\mathcal{S} = \frac{1}{N_1 + N_2 - 2} \sum_{i=1}^{N_1 + N_2} R_i^2 X_i X_i' .$$

- **On $(\widetilde{\mu}_2 - \widetilde{\mu}_1)'(\widehat{\Sigma} + A)^{-1}(\widetilde{\mu}_2 - \widetilde{\mu}_1)$**

Using a rank-1 update formula (we could also use a more general version of the Sherman-Woodbury-Morrison formula), we have

$$\widetilde{\mu}_2'(\widehat{\Sigma} + A)^{-1} = \frac{\widetilde{\mu}_2'(\mathcal{S} + A - \tilde{p}_1 \widetilde{\mu}_1 \widetilde{\mu}_1')^{-1}}{1 - \tilde{p}_2 \widetilde{\mu}_2'(\mathcal{S} + A - \tilde{p}_1 \widetilde{\mu}_1 \widetilde{\mu}_1')^{-1} \widetilde{\mu}_2} .$$

Therefore, we have in particular,

$$\widetilde{\mu}_2'(\widehat{\Sigma} + A)^{-1} \widetilde{\mu}_2 = \frac{1}{\tilde{p}_2} \left(\frac{1}{1 - \tilde{p}_2 \widetilde{\mu}_2'(\mathcal{S} + A - \tilde{p}_1 \widetilde{\mu}_1 \widetilde{\mu}_1')^{-1} \widetilde{\mu}_2} - 1 \right) .$$

We also see by the same token that

$$\widetilde{\mu}_2'(\mathcal{S} + A - \tilde{p}_1 \widetilde{\mu}_1 \widetilde{\mu}_1')^{-1} \widetilde{\mu}_2 = \widetilde{\mu}_2'(\mathcal{S} + A)^{-1} \widetilde{\mu}_2 + \tilde{p}_1 \frac{(\widetilde{\mu}_2'(\mathcal{S} + A)^{-1} \widetilde{\mu}_1)^2}{1 - \tilde{p}_1 \widetilde{\mu}_1'(\mathcal{S} + A)^{-1} \widetilde{\mu}_1} .$$

Now recall that (see Subsubsection 3.3.2)

$$\widetilde{\mu}_2'(\widehat{\Sigma} + A)^{-1} \widetilde{\mu}_1 \simeq 0 .$$

So we conclude that

$$\boxed{\widetilde{\mu}_2'(\widehat{\Sigma} + A)^{-1} \widetilde{\mu}_2 \simeq \frac{\widetilde{\mu}_2'(\mathcal{S} + A)^{-1} \widetilde{\mu}_2}{1 - \tilde{p}_2 \widetilde{\mu}_2'(\mathcal{S} + A)^{-1} \widetilde{\mu}_2} .}$$

Naturally, our work in Subsubsection 3.3.2 allows us to find a deterministic equivalent to

$$\widetilde{\mu}_2'(\mathcal{S} + A)^{-1} \widetilde{\mu}_2$$

and so from then we get a deterministic equivalent to $\widetilde{\mu}_2'(\widehat{\Sigma} + A)^{-1} \widetilde{\mu}_2$. Of course, a similar analysis carries through for $\widetilde{\mu}_1'(\mathcal{S} + A)^{-1} \widetilde{\mu}_1$. To be more precise, if we call 1_{G_i} the vector that has 1 if $\mathfrak{X}_k$ is in group i and 0 otherwise, we see that the α that corresponds to $\widetilde{\mu}_2$ is

$$\alpha = \frac{\sqrt{n}}{N_2} 1_{G_2} ,$$

and we can apply our formulas.

We also need to consider

$$\widetilde{\mu}_2'(\mathcal{S} + A)^{-1} \widetilde{\mu}_1 .$$

Using the rank-1 update formula, we have

$$\widetilde{\mu}_2'(\mathcal{S} + A)^{-1} \widetilde{\mu}_1 = \frac{\widetilde{\mu}_2'(\mathcal{S} + A - \tilde{p}_1 \widetilde{\mu}_1 \widetilde{\mu}_1')^{-1} \widetilde{\mu}_1}{1 - \tilde{p}_2 \widetilde{\mu}_2'(\mathcal{S} + A - \tilde{p}_1 \widetilde{\mu}_1 \widetilde{\mu}_1')^{-1} \widetilde{\mu}_2} .$$

We have already worked out an approximation to the denominator. Now for the numerator, we have obviously

$$\widetilde{\mu}_2'(\mathcal{S} + A - \tilde{p}_1 \widetilde{\mu}_1 \widetilde{\mu}_1')^{-1} \widetilde{\mu}_1 = \frac{\widetilde{\mu}_1'(\mathcal{S} + A)^{-1} \widetilde{\mu}_2}{1 - \tilde{p}_1 \widetilde{\mu}_1'(\mathcal{S} + A)^{-1} \widetilde{\mu}_1} .$$

Hence, again, we see that in the asymptotic limit we consider,

$$\widetilde{\mu}_2'(\mathcal{S} + A - \tilde{p}_1 \widetilde{\mu}_1 \widetilde{\mu}_1')^{-1} \widetilde{\mu}_1 \simeq 0 \ .$$

So we conclude that

$$\boxed{\widetilde{\mu}_2'(\widehat{\Sigma} + A)^{-1} \widetilde{\mu}_1 \simeq 0 \ .}$$

● **On** $(\widetilde{\mu}_2 - \widetilde{\mu}_1)'(\widehat{\Sigma} + A)^{-1}\mu$

The idea is here again to use our rank-1 update formulas. We have

$$\widetilde{\mu}_2'(\widehat{\Sigma} + A)^{-1}\mu = \frac{\widetilde{\mu}_2'(\mathcal{S} + A - \tilde{p}_1 \widetilde{\mu}_1 \widetilde{\mu}_1')^{-1}\mu}{1 - \tilde{p}_2 \widetilde{\mu}_2'(\mathcal{S} + A - \tilde{p}_1 \widetilde{\mu}_1 \widetilde{\mu}_1')^{-1} \widetilde{\mu}_2} \ .$$

We also have

$$\widetilde{\mu}_2'(\mathcal{S} + A - \tilde{p}_1 \widetilde{\mu}_1 \widetilde{\mu}_1')^{-1}\mu = \widetilde{\mu}_2'(\mathcal{S} + A)^{-1}\mu + \tilde{p}_1 \frac{\widetilde{\mu}_2'(\mathcal{S} + A)^{-1}\widetilde{\mu}_1 \widetilde{\mu}_1'(\mathcal{S} + A)^{-1}\mu}{1 - \tilde{p}_1 \widetilde{\mu}_1'(\mathcal{S} + A)^{-1}\widetilde{\mu}_1} \ .$$

So we conclude that if, for instance $\|\mu\|$ stays bounded,

$$\boxed{\widetilde{\mu}_2'(\widehat{\Sigma} + A)^{-1}\mu \simeq 0 \ .}$$

We now have all the elements needed to get an asymptotically deterministic approximation to

$$(\widehat{\mu}_2 - \widehat{\mu}_1)'(\widehat{\Sigma} + A)^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1) \ .$$

4.2.2 LDA: Gaussian case

We recall the (optimal) setup. Suppose we have two groups (or classes). The observations can come from group 1 or group 2. In both groups they are $\mathcal{N}(\mu_{1,2}, \Sigma)$. The probability of belonging to group 1 is π_1. The question is now given an observation, how should it be classified?

It is easy and standard to find the optimal rule in the population. Namely, by doing likelihood computations, one quickly realizes that the optimal classification rule is (Hastie et al. (2009)): classify an observation as belonging to Group 2 if

$$x'\Sigma^{-1}(\mu_2 - \mu_1) \geq \frac{1}{2}(\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 + \mu_1) + \log(\pi_1/\pi_2) \ .$$

Naturally, in practice, Σ and μ_1 and μ_2 need to be estimated. A natural solution is to use the training data (which is labeled, i.e we know to which class each observation belongs) to estimate μ_1 and μ_2 and then use a pooled estimate of covariance for Σ.

In somewhat more details, if we have N_1 observations that belong to class 1 in our training set, and N_2 that belong to class 2, let us denote by $\widehat{\mu}_1$ and $\widehat{\mu}_2$ the sample mean of the observations in group 1 and group 2. If $\widehat{\Sigma}_1$ and $\widehat{\Sigma}_2$ are the sample covariance in each of these groups, then our estimate of Σ is

$$\widehat{\Sigma} = \frac{1}{N_1 + N_2 - 2}\left((N_1 - 1)\widehat{\Sigma}_1 + (N_2 - 1)\widehat{\Sigma}_2\right) \ .$$

(We will assume in the following discussion that $p \leq N_1 + N_2 - 2$ so $\widehat{\Sigma}$ is invertible.) It is now natural to ask the following questions:

1. how does naive LDA perform?

2. how suboptimal is the naive threshold?

3. is it possible to estimate the minimal misclassification rate, even if we cannot find the optimal direction on which to project a new observation?

Naturally, when a Gaussian vector is projected on a direction d, its distribution becomes $\mathcal{N}(\mu'd, d'\Sigma d)$. If our decision rule is to classify x to Group 2 if $x'd > t$, it is clear that the misclassification rate is, if $\mu_1(d) = \mu_1'd$, $\mu_2(d) = \mu_2'd$ and $\sigma^2(d) = d'\Sigma d$,

$$\pi_1(1 - \Phi(\frac{t - \mu_1(d)}{\sigma})) + \pi_2\Phi(\frac{t - \mu_2(d)}{\sigma(d)}) \ .$$

A simple computation therefore shows that the optimal threshold is

$$t^* = \frac{\sigma^2(d)}{\mu_2(d) - \mu_1(d)}\log(\pi_1/\pi_2) + \frac{\mu_2(d) + \mu_1(d)}{2} \ .$$

Hence we have

$$\frac{t^* - \mu_{1,2}}{\sigma} = \pm\frac{\mu_2 - \mu_1}{2\sigma} + \frac{\sigma}{\mu_2 - \mu_1}\log(\frac{\pi_1}{\pi_2}) \ .$$

We can therefore compute the optimal misclassification rate as

$$\pi_1(1 - \Phi(\frac{\mu_2 - \mu_1}{2\sigma} + \frac{\sigma}{\mu_2 - \mu_1}\log(\frac{\pi_1}{\pi_2}))) + \pi_2\Phi(-\frac{\mu_2 - \mu_1}{2\sigma} + \frac{\sigma}{\mu_2 - \mu_1}\log(\frac{\pi_1}{\pi_2})) \ .$$

Note that in LDA in the population, we have $\mu_2(d) - \mu_1(d) = (\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 - \mu_1) = \sigma^2$. Hence some simplifications ensue; in particular, the optimal misclassification rate is, if σ is the Mahalanobis distance between μ_2 and μ_1,

$$\pi_1 - \pi_1\Phi(\frac{\sigma}{2} + \frac{1}{\sigma}\log(\pi_1/\pi_2)) + \pi_2\Phi(-\frac{\sigma}{2} + \frac{1}{\sigma}\log(\pi_1/\pi_2)) \ .$$

Hence, our problems reduce to:

1. Estimate the Mahalanobis distance between μ_1 and μ_2 so we can compute the optimal misclassification rate for the problem

2. Estimate t^* from the data to obtain a procedure that outperforms the naive procedure.

We note that it is good practice to do cross-validation to estimate t^* - and this has been recognized by practitioners, see Hastie et al. (2009). However, even when the data is Gaussian, as we show below, a correction to the naive empirical threshold is needed in high-dimension.

• **Estimation of t^*.** When $d = \widehat{\Sigma}^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1)$, we have

$$\sigma^2(d) = (\widehat{\mu}_2 - \widehat{\mu}_1)'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1) \ ,$$
$$\mu_i(d) = \mu_i'\widehat{\Sigma}^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1) \ .$$

In the Gaussian case, using properties of Wishart matrices (the interested reader is also refered to El Karoui (2009c) for similar computations, but going beyond the Wishart case), we see that, if $\rho = p/N$,

$$\sigma^2(d) \simeq (\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 - \mu_1)\frac{1}{(1 - \rho)^3} + \frac{1}{(1 - \rho)^3}(\frac{p}{N_1} + \frac{p}{N_2})$$

On the other hand,

$$\mu_i(d) \simeq \frac{1}{1 - \rho}\mu_i'\Sigma^{-1}(\mu_2 - \mu_1) \ .$$

Now from the data we can get an estimate of $(\widehat{\mu}_2 - \widehat{\mu}_1)'\widehat{\Sigma}^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1)$. A simple computation, based on properties of Wishart matrices (see e.g El Karoui (2009b) for full details) gives:

$$(\widehat{\mu}_2 - \widehat{\mu}_1)'\widehat{\Sigma}^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1) \simeq \frac{1}{1 - \rho}\left[(\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 - \mu_1) + \frac{p}{N_1} + \frac{p}{N_2}\right] \simeq (1 - \rho)^2\sigma^2(d) \ .$$

On the other hand,

$$\widehat{\mu}_2'\widehat{\Sigma}^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1) \simeq \frac{1}{1 - \rho}\left[\mu_2'\Sigma^{-1}(\mu_2 - \mu_1) + \frac{p}{N_2}\right] \ ,$$
$$\widehat{\mu}_1'\widehat{\Sigma}^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1) \simeq \frac{1}{1 - \rho}\left[\mu_1'\Sigma^{-1}(\mu_2 - \mu_1) - \frac{p}{N_1}\right] \ .$$

So we can estimate $\mu_i(d)$ by

$$\mu_i(d) \simeq \widehat{\mu}_i' \widehat{\Sigma}^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1) \pm \frac{p}{N_i(1-\rho)} ,$$

where $\pm$ is 1 for $i = 1$ and $\pm = -1$ for $i = 2$.

We can now estimate t^* by putting together all these estimators. (We note that we could also do this by using estimate of $\sigma^2(d)$ and $\mu_i(d)$ based on leave-one out procedures. However, the advantage of the procedure proposed here is that the amount of extra computations is extremely small, since the corrections are known in closed form.)

On the other hand, it is clear that the naive threshold value is (in general) suboptimal. As a matter of fact, it is approximately

$$t_{\text{naive}} \simeq \frac{1}{2} \left[\frac{1}{1-\rho}(\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 + \mu_1) + \frac{1}{1-\rho}\left[\frac{p}{N_2} - \frac{p}{N_1} \right] \right] + \log(\pi_1/\pi_2) .$$

On the other hand, if maha $= (\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 - \mu_1)$,

$$t^* \simeq \frac{1}{2}\frac{1}{1-\rho}(\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 + \mu_1) + \log(\pi_1/\pi_2)\frac{1}{(1-\rho)^2}\left[1 + \left(\frac{p}{N_1} + \frac{p}{N_2} \right)\frac{1}{\text{maha}} \right] .$$

Let us further remark that when $N_1 = N_2$, because $\log(\pi_1/\pi_2) = 0$, our correction returns exactly the naive threshold, and hence will not yield improvements. On the other hand, in this situation, the naive threshold is close to optimal and our analysis shows that further numerical investigation of a good threshold is not needed.

In other respects, it is rather easy to estimate $(\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 - \mu_1)$, and hence get the optimal misclassification rate for any classification procedure, in the case where the data is truly Gaussian. Note that this is not available by using cross-validation.

Hence, beside shedding light on the potential (limited) problems of LDA in high-dimension, the computations we showed can be used to establish a benchmark for how well a classification procedure can perform and perhaps helps the user in choosing something better than LDA - or convincing her that LDA (perhaps corrected) in her context is performing quite well and close to the optimum.

4.2.3 "LDA": elliptical case

We are now interested in finding a reasonable classification procedure for elliptical data in high-dimension. We will see that the results obtained in this paper are relevant to shed light on their behavior.

We consider the case here where R_i's have a smooth density. The data is modeled as

$$\mathfrak{X}_i = \mu_{1,2} + R_i X_i .$$

We will focus on the case $X_i \sim \mathcal{N}(0, \Sigma)$, though some of the computations could be carried in a more complex situation. Let us call f the density of R. The density of

$$\mathfrak{X} = \mu_{1,2} + R_i X_i ,$$

is, since it is a continuous scale mixture of normal,

$$\phi(y; \mu) = \int f(r)r^{-p}\frac{\exp\left(-\frac{(y-\mu)'\Sigma-1(y-\mu)}{2r^2} \right)}{\sqrt{\det(2\pi\Sigma)}} dr .$$

Hence, it is difficult to get an exactly optimal classification rule by using a likelihood method. Nonetheless, we can apply Laplace's method to approximate this integral.

We now recall the model from which $\mathfrak{X}$ is generated and we see that $\frac{(\mathfrak{X}-\mu)'\Sigma^{-1}(\mathfrak{X}-\mu)}{p}$ is concentrated around R^2 if $(\mu_{1,2} - \mu)\Sigma^{-1}(\mu_{1,2} - \mu) = O(1)$.

We are going to make the assumption that $(\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 - \mu_1) = O(1)$. Calling, for y a dummy variable assumed to take values only where $\mathfrak{X}$ concentrates,

$$\alpha_p(i) = \frac{(y - \mu_i)'\Sigma^{-1}(y - \mu_i)}{p} ,$$

we see that $\alpha_p(i) = \mathrm{O}(1)$ (indeed $\alpha_p(i) \simeq R^2$; see the remark on $\mathfrak{X}$ above) and

$$|\alpha_p(1) - \alpha_p(2)| = \mathrm{O}(1/p) \ .$$

Hence applying Laplace's method, we see that

$$\phi(y; \mu_i) \sim f(\sqrt{\alpha_p(i)}) \exp(-p/2(\log(\alpha_p(i)) + 1)) \sqrt{\pi \alpha_p(i)/p} \ .$$

Hence, under our assumptions, if $\Delta = p(\alpha_p(2) - \alpha_p(1))$ (which is of order 1),

$$\frac{\phi(y; \mu_1)}{\phi(y; \mu_2)} \simeq \exp(\Delta/(2\alpha_p(1))) \ .$$

Now $-p\Delta = 2(\mu_2 - \mu_1)'\Sigma^{-1}y + \mu_1'\Sigma^{-1}\mu_1 - \mu_2'\Sigma^{-1}\mu_2$. Hence, if the prior probabilities are π_1 and π_2, a reasonable rule for classification appears to be: classify in group 2 if, for a new observation y,

$$y'\Sigma^{-1}(\mu_2 - \mu_1) \geq \alpha_p(1)\log\left(\frac{\pi_1}{\pi_2}\right) + \frac{(\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 + \mu_1)}{2} \ . \tag{35}$$

Here, in what is perhaps a surprise, we see that in high-dimension, in the class of elliptical distribution a procedure similar to LDA seems quite reasonable.

Under our assumptions, it should be noted that in high-dimension, $\alpha_p(1) \simeq R_i^2$ (for new data generated according to our model). Therefore this rule consistent with LDA, since for Gaussian data $R_i^2 = 1$. Now, if $\|\mu_2 - \mu_1\|^2 \ll \mathrm{trace}(\Sigma)$, we see that

$$\frac{\|\mathfrak{X} - \mu_i\|^2}{\mathrm{trace}(\Sigma)} \simeq R_i^2 \ ,$$

hence, the rule is approximately implementable - though situations where R_i has very heavy tails are likely to be very hard on these approximations.

Now in the elliptical case, we know (see El Karoui (2009b)) that there exists $\mathfrak{s}$ such that if $\mathfrak{X}$ is independent of $\widehat{\Sigma}$, $\widehat{\mu}_1$ and $\widehat{\mu}_2$,

$$\mathfrak{X}'\widehat{\Sigma}^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1) \simeq \mathfrak{s}\mathfrak{X}'\Sigma^{-1}(\mu_2 - \mu_1)$$

and we can also find an approximation of

$$\frac{(\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 + \mu_1)}{2}$$

through appropriate corrections, the key computations having been carried out in El Karoui (2009b). Hence, we can design a classification rule by using (nearly) unbiased estimators of the quantities on both sides of Equation (35). This could naturally also be done using leave-one out procedures, though these procedures would not explain what is happening.

●**On changing estimators of covariance**

One advantage of the analyses we have carried out is that they reveal (somewhat explicitly) the role played by the R_i's. Since those are essentially estimable (for instance in the Gaussian, and in general as soon as we have measure concentration), we could also envision different weighting schemes, in particular putting all of them to 1 (which is extremely natural from a convexity standpoint), which amounts to using estimators which are similar in spirit to Tyler's estimator (see Tyler (1987) and El Karoui (2009b) for more details.) Because the paper is already quite long, we will not seek an optimal procedure here, but our various estimates (here and in El Karoui (2009b), El Karoui (2009c)) can in principle be used to assess difference in performance between these estimators of covariance for the statistical tasks at hand.

● **Computing the misclassification rate in the elliptical setting**

Suppose we now use a simple threshold rule, similar to LDA, to classify. Though this is suboptimal, understanding the behavior of this simple rule is interesting, and helps shed light on various procedures and their robustness.

So suppose we classify an observation x to Group 2 if $x'v > t$. Suppose that x is elliptical and call f the density of the R. A computation similar to the ones carried before shows that the misclassification rate is

$$\pi_1 \int f(r)\Phi\left(\frac{\mu_1 - t}{\sqrt{v'\Sigma vr}}\right) dr + \pi_2 \int f(r)\Phi\left(\frac{t - \mu_2}{\sqrt{v'\Sigma vr}}\right) dr \ .$$

Since we are able to estimate μ_i's, and $v'\Sigma v$, as well as (at least coarsely) the density f - since we can estimate the R_i's, we can find the optimal threshold t^*. This gives a principled alternative to cross-validation in this case (though leave-one-out techniques could also be used).

4.2.4 RDA

In Friedman (1989), partly motivated by questions having to do with the variability of LDA procedures (in particular when Σ is ill-conditioned), it was proposed to replace $\widehat{\Sigma}$ by

$$\widetilde{\Sigma} = (1 - w)\widehat{\Sigma} + wA \ ,$$

where A is a matrix towards which Σ is shrunken. The computations done in the first part of the paper allow us to measure the performance of RDA in our asymptotic context.

Our results show that when w varies from 0 to 1, up to a computable scaling factor, forms of the type $v'\widetilde{\Sigma}v$ cover the range of $v'[(1-\lambda)\Sigma + \lambda A]^{-1}v$, for λ varying from 0 to 1, though of course λ is very different from w (and λ depends on the ellipticity of the data). This property is something that is not immediately obvious in high-dimension. This is valid much beyond the Gaussian design case, as we have shown.

Let us now illustrate this in the Gaussian case. In this case, we know how to pick the optimal threshold at given w and can compute the misclassification rate of the corresponding procedure. Our results also show that the naive threshold is suboptimal, and suggests corrections, though those can also be found using leave-one out procedures that do not rely on our understanding of the phenomena. (This is fairly similar to our more detailed LDA discussion.)

Our computations also show that one should probably not use 5 or 10 fold cross-validation methods in high-dimension, since it affects that the ratio p/n, which is key in determining and getting optimal performance.

Here again, a rigorous study of the impact of R_i on the quality of classification and the potential benefits of using robust estimate of scatter is now feasible but we postpone it to other investigations because of the length of this paper.

4.3 Optimization problems

Suppose we consider the optimization problem

$$\begin{cases} \min_w w'\Sigma w \\ \text{subject to } V'w = U \end{cases} ,$$

where V is a $p \times k$ matrix of constraints, and U is a $k \times 1$ vector of values for those constraints. This is a canonical problem in portfolio optimization (see Meucci (2005), Markowitz (1952)). Under minimal invertibility conditions, the solution is

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

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

Suppose that we estimate Σ by $\widetilde{\Sigma} = \lambda\widehat{\Sigma} + A$ and suppose that V contains a constraint involving μ, which is not known and needs to be estimated. Call $\widehat{w}$ the corresponding solution and $\widehat{M} = \widehat{V}'f(\widehat{\Sigma})^{-1}\widehat{V}$. Then our estimates allow us to get a deterministic equivalent to the naive estimate of the risk, namely, $\widehat{w}'\widetilde{\Sigma}\widehat{w} = U'\widehat{M}U$ as well as the true risk of our allocation, i.e $\widehat{w}'\Sigma\widehat{w}$, at least when the number of constraints is fixed.

Let us now be a bit more specific. Suppose $\widetilde{\Sigma} = \widehat{\Sigma} + A$ (scalar constants can easily be dealt with), that the number of constraints is fixed and V contains only fixed constraints (i.e nothing needs to be estimated,

and in particular not the mean - this is for instance the case when users in Finance perform minimum variance optimization, without regards for expected returns). Then,

$$\widehat{M} \simeq V'(\gamma(A)\Sigma + A)^{-1}V = \widetilde{M}_A,$$

so we get as deterministic equivalent of the naive risk

$$U'(V'(\gamma(A)\Sigma + A)^{-1}V)^{-1}U .$$

The interpretation of this result is that the shrinkage procedure essentially produces an estimator which is a dampen shrinkage estimator, the damping factor being $\gamma(A)$.

To compute the realized risk, all one needs to do is look at $U'\widehat{M}^{-1}V'(\widehat{\Sigma} + A)^{-1}\Sigma(\widehat{\Sigma} + A)^{-1}V\widehat{M}^{-1}U$. To understand this, we can just rely on the results of Heuristic 2.2, with $B = \Sigma$. It should be noted that

$$V'(\widehat{\Sigma} + A)^{-1}\Sigma(\widehat{\Sigma} + A)^{-1}V \simeq [1 + \xi(A, \Sigma)]V'\Sigma V = [1 + \xi(A, \Sigma)]M .$$

Hence,

$$U'\widehat{M}^{-1}V'(\widehat{\Sigma} + A)^{-1}\Sigma(\widehat{\Sigma} + A)^{-1}V\widehat{M}^{-1}U \simeq [1 + \xi(A, \Sigma)]U'\widetilde{M}_A^{-1}M\widetilde{M}_A^{-1}U .$$

The situation where V involves μ and is replaced by $\widehat{\mu}$ in $\widehat{V}$ can be investigated using our results on quadratic forms in $DX(X'D^2X + A)^{-1}X'D$ and the other results we developed in the paper specifically for this task.

Finally, to the reader who might wonder why the study of $M^{-1}\Sigma_\epsilon M^{-1}$ is potentially useful, even in the setting where $\mathfrak{X}_i$ are i.i.d and hence have the same covariance Σ, let us give a "practical" example: it is sometimes the case that in the context of portfolio optimization, one uses log-returns instead of returns to find the portfolio weights. This is found to be natural when the stock prices follow geometric brownian motions, as in the Black-Scholes model. But clearly, in that setting of log-normal prices, the risk exposure should be computed using the covariance of the returns and not that of the log returns - two matrices that are in general different. (Note that our results (and our work on log-normal distributions) also give risk predictions when using returns instead of log returns when working with log-normal data.)

4.4   Ridge regression

Suppose we consider ridge regression with a general quadratic penalty (a.k.a Tikhonov regularization). Then $\widehat{\beta}$ is found by solving

$$\widehat{\beta}_{\text{ridge}} = \text{argmin}_\beta \|Y - \frac{1}{\sqrt{n}}X\beta\|_2^2 + \lambda\beta'\Gamma\beta ,$$

where Y is our response, X is the design matrix and Γ is a psd matrix. It is easy to verify that

$$\widehat{\beta}_{\text{ridge}} = \frac{1}{\sqrt{n}}(\frac{1}{n}X'X + \lambda\Gamma)^{-1}X'Y .$$

Suppose that $Y = \frac{1}{\sqrt{n}}[X\beta_0 + \epsilon]$. Then,

$$\widehat{\beta}_{\text{ridge}} = (\frac{X'X}{n} + \lambda\Gamma)^{-1}(\frac{X'X}{n}\beta_0 + \frac{X'}{n}\epsilon) .$$

Hence,

$$\widehat{\beta}_{\text{ridge}} - \beta_0 = -\lambda(\frac{X'X}{n} + \lambda\Gamma)^{-1}\Gamma\beta_0 + (\frac{X'X}{n} + \lambda\Gamma)^{-1}\frac{X'}{n}\epsilon .$$

The situation where the design is random can now be studied with our tools, provided the assumptions of our theorems are satisfied.

For instance if ϵ has covariance Σ_ϵ and mean 0, we have

$$\mathbf{E}\left(\|\widehat{\beta}_{\text{ridge}} - \beta_0\|_2^2|X\right) = \lambda^2\beta_0'\Gamma'(\frac{X'X}{n} + \lambda\Gamma)^{-2}\Gamma\beta_0 + \frac{1}{n}\text{trace}\left((\frac{X'X}{n} + \lambda\Gamma)^{-1}\frac{X'\Sigma_\epsilon X}{n}(\frac{X'X}{n} + \lambda\Gamma)^{-1}\right) .$$

The first quantity can be analyzed using our results in this paper. The second one is comparatively simpler and comes out of random matrix arguments. For instance, when $\Sigma_\epsilon = \mathrm{Id}_n$, we see that we are left with

$$\mathrm{trace}\left((\frac{X'X}{n} + \lambda\Gamma)^{-1}\frac{X'X}{n}(\frac{X'X}{n} + \lambda\Gamma)^{-1}\right) = \mathrm{trace}\left((\frac{X'X}{n} + \lambda\Gamma)^{-1}\right) - \lambda\mathrm{trace}\left(\Gamma(\frac{X'X}{n} + \lambda\Gamma)^{-2}\right) ,$$

and these quantities can be analyzed using standard results on Stieltjes transforms (as well as the derivation trick we use repeatedly in this paper).

We also note that if X has a symmetric distribution (we could relax of course this assumption with some work done along the lines of what is done in the paper),

$$\frac{1}{n}\mathbf{E}\left(\mathrm{trace}\left((\frac{X'X}{n} + \lambda\Gamma)^{-1}\frac{X'X}{n}(\frac{X'X}{n} + \lambda\Gamma)^{-1}\right)\right) = \mathbf{E}\left(\frac{1'}{\sqrt{n}}\frac{X}{\sqrt{n}}(\frac{X'X}{n} + \lambda\Gamma)^{-2}\frac{X'}{\sqrt{n}}\frac{1}{\sqrt{n}}\right) ,$$

and we can therefore use the work done in 3.4.2. A similar argument would hold if Σ_ϵ were diagonal, with 1 replaced by u, with $u_i^2 = \Sigma_\epsilon(i,i)$.

The arguments presented in this paper can also be used to understand the quantities $\|\widehat{\beta}_{\mathrm{ridge}} - \beta_0\|_2^2$ directly, before taking expectation, if for instance we have a bound (with high-probability) on $\|\epsilon\|$.

Our concentration arguments also allow us to show that

$$\frac{1}{p}\mathbf{E}\left(\|\widehat{\beta}_{\mathrm{ridge}}(X,Y) - \beta_0\|_2^2\right)$$

has the same limit as the conditional version.

5 Conclusion

Our study aimed at showing that the tools of random matrix theory could be used to further our understanding of various statistical procedures based on shrinkage estimators of covariance. Despite the great recent interest in l_1-type regularizations, these more classical methods are still very useful and very much in use, which is why we undertook the task of explaining what they actually did (at least asymptotically) in high-dimension. We also note that our study has moved us now quite far away from "linear" models for the data and we have obtained results for distributions with genuinely non-linear structures, something that is very much needed to understand various practical applications.

We have both shown what we think is a great distributional robustness of random matrix based results in this context and a great geometric fragility of those models: distributional assumptions are largely irrelevant as long as they have the same geometric implications for the data; when two models yield a different geometry, the limiting approximations can change completely. Hence it seems to us that our study highlights a basic applied fact: namely users of random matrix results should run diagnostic tests before they apply (or rely on) results obtained in Gaussian or Gaussian like situations (which are the only ones covered by the "classical" random matrix models). For otherwise, if there is e.g correlation between our n observations, or if the geometry of the dataset does not conform to "i.i.d Gaussian" geometry, naive random matrix predictions will prove unhelpful and uninformative at best.

On a technical note, our results are quite general, thanks in large part to the approach we used, which does not require us to compute the limit (or deterministic equivalent) of various quantities to show it is the same when our data come from a wide class of possible distributions. It should be noted that our results encompass many distributions for which natural questions in random matrix theory (such as behavior of largest and smallest eigenvalues) have not yet been settled or even investigated. In the future, it might also be of interest to look into more general estimates of covariance, namely matrix functions of the (shrunken) covariance matrix, i.e estimates that apply a certain fixed function to the eigenvalues of the shrunken matrix and leave the eigenvectors as is. This seems very approachable by our methods, using Cauchy's formula for instance, but because this might be considered a bit less central to multivariate statistics we postpone a rigorous study of these questions to a possible future paper.

APPENDIX

A A remark on robustness of spectral distributions

This technical appendix is not directly related to the rest of the paper but shows how the methods we used can be utilized to analyze the robustness of another quantity of interest in random matrix theory, namely the spectral distribution of the matrix. (We put the result here because it fits our theme of robustness and is interesting but of course does not warrant its own paper.)

In El Karoui (2009a), we investigated the robustness properties of generalizations of the Marčenko-Pastur equation and showed that it held under mild concentration requirements on the data.

As a first step we showed that we could use Azuma's inequality to control the fluctuations of the Stieltjes transform for a very broad class of distributions. Now to show robustness, all we have to do is show that the expectation of the Stieltjes transform is the same for all the models we consider. In El Karoui (2009a), we limited ourselves to models for which the data $\mathfrak{X}_i$ had the same covariance for all i. We can now use similar ideas to the ones we have developed in this paper to do it in a more general case. We call

$$\mathcal{S}_X = \frac{1}{n}\sum_{i=1}^n R_i^2 X_i X_i'$$

and the corresponding Stieltjes transform (for $\mathcal{S}_X + A$)

$$m_{p,X}(z) = \frac{1}{p}\text{trace}\left(\left(\mathcal{S}_X + A - z\text{Id}\right)^{-1}\right)\,,$$

where A is a (deterministic) psd matrix, $z \in \mathbb{C}^+$ and $\text{Im}\,[z] = v > 0$. We call $u = \text{Re}\,[z]$.

We have the following theorem.

Theorem A.1. *Under the usual assumptions of this paper (see Subsection 3.2), assuming that the R_i's are deterministic, and $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ have mean 0 and are such that $\text{cov}\,(X_i) = \text{cov}\,(Y_i)$, we have, for any fixed $z \in \mathbb{C}^+$,*

$$\left|\mathbf{E}\left(m_{p,X}(z) - m_{p,Y}(z)\right)\right| \leq \frac{1}{p}\sum_{i=1}^n\left(\frac{K|z|}{v^3}\frac{R_i^2}{n}\left[b_{Q_2}(1;X_i) + b_{Q_2}(1;Y_i)\right]\right) \wedge \frac{2}{v}\,.$$

This extends some of the results of El Karoui (2009a), since under various concentration assumptions we will be able to control $b_{Q_2}(1;Y_i)$ and $b_{Q_2}(1;X_i)$ (recall that when X_i are Gaussian with covariance bounded in operator norm, $b_{Q_2}(1;X_i)$ is of order $\sqrt{p}$). Note once again that the models considered here are richer than the ones considered in El Karoui (2009a). The main difference with the results of El Karoui (2009a) is that this new theorem covers cases where we cannot describe the limit, whereas in El Karoui (2009a) we described the limit "explicitly".

We refer the reader to El Karoui (2009a) (or Bai and Silverstein (2010)) for details explaining why showing a.s convergence of the Stieltjes transform at each z (and a mass preservation condition) gives a.s weak convergence of the spectral distribution. Essentially our theorem says that the existence of a limit needs to be checked only in the Gaussian case and that such a result would transfer over to more general distributions for which we control $b_{Q_2}(1;Y_i)$.

Proof. We go quick on the details of the proof because we have done many similar ones in the paper. We take a Lindeberg approach, naturally. It is clear that if $B_j = \frac{1}{n}\sum_{k=1}^{j-1}R_k^2 X_k X_k' + \frac{1}{n}\sum_{k=j+1}^n R_k^2 Y_k Y_k' + A$ (with obvious adjustments mentioned in the paper for $j=1$ and $j=n$), all we have to do is understand

$$\frac{1}{p}\mathbf{E}\left(\text{trace}\left(\left(B_j + \frac{R_j^2}{n}X_j X_j' - z\text{Id}\right)^{-1}\right) - \text{trace}\left(\left(B_j + \frac{R_j^2}{n}Y_j Y_j' - z\text{Id}\right)^{-1}\right)\right)\,.$$

Let us call $B_j(z) = B_j - z\text{Id}$. Note that B_j is psd. By standard rank-1 updates arguments, we have

$$\Delta_j = \text{trace}\left(\left(B_j(z) + \frac{R_j^2}{n}X_j X_j'\right)^{-1}\right) - \text{trace}\left(\left(B_j(z) + \frac{R_j^2}{n}Y_j Y_j'\right)^{-1}\right)$$

$$= \frac{R_j^2}{n}\left[-\frac{X_j' B_j^{-2}(z)X_j}{1 + \frac{R_j^2}{n}X_j' B_j^{-1}(z)X_j} + \frac{Y_j' B_j^{-2}(z)Y_j}{1 + \frac{R_j^2}{n}Y_j' B_j^{-1}(z)Y_j}\right]$$

Let us call $d_j(z) = \mathrm{trace}\left(B_j(z)^{-1}\Sigma_j\right)$, where Σ_j is the covariance of X_j and Y_j. Clearly, since $d_j(z)$ is independent of X_j and Y_j,

$$\mathbf{E}\left(\frac{X_j'B_j^{-2}(z)X_j}{1+\frac{R_j^2}{n}d_j(z)}\right) = \mathbf{E}\left(\frac{Y_j'B_j^{-2}(z)Y_j}{1+\frac{R_j^2}{n}d_j(z)}\right) = \mathbf{E}\left(\frac{\mathrm{trace}\left(B_j^{-2}(z)\Sigma_j\right)}{1+\frac{R_j^2}{n}d_j(z)}\right).$$

So to control $|\mathbf{E}\left(\Delta_j\right)|$, all we have to do is control, if we call $q_j(z) = X_j'B_j^{-1}(z)X_j$,

$$\left|\mathbf{E}\left(\frac{R_j^2}{n}\frac{X_j'B_j^{-2}(z)X_j}{1+\frac{R_j^2}{n}d_j(z)} - \frac{R_j^2}{n}\frac{X_j'B_j^{-2}(z)X_j}{1+\frac{R_j^2}{n}q_j(z)}\right)\right|.$$

The quantity inside the expectation can be rewritten

$$\Omega_j = \frac{R_j^2}{n}X_j'B_j^{-2}(z)X_j\frac{R_j^2}{n}\frac{q_j(z)-d_j(z)}{(1+\frac{R_j^2}{n}d_j(z))(1+\frac{R_j^2}{n}q_j(z))}.$$

Lemma 2.6 in Silverstein and Bai (1995) shows that

$$\left|\frac{R_j^2}{n}\frac{X_j'B_j^{-2}(z)X_j}{1+\frac{R_j^2}{n}q_j(z)}\right| \le \frac{1}{v}.$$

Hence, $|\Delta_j| \le 2/v$ and

$$|\mathbf{E}\left(\Omega_j\right)| \le \frac{1}{v}\frac{R_j^2}{n}\mathbf{E}\left(\frac{|d_j(z)-q_j(z)|}{|1+\frac{R_j^2}{n}d_j(z)|}\right).$$

By writing $B_j^{-1}(z)$ in terms of its eigenvalues and eigenvectors, we note that $\mathrm{Im}\left[z\mathrm{trace}\left(B_j^{-1}(z)\Sigma_j\right)\right] \ge 0$ because B_j and Σ_j are psd and $z \in \mathbb{C}^+$ (alternatively, $\mathrm{Im}\left[zB_j^{-1}(z)\right]$ is psd). Therefore $\mathrm{Im}\left[zd_j(z)\right] \ge 0$. Hence,

$$\frac{1}{|z(1+\frac{R_j^2}{n}d_j(z))|} \le \frac{1}{v}.$$

So finally,

$$|\mathbf{E}\left(\Omega_j\right)| \le \frac{|z|}{v^2}\frac{R_j^2}{n}\mathbf{E}\left(|d_j(z)-q_j(z)|\right).$$

We now have to analyze $d_j(z) - q_j(z)$. We notice that

$$q_j(z) - d_j(z) = X_j'M_1X_j + iX_j'M_2X_j - \mathbf{E}_j\left(X_j'M_1X_j + iX_j'M_2X_j\right),$$

where if α_k's are the eigenvectors of B_j and λ_k its eigenvalues, we have

$$\mathrm{Re}\left[B_j^{-1}(z)\right] = M_1 = \sum_{k=1}^{p}\frac{\lambda_k - u}{(\lambda_k - u)^2 + v^2}\alpha_k\alpha_k'$$

and

$$\mathrm{Im}\left[B_j^{-1}(z)\right] = M_2 = \sum_{k=1}^{p}\frac{v}{(\lambda_k - u)^2 + v^2}\alpha_k\alpha_k'.$$

M_1 can be written as $M_1 = M_{1,+} - M_{1,-}$, where $M_{1,+}$ is formed by keeping the non-negative eigenvalues of M_1 and replacing the negative ones by 0. Of course, $M_{1,+}$ and $M_{1,-}$ are psd (technically we should index them by u, but we do not do it to alleviate the notation). We now remark that $M_{1,+}$, $M_{1,-}$ and M_2 are psd with $\|M_{1,\pm}\| \le 1/v$ and $\|M_2\| \le 1/v$. We can therefore conclude, using the fact that $|z| \le |\mathrm{Re}\left[z\right]| + |\mathrm{Im}\left[z\right]|$ as well as the fact that M_1 and M_2 are independent of X_j that

$$\mathbf{E}\left(|d_j(z)-q_j(z)|\right) \le \frac{K}{v}b_{Q_2}(1;X_j).$$

Putting everything together we obtain the result announced in the theorem. $\qquad\square$

References

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

BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, New York, second edition. URL http://dx.doi.org/10.1007/978-1-4419-0661-8.

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.

BHATIA, R. (1997). *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York.

BICKEL, P. and LEVINA, E. (2003). Some theory for Fisher's Linear Discriminant function, naive "Bayes", and some alternatives when there are many more variables than observations. Technical Report 404, University of Michigan, Department of Statistics.

BURKHOLDER, D. L. (1973). Distribution function inequalities for martingales. *Ann. Probability* **1**, 19–42.

CHATTERJEE, S. (2005). A simple invariance principle Available at http://arxiv.org/abs/math/0508213.

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

EATON, M. L. (2007). *Multivariate statistics*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 53. Institute of Mathematical Statistics. Reprint of the 1983 original.

EFRON, B. and STEIN, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9**, 586–596.

EL KAROUI, N. (2009a). Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability* **19**, 2362–2405.

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.

EL KAROUI, N. (2009c). On the realized risk of high-dimensional Markowitz portfolios. Technical Report 784, Department of Statistics, UC Berkeley.

EL KAROUI, N. (2011). *Handbook of random matrix theory*, chapter Multivariate Statistics (28). Oxford University Press.

FRIEDMAN, J. H. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.* **84**, 165–175.

GIRKO, V. L. (1990). *Theory of random determinants*, volume 45 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht. Translated from the Russian.

HAFF, L. R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann. Statist.* **8**, 586–597.

HASTIE, T., TIBSHIRANI, R., and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer-Verlag, New York, 2nd ed. edition. Data mining, inference, and prediction.

HORN, R. A. and JOHNSON, C. R. (1990). *Matrix analysis*. Cambridge University Press, Cambridge. Corrected reprint of the 1985 original.

HORN, R. A. and JOHNSON, C. R. (1994). *Topics in matrix analysis*. Cambridge University Press, Cambridge. Corrected reprint of the 1991 original.

JOHNSTONE, I. (2001). On the distribution of the largest eigenvalue in principal component analysis. *Ann. Statist.* **29**, 295–327.

JOHNSTONE, I. M. (2007). High dimensional statistical inference and random matrices. In *International Congress of Mathematicians. Vol. I*, pp. 307–333. Eur. Math. Soc., Zürich.

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.

LINDEBERG, J. W. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Math. Z.* **15**, 211–225. URL http://dx.doi.org/10.1007/BF01494395.

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.

MEUCCI, A. (2005). *Risk and asset allocation*. Springer Finance. Springer-Verlag, Berlin.

PISIER, G. (1986). Probabilistic methods in the geometry of banach spaces. In *Probability and Analysis* (LETTA, G. and PRATELLI, M., editors), volume 1206 of *Lecture Notes in Mathematics*, pp. 167–241. Springer Berlin / Heidelberg. URL http://dx.doi.org/10.1007/BFb0076302. 10.1007/BFb0076302.

SCHECHTMAN, G. and ZINN, J. (2000). Concentration on the l_p^n ball. In *Geometric aspects of functional analysis*, volume 1745 of *Lecture Notes in Math.*, pp. 245–256. Springer, Berlin.

SILVERSTEIN, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.* **55**, 331–339.

SILVERSTEIN, J. W. and BAI, Z. D. (1995). On the empirical distribution of eigenvalues of a class of large-dimensional random matrices. *J. Multivariate Anal.* **54**, 175–192.

STROOCK, D. W. (1993). *Probability theory, an analytic view*. Cambridge University Press, Cambridge.

TYLER, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *Ann. Statist.* **15**, 234–251. URL http://dx.doi.org/10.1214/aos/1176350263.