HIGH-DIMENSIONAL INFERENCE BY UNBIASED RISK ESTIMATION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Didier Chételat

May 2015

© 2015 Didier Chételat

ALL RIGHTS RESERVED

HIGH-DIMENSIONAL INFERENCE BY UNBIASED RISK ESTIMATION

Didier Chételat, Ph.D.

Cornell University 2015

This thesis derives natural and efficient solutions of three high-dimensional statistical problems by exploiting unbiased risk estimation. They exemplify a general methodology that provides attractive estimators in situations where classical theory is unsuccessful, and that could be exploited in many other problems.

First, we extend the classical James-Stein shrinkage estimator to the context where the number of covariates is larger than the sample size and the covariance matrix is unknown. The construction is obtained by manipulating an unbiased risk estimator and shown to dominate in invariant squared loss the maximum likelihood estimator. The estimator is interpreted as performing shrinkage only the random subspace spanned by the sample covariance matrix.

Second, we investigate the estimation of a covariance and precision matrix, and discriminant coefficients, of linearly dependent data in a normal framework. By bounding the difference in risk over classes of interest using unbiased risk estimation, we construct interesting estimators and show domination over naive solutions.

Finally, we explore the problem of estimating the noise coefficient in the spiked covariance model. By decomposing an unbiased risk estimator and minimizing its dominant part using calculus of variations, we obtain an estimator in closed form that approximates the optimal solution. Several attractive properties are proven about the proposed construction. We conclude by showing that the associated spiked covariance estimators possess excellent behavior under the Frobenius loss.

BIOGRAPHICAL SKETCH

Didier Chételat was born and raised in Québec. He received his bachelor's degree in Mathematics from McGill University in 2010. Following this, he moved to the United States to attend Cornell University, where he received his master's and doctorate degrees in Statistics in 2013 and 2015. His advisor was Martin T. Wells. He subsequently moved back to Montréal as an assistant professor in the department of Decision Sciences at HEC Montréal.

To Natalia.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Martin Wells, for his continued guidance and support. His advising provided me the freedom to learn and discover new interests. I would also like to thank Johannes Lederer who, as a friend and collaborator, introduced me to new topics in a fresh way. I am also deeply grateful to Michael Nussbaum for his insight as a current collaborator and teacher. I am thankful to Florentina Bunea and Jacob Bien for their help and time as committee members. I was also lucky to count many fellow grad students as friends, and I learned much more from them than from any book or article. They include Will, Ben, Josh, Lucas, Jón, Sasha, Dave, Dan and Kerstin, among many others. Finally, but most importantly, I am deeply grateful to Natalia for her continued support and patience – I couldn't have done it without you!

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The past two decades have seen a phenomenal rise in very large datasets and have brought much attention to challenges specific to high-dimensional statistics. In this framework, traditionally successful methodologies to derive estimators, such as the method of maximum likelihood and the method of moments, often lead to inadequate solutions. For example, the resulting constructions may fail to reach the optimal mimimax rate, be inconsistent or, in "p greater than n" settings, simply fail to exist.

As a consequence, high-dimensional statistical estimation problems have mostly been dealt with on a case to case basis. Over time, a general methodology developed around the concept of regularized optimization, which relies on finding a plausible goodness of fit criterion and minimizing some regularized version. Despite its generality, this approach suffers from a few pitfalls. Problematically, regularization introduces tuning parameters that must be calibrated. In addition, the estimators rarely possess closed forms, and this can make the study of their theoretical properties difficult.

A popular but computationally expensive approach to the calibration problem is cross-validation. In its simplest form, the training set is divided into folds. For each fold, the model is trained on the complement and evaluated on the fold. The tuning parameter is chosen as the one that best performed on average.

Recently, there has been a revival of interest in alternative approaches based on minimizing an unbiased risk estimator (URE). These methods usually require strong parametric assumptions and are limited to predictive tasks. However, within these restrictions, these methods usually offer superior performance for tuning parameter calibration compared to cross-validation, and are computationally light.

Moreover, the selected tuning parameters are more amenable to theoretical analysis.

This dissertation consists of three chapters that, in different ways, exploits this idea further. The usual approach is to derive in some way a parametrized estimator and calibrate it using unbiased risk estimation. The three chapters cut the middle step: in every project we derive estimators *by manipulating the unbiased risk estimator itself*. The last chapter is the most sophisticated example, in which an estimator is obtained by calculus of variations on an appropriate unbiased risk estimator. But in every case, the resulting estimators turn out to yield big gains in performance and possess closed forms, which makes them computationally trivial and amenable to theoretical analysis. These three chapters exemplify how, in our opinion, the unbiased risk minimization approach could succeed as a general methodology for deriving high-dimensional estimators, in the same way that regularization has become.

The second chapter concerns the extension of the James-Stein phenomenon to the high-dimensional setting where the number of covariates exceed the sample size. In James and Stein [1961], the natural solution to arguably the simplest problem in multivariate statistics, the estimation of the mean of a p-dimensional normal distribution, was shown inadmissible when $p > 3$. The authors derived an estimator by an empirical Bayes argument and showed that it dominated the maximum likelihood estimator. The key to the argument was to express the risk of the James-Stein estimator using an unbiased risk estimator, and show that its difference with the risk of the naive estimator had to be negative.

Using a similar manipulation of an unbiased risk estimator, we derive a proper large-scale analogue of this estimator. Our construction can be thought of as performing shrinkage only on the random subspace spanned by the sample covariance

matrix, which is n-dimensional. In this way, second-moment information is used to bring the $p > n$ setting down to a classical first-moment problem, in a novel way. This suggests that novel gains in performance in regularized methods could be realized through a similar projection.

The third chapter concerns the estimation of covariance matrices, precision matrices and discriminant coefficients in a singular multivariate normal model. Using recent results of Tsukuma and Kubokawa [2014], we construct an unbiased risk estimator for a loss for each problem. We then show how to improve over naive estimators over different classes by bounding the difference in risk between the estimators and minimizing this bound to obtain a positive gain.

The fourth chapter concerns covariance matrix estimation. A common theoretical setting to study high-dimensional principal components analysis (PCA) is the spiked model of Johnstone [2001]. An important quantity that must be estimated is the smallest eigenvalue of the covariance matrix, also known as the noise. A closely related problem is the estimation of the spiked covariance matrix itself under Frobenius loss, which heavily relies on the estimation of its noise.

An interesting aspect of the second problem is that it admits an unbiased risk estimator for a large class of interesting estimators. We propose to minimize this URE with respect to the noise using calculus of variations. Neglecting the asymptotically negligible part of the optimum that depends on the truth, we obtain a closed form estimator that performs well in practice. Moreover, we prove that it is consistent, essentially achieves the minimax rate and is almost asymptotically normal.

The remainder of this introduction will elaborate on the generic problem of model calibration that underlies each chapter. We hope these sections can provide the reader a better understanding of the wider context of this work.

1.1 Framework

A statistical model is some collection of probability measures $\mathbb{P}_\theta$ with parameters $\theta \in \Theta$. One has usually at disposal independent and identically distributed data $X_1, ..., X_n \sim \mathbb{P}_\theta$ for some θ. Estimation consists in recovering the true parameter θ that generated our sample. The problem is high-dimensional when the probability measures $\mathbb{P}_\theta$ live on $\mathbb{R}^p$ for p of the same order, or larger than, the sample size n. By high-dimensional asymptotics we mean behavior of some underlying statistical object as both p and n tend to the limit to infinity.

An estimator is some function of the sample $\hat{\theta}(X_1, ..., X_n)$ that is functionally independent of θ, but is used to approximate it. As there are many possible estimators, one must choose some criterion $\mathrm{L}(\hat{\theta}, \theta) : \Theta \times \Theta \mapsto (0, \infty)$ to measure performance, called a loss function. The loss of an estimator is a random quantity, so it is common to focus on the risk of an estimator, defined as $\mathrm{R}(\hat{\theta}, \theta) = \mathrm{E}_\theta \left[\mathrm{L}(\hat{\theta}, \theta) \right]$. An estimator $\hat{\theta}_1$ is said to dominate some other estimator $\hat{\theta}_2$ if $\mathrm{R}(\hat{\theta}_1, \theta) \leq \mathrm{R}(\hat{\theta}_2, \theta)$ for all θ, with this inequality strict for some θ.

Instead of a single estimator, it is common to consider a family of estimators $\hat{\theta}_\lambda$, where $\lambda \in \Lambda$ is called a tuning parameter. The tuning space Λ is often finite or countable. Once the family is chosen to estimate θ, it is common to choose λ in a data-dependent way $\hat{\lambda}$ to minimize the risk, a step called calibration or selection. The final estimator is $\hat{\theta}_{\hat{\lambda}}$.

The oracle tuning parameter is the data- and truth-dependent value that minimizes the loss,

$$\lambda^* = \lambda^*(X_1, ..., X_n, \theta) = \arg\min_\lambda \mathrm{L}(\hat{\theta}_\lambda, \theta),$$

and the oracle of the family is the truth-dependent estimator that minimizes the loss, $\hat{\theta}_* = \hat{\theta}_{\lambda^*(\theta)}$. The risk of the oracle puts a lower bound on the quality of the

estimation within the family, called the *approximation error*,

$$\text{Approximation error at } \theta = R(\hat{\theta}_*, \theta) \;\leq\; R(\hat{\theta}_{\hat{\lambda}}, \theta).$$

Since it is usually impossible for a data-dependent scheme $\hat{\lambda}$ to achieve the oracle risk, there is additional *estimation error*

$$\text{Estimation error at } \theta = R(\hat{\theta}_{\hat{\lambda}}, \theta) - R(\hat{\theta}_*, \theta) \geq 0,$$

so that $R(\hat{\theta}_{\hat{\lambda}}, \theta) = \text{Approximation error} + \text{Estimation error}$. Good calibration schemes minimize the estimation error.

1.2 Calibration as risk estimation

There are many approaches to parameter calibration, but two are particularly popular: unbiased risk estimator minimization and cross-validation. In both cases, the scheme selects the tuning parameter by minimizing an estimator of the risk.

The first scheme relies on the existence of an unbiased risk estimator for the family $\hat{\theta}_\lambda$, that is, a functional $U(\hat{\theta}_\lambda)$ independent of θ such that

$$R(\hat{\theta}_\lambda, \theta) = \mathrm{E}_\theta \left[U(\hat{\theta}_\lambda) \right] \qquad\qquad \forall \theta \in \Theta.$$

Examples include Mallow's C_p and Akaike's Information Criterion. Since $U(\hat{\theta}_\lambda)$ does not depend on θ, the scheme suggests to take

$$\hat{\lambda}^{\mathrm{U}} = \arg\min_\lambda U(\hat{\theta}_\lambda).$$

The heuristic argument that underlies this choice is that

$$R(\hat{\theta}_{\hat{\lambda}^{\mathrm{U}}}, \theta) \approx \mathrm{E}_\theta \left[U(\hat{\theta}_{\hat{\lambda}^{\mathrm{U}}}) \right] = \mathrm{E}_\theta \left[\min_\lambda U(\hat{\theta}_\lambda) \right] \approx \mathrm{E}_\theta \left[\min_\lambda \mathrm{L}(\hat{\theta}_\lambda, \theta) \right] = R(\hat{\theta}_*, \theta).$$

Therefore, minimizing the unbiased risk estimator should yield performance comparable to the oracle as long as the approximations are accurate.

In contrast, the second scheme relies on the existence of a contrast function, that is, a function γ such that

$$\mathrm{L}\Big(\hat{\theta}, \theta\Big) = \mathrm{E}_X\left[\gamma(\hat{\theta}, X)\right]$$

for $X \sim \mathbb{P}_\theta$, $X \perp\!\!\!\perp X_1, ..., X_n$. Cross-validation proposes to estimate the risk by repeatedly dividing the data into training and validation sets, namely, by dividing the n-sample into k folds of size n/k. Popular choices include ten-fold ($k = 10$), five-fold ($k = 5$) and leave-one-out ($k = 1$). If $I_1, ..., I_k \subset [n]$ represent the k folds that partition the sample, i.e. so that $I_i \cap I_j = \emptyset$ for $i \neq j$ and $\bigcup_{i=1}^{k} U_i = [n]$, the cross-validative risk estimator is

$$CV(\hat{\theta}_\lambda) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j \in I_i} \gamma\Big(\hat{\theta}_\lambda\big(X_l \,|\, l \in I_i^C\big), X_j\Big).$$

Now, quite often $\mathrm{E}\Big[CV(\hat{\theta}_\lambda)\Big] \to \mathrm{R}(\hat{\theta}_\lambda, \theta)$ as n and k jointly tend to infinity in some way. Since $CV(\hat{\theta}_\lambda)$ does not depend on θ, this suggests to take

$$\hat{\lambda}^{\mathrm{CV}} = \arg\min_{\lambda} CV(\hat{\theta}_\lambda)$$

which is the cross-validated tuning parameter. Similarly to the previous scheme, the heuristic argument that underlies this choice is that

$$\mathrm{R}(\hat{\theta}_{\hat{\lambda}^{\mathrm{CV}}}, \theta) \approx \mathrm{E}_\theta\left[CV(\hat{\theta}_{\hat{\lambda}^{\mathrm{CV}}})\right] = \mathrm{E}_\theta\left[\min_\lambda CV(\hat{\theta}_\lambda)\right] \approx \mathrm{E}_\theta\left[\min_\lambda \mathrm{L}(\hat{\theta}_\lambda, \theta)\right] = \mathrm{R}(\hat{\theta}_*, \theta).$$

Therefore, minimizing the cross-validation risk estimator should yield performance comparable to the oracle as long as the approximation is accurate.

In both cases, we choose the tuning parameter by minimizing an estimator of the risk $\mathrm{R}(\hat{\theta}_\lambda, \theta)$. In the first scheme, the estimator $U(\hat{\theta}_\lambda)$ is unbiased, in the sense that $\mathrm{E}\Big[U(\hat{\theta}_\lambda)\Big] = \mathrm{R}(\hat{\theta}_\lambda, \theta)$. In contrast, the cross-validation risk estimator $CV(\hat{\theta}_\lambda)$ is generally biased, $\mathrm{E}\Big[CV(\hat{\theta}_\lambda)\Big] \neq \mathrm{R}(\hat{\theta}_\lambda, \theta)$. This leads to some kind of second-order estimation theory: if the performance of risk estimators is measured in terms of

the mean-squared error $\left[\hat{R}(\hat{\theta}_\lambda) - \mathrm{R}(\hat{\theta}_\lambda, \theta)\right]^2$, then whether cross-validation is more effective at measuring the risk than URE minimization depends on whether the variance of $CV(\hat{\theta}_\lambda)$ is small enough to compensate its bias. This varies according to the problem at hand.

In practice, the functions $\lambda \to U(\hat{\theta}_\lambda)$ and $\lambda \to CV(\hat{\theta}_\lambda)$ can be very difficult to minimize. They are often non-convex, and even if they are differentiable it might be very difficult to compute their gradient. Consequently, tuning parameter spaces are usually kept finite or one-dimensional, and multiple tuning parameters are best avoided.

In this regard, expressions that minimize risk estimators in closed form are especially valuable, since they automatically approximate the oracle yet are easy to compute. This is especially the case in contexts where little methodology to derive estimators exists, such as in high-dimensional statistics.

CHAPTER 2

IMPROVED MULTIVARIATE NORMAL MEAN ESTIMATION WITH UNKNOWN COVARIANCE WHEN P IS GREATER THAN N

2.1 Introduction

Suppose a p-dimensional random vector X is observed which is normally distributed, with mean vector θ and unknown positive definite covariance matrix Σ, and we wish to estimate θ under the invariant quadratic loss

$$L(\theta, \delta) = (\delta - \theta)'\Sigma^{-1}(\delta - \theta). \tag{2.1}$$

Since the covariance matrix Σ is unknown, a random matrix S is observed along with X, which is assumed to be independent of X, and has a Wishart distribution with n degrees of freedom, where $p > n$. In high-dimensional estimation problems, where p, the number of features, is nearly as large as, or larger than, n, the number of observations, the ordinary least squares estimator does not typically provide a satisfactory estimate of θ.

Modern data sets are increasingly becoming characterized by a number of features that is much larger than the number of sample units (large-p, small-n) in contrast to classical data sets where the number of sample units is often much larger than the number of random variables (small-p, large-n). Modern applications in the $p > n$ setting include examples from microarrays, association mapping, proteomics, radiology, biomedical imaging, signal processing, climate modeling, and finance. For instance, in the case of microarray data, the dimensionality is frequently in thousands or beyond, while the sample size is typically in the order of tens. The large-p, small-n scenario poses challenges in most inferential settings. We are considering a canonical setting. For the usual multivariate location-scale

estimation problem let $W = (W_1, ..., W_p)$ denote an $N \times p$ matrix of data (N is the number of observations and p the number of features) where W_i are taken from a p-dimensional normal distribution with mean vector θ and covariance matrix Ξ. In this article we let the X and S be the sample mean and covariance of the features, respectively. In the context of this notation, $\Sigma = N^{-1}\Xi$ and $n = N - 1$.

The usual estimator under invariant quadratic loss is $\delta_0(X) = X$. It is minimax and admissible when $p \leq 2$ and $p \leq n$. However, when $p \geq 3$ and $p \leq n$, $\delta_0(X)$ remains minimax but is no longer admissible. Explicit improvements are known in the multivariate normal case [James and Stein [1961], Berger and Bock [1976], Berger et al. [1977], Gleser [1979], Berger and Haff [1983], Gleser [1986]] and in the case of elliptically symmetric distribution [Srivastava [1989], Fourdrinier et al. [2003]].

In this article we primarily concentrate on the case $p > n$ and construct a class of estimators, depending on the sufficient statistics (X, S), of the form

$$\delta(X, S) = X + g(X, S) \tag{2.2}$$

which dominate $\delta_0(X)$ under invariant quadratic loss. Note that, although the loss in (2.1) is invariant, the estimate in (2.2) may not be (except for $\delta_0(X)$). This class generalizes several estimators studied previously for the multivariate normal distribution to the $p \leq n$ setting [James and Stein [1961], Berger and Bock [1976], Berger et al. [1977], Gleser [1979], Berger and Haff [1983], Gleser [1986]]. Examples of estimators we study here in this setting extend the class of so-called Baranchik estimators and includes a new high dimensional James-Stein estimator

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

where $0 \leq a \leq \frac{2(n-2)}{p-n+3}$ and S^+ is the Moore-Penrose inverse of S.

The estimation of the inverse covariance matrix, namely the precision matrix Σ^{-1}, of a multivariate normal distribution has been an important problem in

practical situations as well as from a theoretical perspective. But, when $p > n$, the Wishart-distributed sample covariance matrix is singular; in this case, one is tempted to construct estimators using the Moore-Penrose generalized inverse S^+. Recently there has been an increased interest in the problem of estimating the covariance matrix of large dimension given variables of dimension larger than the number of observations [Bickel and Levina [2008a], Aspremont et al. [2008], Konno [2009], Ledoit and Wolf [2004], Levina et al. [2008], Rothman et al. [2008]].

Our method of proof relies on an unbiased estimator of risk difference, say $\rho(X, S)$. Specifically, we show that, for $g(X, S)$ of the form $-\frac{r(X'S^+X)SS^+}{X'S^+X}X$, the estimator $\delta(X, S) = X + g(X, S)$ dominates X provided $\rho(X, S) \leq 0$. In the next section we present the main results and their proofs are given in Section 2.3. We need Stein's integration-by-parts identity [Stein [1981]] and the so-called Stein-Haff identity for the singular Wishart distribution. The Stein-Haff identity was derived by Haff [1979a] and Stein [1977] for the full rank Wishart distribution. A similar identity for the elliptically contoured model has been given by Fourdrinier et al. [2003]. We make some concluding comments in Section 2.4.

2.2 Main results

Let X be a random vector distributed as $N_p(\theta, \Sigma)$ with unknown θ and Σ. Suppose an estimator of Σ is available, say $S \sim \text{Wishart}_p(n, \Sigma)$, with S independent of X. By definition of the Wishart distribution, we can write $S = Y'Y$ for some matrix normal $Y \sim N_{n\times p}(0, I \otimes \Sigma)$. An elementary property of this distribution is that S is (almost surely) invertible if $p \leq n$, and (almost surely) singular if $p > n$ [cf. Srivastava and Khatri [1979]].

An usual estimator of θ is $\delta^0(X, S) = X$; however, it turns out that this estimator is inadmissible under quadratic loss. If some estimator $S \sim \text{Wishart}_p(n, \Sigma)$ is

available, with $n \geq p \geq 3$, δ^0 is dominated by the so-called James-Stein estimator

$$\delta^{JS}(X, S) = \left(1 - \frac{(p-2)/(n-p+3)}{X'S^{-1}X}\right) X.$$

The main contribution of this article is to extend this type of result to a more general class of estimators in the $p > n$ setting.

For some positive, bounded, and differentiable function $r : \mathbb{R} \to \mathbb{R}$, define the Baranchik-type estimator

$$\begin{aligned} \delta_r(X, S) &= \left(I - \frac{r(X'S^+X)SS^+}{X'S^+X}\right) X \\ &= X + g(X, S) \end{aligned} \tag{2.3}$$

where I is the identity matrix and S^+ denotes the Moore-Penrose inverse of S. This estimator generalizes the usual Baranchik [1970] estimator to the unknown covariance setting for $p > n$.

Theorem 1. *Let* $\min(p, n) \geq 3$. *Suppose that:*

(i) r *satisfies* $0 \leq r \leq \frac{2\,(\min(n,p)-2)}{n+p-2\,\min(n,p)+3}$;

(ii) r *is nondecreasing; and*

(iii) r' *is bounded.*

Then under invariant quadratic loss, δ_r *dominates* δ^0.

Throughout the article we will use the expression $tr(SS^+)$, which of course equals $\min(n, p)$. This notation allows us to simultaneously handle both the $p > n$ and $n \geq p$ cases. The condition $\min(p, n) \geq 3$ merely guarantees that condition (i) of Theorem 1 holds for some r and is reminiscent of the dimension cut-off in classical Stein estimation.

Proof. The hypotheses of the theorem imply that r is differentiable almost everywhere. Under invariant quadratic loss, the difference in risk between δ_r and δ^0 is given by

$$\begin{aligned}\Delta_\theta =& E_\theta\left[(X+g(X,S)-\theta)'\Sigma^{-1}(X+g(X,S)-\theta)\right]\\ &- E_\theta\left[(X-\theta)'\Sigma^{-1}(X-\theta)\right]\\ =& 2E_\theta\left[g(X,S)'\Sigma^{-1}(X-\theta)\right]+E_\theta\left[g(X,S)'\Sigma^{-1}g(X,S)\right]. \end{aligned} \tag{2.4}$$

In order to show the domination result we need to show that under the sufficient conditions on r, (2.4) is nonpositive for all θ. First, for the left most term of (2.4) it can be shown that

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

Fourdrinier et al. [2003] give a more general form of this result in their Lemma 1(i); it is essentially an extension of Stein's classical integration by parts identity. By using Lemma 2 in Section 3, we have that

$$\begin{aligned}2E_\theta\left[\mathrm{div}_X g(X,S)\right] &= -2E_\theta\left[\mathrm{div}_X \frac{r(X'S^+X)SS^+X}{X'S^+X}\right]\\ &= -2E_\theta\left[2r'(X'S^+X)+r(X'S^+X)\frac{\mathrm{tr}(SS^+)-2}{X'S^+X}\right]. \end{aligned} \tag{2.5}$$

For the right term of (2.4), we have

$$\begin{aligned}E_\theta\left[g(X,S)'\Sigma^{-1}g(X,S)\right] &= \mathrm{tr}(E_\theta\left[g(X,S)'\Sigma^{-1}g(X,S)\right])\\ &= E_\theta\left[\mathrm{tr}\left(g(X,S)'\Sigma^{-1}g(X,S)\right)\right]\\ &= E_\theta\left[\mathrm{tr}\left(\Sigma^{-1}g(X,S)g(X,S)'\right)\right]\\ &= E_\theta\left[\mathrm{tr}\left(\Sigma^{-1}Sr^2(X'S^+X)\frac{S^+XX'S^+S}{(X'S^+X)^2}\right)\right]. \end{aligned}$$

Through Lemma 3 in Section 3, we will find

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

The finiteness of the risk of δ_r is guaranteed to hold by Theorem 2 in Section 3 for all p and n.

Now applying Lemma 1 in Section 3, we find

$$\begin{aligned}
&E_\theta\left[n\operatorname{tr}\left(r^2(X'S^+X)\frac{S^+XX'S^+S}{(X'S^+X)^2}\right)\right.\\
&\quad\left.+\operatorname{tr}\left(Y'\nabla_Y\left\{r^2(X'S^+X)\frac{SS^+XX'S^+}{(X'S^+X)^2}\right\}\right)\right]\\
=&E_\theta\left[n\frac{r^2(X'S^+X)}{X'S^+X}-4r(X'S^+X)r'(X'S^+X)\right.\\
&\quad\left.+r^2(X'S^+X)\frac{p-2\operatorname{tr}(SS^+)+3}{X'S^+X}\right]\\
=&E_\theta\left[r^2(X'S^+X)\frac{n+p-2\operatorname{tr}(SS^+)+3}{X'S^+X}-4r(X'S^+X)r'(X'S^+X)\right]. \qquad (2.6)
\end{aligned}$$

Replacing (2.5) and (2.6) back into (2.4), we obtain

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

Since r is nonnegative and nondecreasing, $-4r'(X'S^+X)\{1+r(X'S^+X)\}\leq 0$

follows. Finally, for the X and S such that $r(X'S^+X) \neq 0$,

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

Therefore, under the three sufficient conditions on r, it follows that $\Delta_\theta \leq 0$ for any θ, that is, the domination result holds. □

In the $p > n$ setting, we obtain the following two corollaries.

Corollary 1. *For $p > n \geq 3$, δ_r dominates δ^0 under invariant quadratic loss for all r nondecreasing, differentiable and satisfying*

$$0 \leq r \leq \frac{2(n-2)}{p-n+3}. \tag{2.7}$$

Corollary 2 (James-Stein estimator with large p and small n)**.** *For $p > n \geq 3$ and $a \in \mathbb{R}$, the James-Stein-like estimator*

$$\delta_a^{JS}(X,S) = \left(I - \frac{aSS^+}{X'S^+X}\right)X \tag{2.8}$$

dominates δ^0 under invariant quadratic loss for all

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

Note that if p is only moderately larger than n, Corollary 1 implies that one can construct an estimator with substantial improvement over δ^0. However, in the ultra-high dimensional setting the denominator in (2.7) could be quite large and consequently the amount of improvement over δ^0 could be quite small. The estimator in (2.8) generalizes the classical James-Stein with unknown covariance matrix,

$$\delta_a^{JS}(X,S) = \left(1 - \frac{a}{X'S^{-1}X}\right)X$$

which is restricted to the case $p \leq n$, for $a \in \mathbb{R}_+$. In this setting, this result is consistent with previously bounds in Fourdrinier et al. [2003] (where $n-1$ is used instead of our n.)

2.3 Technical results and proofs

It remains to clarify several of the somewhat technical computations used in the proof of Theorem 1. We provide them in this section; these computations are likely to be of independent interest and showcase several technical maneuvers that the reader could find useful in dealing with singular Wishart matrices.

Proposition 1. *Let Y be an $n \times p$ matrix, $S = Y'Y$, S^+ be its Moore-Penrose pseudo-inverse, X a p-vector, and $F = X'S^+X$. It then follows that*

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

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

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

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

Proof. First, notice that from the usual chain-rule that

$$\left\{ \frac{\partial S}{\partial Y_{\alpha\beta}} \right\}_{kl} = \frac{\partial}{\partial Y_{\alpha\beta}} S_{kl} = \frac{\partial}{\partial Y_{\alpha\beta}} \sum_q Y_{qk} Y_{ql} = \delta_{\beta k} Y_{\alpha l} + \delta_{\beta l} Y_{\alpha k}.$$

This shows (i).

Let A be a symmetric matrix and $t \in \mathbb{R}$, then

$$\frac{\partial A^+}{\partial t} = -A^+ \frac{\partial A}{\partial t} A^+ + (I - AA^+) \frac{\partial A}{\partial t} A^+ A^+ + A^+ A^+ \frac{\partial A}{\partial t} (I - AA^+).$$

This result was, it seems, first proved in Golub and Pereyra [1973], as their Theorem 4.3, but can be found in standard textbooks on elementary linear algebra. Also, again for A symmetric, we have $AA^+ = A^+A$ and $A(I - AA^+) = (I - AA^+)A = A^+(I - AA^+) = (I - AA^+)A^+ = 0$. This easily follows from elementary properties of the Moore-Penrose pseudoinverse.

Since $S = Y'Y$, notice through a singular value decomposition argument that $SS^+Y' = Y'$ and thus $(I - SS^+)Y' = 0$. Using (i) we find that

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

which gives (ii).

Using (i) we have that for any conformable matrices A and B

$$\left(A \frac{\partial S}{\partial Y_{\alpha\beta}} B\right)_{kl} = \sum_{i,j} A_{ki} \frac{\partial S}{\partial Y_{\alpha\beta}}_{ij} B_{jl}$$

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

Therefore, using again $(I - SS^+)Y' = 0$:

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

which gives (iii). □

Lemma 1. *Under the hypotheses of Theorem 1 we have*

$$
\begin{aligned}
&tr \left(Y' \nabla_Y \left\{ r^2(X'S^+X) \frac{SS^+XX'S^+}{(X'S^+X)^2} \right\} \right) \\
&\quad = -4r(X'S^+X) r'(X'S^+X) + r^2(X'S^+X) \frac{p - 2tr(SS^+) + 3}{X'S^+X}
\end{aligned}
$$

where ∇_Y *is interpreted as the matrix with components* $(\nabla_Y)_{ij} = \frac{\partial}{\partial Y_{ij}}$.

Proof. To simplify computations in what will follows, we let $F \equiv X'S^+X$. We then have

$$\begin{aligned}
&\left[Y'\nabla_Y \left\{r^2(F)\frac{SS^+XX'S^+}{F^2}\right\}\right]_{ij} \\
=&\sum_{\alpha,\beta}(Y')_{i\alpha}\frac{\partial}{\partial Y_{\alpha\beta}}\left\{r^2(F)\frac{(SS^+XX'S^+)_{\beta j}}{F^2}\right\} \\
=&2\sum_{\alpha,\beta}(Y')_{i\alpha}r(F)r'(F)\frac{\partial F}{\partial Y_{\alpha\beta}}\cdot\frac{(SS^+XX'S^+)_{\beta j}}{F^2} \qquad (2.12)\\
&+\sum_{\alpha,\beta}(Y')_{i\alpha}r^2(F)\frac{\frac{\partial}{\partial Y_{\alpha\beta}}\{(SS^+XX'S^+)_{\beta j}\}}{F^2} \qquad (2.13)\\
&+\sum_{\alpha,\beta}(Y')_{i\alpha}r^2(F)\frac{-2\frac{\partial F}{\partial Y_{\alpha\beta}}(SS^+XX'S^+)_{\beta j}}{F^3}. \qquad (2.14)
\end{aligned}$$

To simplify (2.12) and (2.14) we apply Proposition 1 (ii) to get

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

Using this, we get for (2.12)

$$\begin{aligned}
&2\sum_{\alpha,\beta}(Y')_{i\alpha}r(F)r'(F)\frac{\partial F}{\partial Y_{\alpha\beta}}\cdot\frac{(SS^+XX'S^+)_{\beta j}}{F^2} \\
&\qquad =-4r(F)r'(F)\frac{(SS^+XX'S^+)_{ij}}{F} \qquad (2.15)
\end{aligned}$$

and (2.14) becomes

$$\begin{aligned}
&\sum_{\alpha,\beta}(Y')_{i\alpha}r^2(F)\frac{-2\frac{\partial F}{\partial Y_{\alpha\beta}}\cdot(SS^+XX'S^+)_{\beta j}}{F^3} \\
&\qquad =4r^2(F)\frac{(SS^+XX'S^+)_{ij}}{F^2}. \qquad (2.16)
\end{aligned}$$

This leaves the term (2.13) to analyze. Using Proposition 1 (iii):

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

Next applying this computation in (2.13), we obtain

$$\sum_{\alpha,\beta}(Y')_{i\alpha}r^2(F)\frac{\frac{\partial}{\partial Y_{\alpha\beta}}\{(SS^+XX'S^+)_{\beta j}\}}{F^2}$$
$$= (p-\operatorname{tr}(SS^+)-1)r^2(F)\frac{(SS^+XX'S^+)_{ij}}{F^2}$$
$$- r^2(F)\frac{(SS^+)_{ij}}{F}. \tag{2.17}$$

Now we can combine (2.15), (2.17) and (2.16) together to complete the proof. That is, we have

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

as desired. □

Lemma 2. *Under the hypotheses of Theorem 1 we have*

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

Proof. Again, to simplify computations, let us denote $X'S^+X$ by F. We find

$$\begin{aligned}
&\operatorname{div}_X\left\{r(F)\frac{SS^+X}{F}\right\}=\sum_i\frac{\partial}{\partial X_i}\left\{r(F)\frac{(SS^+X)_i}{F}\right\}\\
=&\sum_i r'(F)\frac{\partial F}{\partial X_i}\frac{(SS^+X)_i}{F}\\
&+r(F)\frac{\frac{\partial}{\partial X_i}\{(SS^+X)_i\}}{F}-r(F)\frac{\frac{\partial F}{\partial X_i}(SS^+X)_i}{F^2}
\end{aligned}$$

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

as desired. □

The following result is an extension of a result in Konno [2009]. This type of result was first obtained by Kubokawa and Srivastava [2008] and then was extended by Konno [2009]. In our generalization we make use of a divergence version of Stein's lemma that comes with somewhat weaker moment conditions, rather than the element-by-element assumptions in Konno [2009]. These weaker moment conditions allow us to cover the p equals n and $n+1$ cases.

Lemma 3. *Let $Y \sim N_{n\times p}(0, I_n \otimes \Sigma)$, let $S = Y'Y$ which has, by definition, a Wishart$_p(n, \Sigma)$ distribution, and let $G(S)$ be a $p \times p$ random matrix that depends on S. Let ∇_Y be interpreted as the matrix with components $(\nabla_Y)_{ij} = \frac{\partial}{\partial Y_{ij}}$, and for A the symmetric positive definite square root of Σ, define $\tilde{Y} = YA^{-1}$ and $H = AGA^{-1}$. Then*

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

under the conditions

$$
E\left[\left| div_{vec(\tilde{Y})} \cdot vec\left(\tilde{Y} H\right) \right| \right] < \infty \tag{2.18}
$$

where $vec(M)$ denotes the vectorization of a matrix M.

Proof. Define $\tilde{S} = \tilde{Y}'\tilde{Y} = A^{-1}SA^{-1}$. Notice that, by construction, $\tilde{Y} \sim N_{n\times p}(0, I_n \otimes I_p)$ which means, by definition of the matrix normal distribution, that $\text{vec}(\tilde{Y}) \sim N_{np}(0, I_{np})$. We can write

$$E\left[\text{tr}\left(\tilde{S}H\right)\right] = E\left[\sum_{\alpha,i,j} \tilde{Y}_{\alpha i}\tilde{Y}_{\alpha j}H_{ji}\right] = E\left[\text{vec}(\tilde{Y}) \cdot \text{vec}\left(\tilde{Y}H\right)\right].$$

Using the divergence form of Stein's lemma, which can be found in Lemma A.1 in Fourdrinier and Strawderman [2003], we obtain, under the moment conditions outlined in (2.18)

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

This last expression can be expressed in a compact matrix form as

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

Finally, we notice

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

which concludes the proof. □

Theorem 2. *Let $Y \sim N_{n\times p}(0, I_n \otimes \Sigma)$ and for A the symmetric positive definite square root of Σ, let $\tilde{Y} = YA^{-1}$. Let r be any bounded differentiable nonnegative*

function $r : \mathbb{R} \to [0, C_1]$ *with bounded derivative* $|r'| \leq C_2$. *Define*

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

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

$$E\left[\left|div_{vec(\tilde{Y})}\ vec\left(\tilde{Y}H\right)\right|\right] < \infty. \tag{2.19}$$

Proof. We first compute $\mathrm{div}_{\mathrm{vec}(\tilde{Y})}\ \mathrm{vec}\left(\tilde{Y}H\right)$. As always, to ease notation we shall write $F = X'S^+X$. We have

$$\begin{aligned}
\mathrm{div}_{\mathrm{vec}(\tilde{Y})}\ \mathrm{vec}\left(\tilde{Y}H\right) =& \sum_{\alpha,i,j}\frac{\partial}{\partial\tilde{Y}_{\alpha i}}\left\{\tilde{Y}_{\alpha j}H_{ji}\right\} \\
=& n\sum_i H_{ii} + \sum_{\alpha,j}\tilde{Y}_{\alpha j}\frac{\partial H_{ji}}{\partial\tilde{Y}_{\alpha i}} \\
=& n\sum_i H_{ii} + \sum_{\alpha,\beta,i,j}\tilde{Y}_{\alpha j}A_{\beta i}\frac{\partial}{\partial Y_{\alpha\beta}}\left\{r^2(F)\frac{\{AS^+XX'SS^+A^{-1}\}_{ji}}{F^2}\right\} \\
=& n\sum_i H_{ii} + \sum_{\alpha,\beta,i,j}\tilde{Y}_{\alpha j}A_{\beta i}\cdot \\
&\left\{2r(F)r'(F)\frac{\partial F}{\partial Y_{\alpha\beta}}\frac{\{AS^+XX'SS^+A^{-1}\}_{ji}}{F^2}\right. && (2.20) \\
&+ \frac{r^2(F)}{F^2}\sum_{k,l}A_{jk}\frac{\partial\{S^+XX'SS^+\}_{kl}}{\partial Y_{\alpha\beta}}A^{-1}_{li} && (2.21) \\
&\left.-r^2(F)\left\{AS^+XX'SS^+A^{-1}\right\}_{ji}\frac{2\frac{\partial F}{\partial Y_{\alpha\beta}}}{F^3}\right\}. && (2.22)
\end{aligned}$$

We simplify each part of the expression. For (2.20), using Proposition 1 (ii), we find

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

$$
\begin{aligned}
&= -4\frac{r(F)r'(F)}{F^2}\left(X'S^+Y'YA^{-1}AS^+XX'SS^+A^{-1}AS^+X\right)\\
&\quad +4\frac{r(F)r'(F)}{F^2}\left(X'S^+S^+Y'YA^{-1}AS^+XX'SS^+A^{-1}A(I-SS^+)X\right)\\
&= -4r(F)r'(F). \qquad (2.23)
\end{aligned}
$$

Similarly, for (2.22)

$$
\begin{aligned}
&\sum_{\alpha,\beta,i,j}\tilde{Y}_{\alpha j}A_{\beta i}r^2(F)\left\{AS^+XX'SS^+A^{-1}\right\}_{ji}\frac{2\frac{\partial F}{\partial Y_{\alpha\beta}}}{F^3}\\
&\quad =4\frac{r^2(F)}{F^3}\sum_{\alpha,\beta,i,j}(X'S^+Y')_\alpha\tilde{Y}_{\alpha j}\left\{AS^+XX'SS^+A^{-1}\right\}_{ji}A_{i\beta}(S^+X)_\beta\\
&\quad =4\frac{r^2(F)}{F^3}\left(X'S^+Y'YA^{-1}AS^+XX'SS^+A^{-1}AS^+X\right)\\
&\quad =4\frac{r^2(F)}{F}. \qquad (2.24)
\end{aligned}
$$

This leaves us with (2.21). Using Proposition 1 (iii) we obtain

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

$$
\begin{aligned}
&A_{jk}(S^+S^+Y)_{k\alpha}\tilde{Y}_{\alpha j}A_{i\beta}((I-SS^+)XX'SS^+)_{\beta l}A^{-1}_{li}\\
&-\tilde{Y}'_{j\alpha}(YS^+XX'SS^+)_{\alpha l}A^{-1}_{li}A_{i\beta}S^+_{\beta k}A_{kj}\\
&-A_{jk}(S^+Y)_{k\alpha}\tilde{Y}_{\alpha j}A_{i\beta}(S^+XX'SS^+)_{\beta l}A^{-1}_{li}\\
&+\tilde{Y}'_{j\alpha}(YS^+S^+XX'SS^+)_{\alpha l}A^{-1}_{li}A_{i\beta}(I-SS^+)_{\beta k}A_{kj}\\
&+\tilde{Y}'_{j\alpha}(YS^+)_{\alpha l}A^{-1}_{li}A_{i\beta}(XX'S^+)_{\beta k}A_{kj}\\
&+A_{jk}(S^+XX'Y')_{k\alpha}\tilde{Y}_{\alpha j}A_{i\beta}(S^+)_{\beta l}A^{-1}_{li}\\
&+A_{jk}(S^+XX'S^+Y')_{k\alpha}\tilde{Y}_{\alpha j}A_{i\beta}(I-SS^+)_{\beta l}A^{-1}_{li}\\
&-\tilde{Y}'_{j\alpha}(YS^+)_{\alpha l}A^{-1}_{li}A_{i\beta}(SS^+XX'S^+)_{\beta k}A_{kj}\\
&-A_{jk}(S^+XX'SS^+Y')_{k\alpha}\tilde{Y}_{\alpha j}A_{i\beta}(S^+)_{\beta l}A^{-1}_{li}\Big\}\\
=&\frac{r^2(F)}{F^2}\big\{\mathrm{tr}(AS^+S^+Y'YA^{-1})\cdot\mathrm{tr}(A(I-SS^+)XX'SS^+A^{-1})\\
&-\mathrm{tr}(A^{-1}Y'YS^+XX'SS^+A^{-1}AS^+A)\\
&-\mathrm{tr}(AS^+Y'YA^{-1})\mathrm{tr}(AS^+XX'SS^+A^{-1})\\
&+\mathrm{tr}(A^{-1}Y'YS^+S^+XX'SS^+A^{-1}A(I-SS^+)A)\\
&+\mathrm{tr}(A^{-1}Y'YS^+A^{-1}AXX'S^+A)\\
&+\mathrm{tr}(AS^+XX'Y'YA^{-1})\cdot\mathrm{tr}(AS^+A^{-1})\\
&+\mathrm{tr}(AS^+XX'S^+Y'YA^{-1})\mathrm{tr}(A(I-SS^+)A^{-1})\\
&-\mathrm{tr}(A^{-1}Y'YS^+A^{-1}ASS^+XX'S^+A)\\
&-\mathrm{tr}(AS^+XX'SS^+Y'YA^{-1})\mathrm{tr}(AS^+A^{-1})\big\}\\
=&\frac{r^2(F)}{F^2}\cdot\big\{-X'S^+X-\mathrm{tr}(SS^+)\cdot X'S^+X\\
&+X'S^+X+X'SS^+X\cdot\mathrm{tr}(S^+)\\
&-X'S^+X-X'SS^+X\cdot\mathrm{tr}(S^+)\big\}\\
=&-\frac{r^2(F)}{F}\Big(1+\mathrm{tr}(SS^+)\Big). \qquad (2.25)
\end{aligned}
$$

Having reexpressed $\mathrm{div}_{\mathrm{vec}(\tilde{Y})}\ \mathrm{vec}\left(\tilde{Y}H\right)$, we now need to bound it above. By virtue

of (2.23), (2.24) and (2.25), we have

$$\begin{aligned}
&E\left[\left|\mathrm{div}_{\mathrm{vec}(\tilde{Y})}\ \mathrm{vec}\left(\tilde{Y}H\right)\right|\right] \\
&= E\left[\left| n\mathrm{tr}(H) + 4\frac{r^2(F)}{F}\right.\right. \\
&\qquad \left.\left. +\ \left(1+\mathrm{tr}(SS^+)\right)\frac{r^2(F)}{F} - 4r(F)r'(F)\right|\right] \\
&\leq C_1^2\left|3-\mathrm{tr}(SS^+)+n\right| E\left[\frac{1}{F}\right] + 4C_1C_2. \qquad (2.26)
\end{aligned}$$

It only remains to show that $E\left[\frac{1}{F}\right]$ is finite. By definition of the Wishart matrix distribution we can define a $T \sim \mathrm{Wishart}_p(n, I_n)$ such that $S = ATA$. Let $T = H'DH$ be the spectral decomposition of T, with $D = \mathrm{diag}(\lambda_i)$. Write the eigenvalues of T^+ as λ_i^+, so that $D^{-1} = \mathrm{diag}(\lambda_i^+)$, and let $\lambda_{\min}^+$ be the smallest nonzero eigenvalue of T^+. The following two identities follow from Tian and Cheng [2004] [Theorem 1.1, equations (1.2) and (1.4)] and symmetry of T:

$$\begin{aligned}
(ATA)^+ &= (T^+TA)^+T^+(AT^+T)^+ \\
(T^+TA)^+(T^+T) &= (T^+TA)^+.
\end{aligned}$$

Using these identities we have

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

Applying Cauchy-Schwartz provides us the bound

$$X'(T^+TA)^+(T^+TA)X \leq\ X'(T^+TA)^+(AT^+T)^+XX'(AT^+T)(T^+TA)X$$

so that we then have

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

To ease notation let us write $Q = AT^+TA$ and $R = (T^+TA)^+(T^+TA)$. Collecting the results together we bound (2.26) by

$$\leq C_1^2 \left|3 - \mathrm{tr}(SS^+) + n\right| E\left[\frac{1}{\lambda^+_{\min}} \frac{X'QX}{X'RX}\right] + 4C_1C_2. \tag{2.27}$$

We now use some independence results. We can write the singular value decomposition of T as $T = H'DH$, but we can also write it as $T = H_1'D_1H_1$, where H_1 is semi-orthogonal ($H_1H_1' = I$), and D_1 is the matrix of the positive eigenvalues of T. If T has full rank (i.e., $n \geq p$) then this coincide with the singular value decomposition of T. In the full rank case, Srivastava and Khatri [1979] [Section 3.4, equation (3.4.3)] provides the joint density of H and $D = \mathrm{diag}(d_i)$ in the standard Wishart case (which applies to T) as

$$f_{H,D}(H, D) = C(p,n)|D|^{\frac{1}{2}(n-p-1)} \left[\mathrm{etr}\left(-\frac{1}{2}D\right)\right] \left[\prod_{i<j}(d_i - d_j)\right] g_p(H) \tag{2.28}$$

for constants $C(p,n)$ and functions g_p. Therefore, H and D are independent. In the rank-deficient case ($p > n$), Srivastava [2003] [Section 3] provides an equivalent expression which, in the singular Wishart case gives

$$f_{H_1,D_1}(H_1, D_1) = K(p,n)|D_1|^{\frac{1}{2}(p-n-1)} \left[\mathrm{etr}\left(-\frac{1}{2}D_1\right)\right] \left[\prod_{i<j}(d_i - d_j)\right] g_{n,p}(H_1) \tag{2.29}$$

for constants $K(p,n)$ and functions $g_{n,p}$ so again, we find H_1 and D_1 independent by factorization. Now, $\lambda^+_{\min}$ is a function, in the full rank case (resp. rank-deficient case), of only D^{-1} (resp. D_1^{-1}), and we can write $T^+T = H'H$ (resp.

$T^+T = H_1'H_1$), so $\lambda_{\min}^+$ and T^+T are independent. Being functions of S, they are also both independent of X. Now, the nonzero eigenvalues of T^+ are the inverses of the nonzero eigenvalues of T, a general fact about Moore-Penrose pseudo-inverses. Therefore, denoting the largest eigenvalue of T as $\lambda_{\max}$, we can split up the expectations in (2.27) and get the bound

$$\leq C_1^2 \left|3 - \operatorname{tr}(SS^+) + n\right| E\left[\lambda_{\max}\right] E\left[\frac{X'QX}{X'RX}\right] + 4C_1C_2. \tag{2.30}$$

Now, it follows from positive semi-definiteness of T that $E\left[\lambda_{\max}\right] \leq E\left[\operatorname{tr}(T)\right]$. If $n \geq p$, $\operatorname{tr}(T) \sim \chi^2_{pn}$ (cf. Muirhead [1982], Theorem 3.2.20) and so $E\left[\operatorname{tr}(T)\right] = pn < \infty$. If $p > n$, recall we can write $T = Z'Z$ for $Z \sim N_{n\times p}(0, I_n \otimes I_p)$ by definition of the Wishart distribution; and $ZZ' \sim \text{Wishart}_n(p, I_n)$ so that $\operatorname{tr}(T) = \operatorname{tr}(ZZ') \sim \chi^2_{pn}$; so again, $E\left[\operatorname{tr}(T)\right] = pn < \infty$. Therefore, in either case,$E\left[\lambda_{\max}\right] \leq pn < \infty$.

We still have to check that the expectation involving X, Q and R in (2.30) is finite. Let $r = \operatorname{rk}(R) = \operatorname{rk}(Q) = \operatorname{rk}(S)$ and write the spectral decomposition of (T^+TA) as $U\Lambda U'$, with $\Lambda = \operatorname{diag}(L, 0_{(p-r)})$ where L is the vector of the r nonzero eigenvalues of (T^+TA). Then $R = (T^+TA)^+(T^+TA) = U\operatorname{diag}(I_r, 0_{(p-r)})U'$; let us define the $p \times (p-r)$ matrix $E = U\left[0_{(p-r)\times r}\ I_{(p-r)}\right]'$ i.e. so that $RE = 0$ and E has full column rank $p - r$. Notice that $QE = AT^+TAU\left[0_{(p-r)\times r}\ I_{(p-r)}\right]' = AU\Lambda U'U\left[0_{(p-r)\times r}\ I_{(p-r)}\right]' = 0$. Since Q and R are symmetric positive semidefinite, we can use results in Magnus [1990] [Theorem 1(i) with $A = Q$ and $B = R$] to conclude that

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

This concludes the proof of the theorem. □

2.4 Numerical study

This section provides some numerical results to showcase the improvement in risk of the minimax estimator over the usual estimator. More precisely, we compared the James-Stein estimator in (2.8) given by

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

and the usual estimator $\delta^0 = X$ under invariant loss. (In addition, we considered the positive James-Stein estimator to be discussed in Section 2.5.) The empirical approximations of the invariant risk of these estimators were plotted for $p = 10, 20, 50$ and $n = \frac{p}{2}, p-1$. Three covariance matrix structures were considered:

Spiked A diagonal matrix with the first $p/2$ diagonal elements equal to 1, and the last $p/2$ equal to 10.

Autoregressive Autoregressive covariance matrices of the form

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho^2 & \\ \rho & 1 & \rho & \\ \rho^2 & \rho & 1 & \\ & & & \ddots \end{pmatrix}$$

for $\rho = 0.5$.

Block diagonal Block diagonal matrices with $p/2$ blocks of the form $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ for $\rho = 0.5$.

In all cases, the true mean was chosen as $\theta \propto (1, ..., 1)$.

We remind the reader that the risk of the trivial estimator is always p, regardless of θ or Σ. With this in mind, we see from Figure 2.1 that in all six scenarios the pattern of domination of the new estimator is similar to one of the usual James-Stein estimator. Also note that, as predicted by the theoretical results, the domination decreases as the smaller n tends to p.

2.5 Comments

An interesting property of the Moore-Penrose inverse is that for any A, AA^+ is the matrix that projects onto the subspace spanned by A (its column space.) It follows that the proposed generalized Baranchik estimator can be expressed as

$$\begin{aligned}\delta_r(X,S) &= (I - SS^+)X + \left(1 - \frac{r(X'S^+X)}{X'S^+X}\right) SS^+X \\ &= P_{S^\perp}X + \left(1 - \frac{r(X'S^+X)}{X'S^+X}\right) P_S X \end{aligned} \tag{2.31}$$

where $P_S = SS^+$ and $P_{S^\perp} = I - SS^+$ are the projection matrices onto the column space of S and its orthogonal complement, respectively. In terms of the kernel and image of the symmetric matrix S, $\mathrm{Ker}(P_{S^\perp}) = \mathrm{Im}(S)$ and $\mathrm{Im}(P_{S^\perp}) = \mathrm{Ker}(S^+)$. When $p > n$, this means we can interpret our estimator as applying shrinkage only on the component of X in the subspace spanned by our covariance matrix estimator S. In particular, note that the estimator $P_S\delta_r(X,S) = \left(1 - \frac{r(X'S^+X)}{X'S^+X}\right) P_S X$ dominates $P_S X$ under invariant loss function (1.1), since $R(P_S\delta_r,\theta) - R(P_S X,\theta) = R(\delta_r,\theta) - R(X,\theta) \geq 0$ if r satisfies the conditions of Theorem 1. This suggests there might be an easier, more abstract proof of Theorem 1, one not relying on brute computations but on the already known full rank S case, although we have not been able to obtain such a result.

A natural extension of the James-Stein estimator, δ_a^{JS} in (2.8), is a positive-

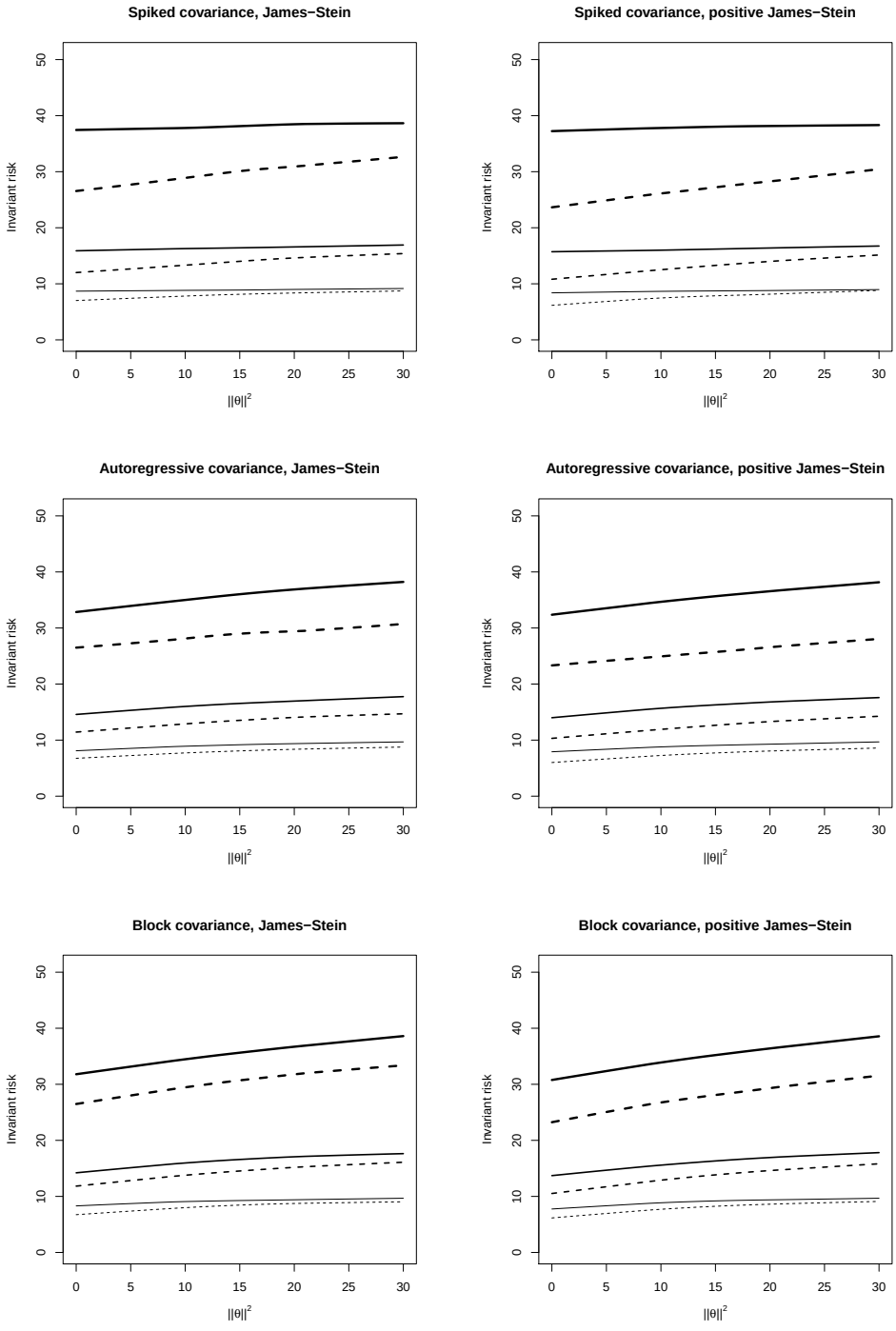


Figure 2.1: The risk function plots of δ_a^{JS} and $\delta_a^{\mathrm{JS}+}$ for $a = (n-2)/(p-n+3)$ are in the left and right columns, respectively. The lines, from thinnest to thickest, are for $p = 10, 20$ and 50. The solid and dashed lines are respectively for $n = p/2$ and $n = p - 1$.

part-type James-Stein estimator. The form of the estimator in (2.31) suggests

$$\delta_a^{\mathrm{JS}+} = \left(I - SS^+\right) X + \left(1 - \frac{a}{X'S^+X}\right)_+ SS^+X, \tag{2.32}$$

where $b_+ = \max(b, 0)$. Simulation evidence from Figure 2.1 suggests that for $a = (n-2)/(p-n+3)$, $\delta_a^{\mathrm{JS}+}$ dominates δ_a^{JS} under invariant loss.

One of the interesting differences between the $n > p$ and $p > n$ cases is the reversal of the roles of p and n. This is essentially due to the distribution of the singular values of S. Recall that for $S = ATA$, $T \sim W_p(n, I_n)$. We can write the singular value decomposition of T as $T = H'DH$, but we can also write it as $T = H_1'D_1H_1$, where H_1 is semi-orthogonal ($H_1H_1' = I$), and D_1 is the matrix of the positive eigenvalues of T. If T has full rank (i.e., $n \geq p$) this coincides with the singular value decomposition of T. In the full rank case the joint density of H and D is given in (2.28), whereas in the rank-deficient case ($p > n$) joint density is given by (2.29), from which stems the reversal of the roles of p and n.

In the heteroscedastic normal mean estimation problem, James and Stein [1961] used the loss function that was weighted by the inverse of the variances, and consequently the problem is essentially transformed to the homoscedastic case under ordinary squared error loss. Similarly in this article, we used the invariant loss function in (2.1), therefore skirting a somewhat subtle issue. In the heteroscedastic setting where there are differing coordinate variances, minimax estimation and Bayes (or empirical Bayes) estimates can be qualitatively different. It turns out that minimax estimators in general shrink most on the coordinates with smaller variances, while Bayes estimators shrink most on large variance coordinates. Brown [1975] shows that the James-Stein shrinkage estimator does not dominate X when the largest variance is larger than the sum of the rest. Moreover, Casella [1980] points out that the James-Stein shrinkage estimator may not be a desirable shrinkage estimator under heteroscedasticity even when it is minimax. Morris and Lysy

[2009] and Brown et al. [2012] give an excellent perspective on minimaxity of shrinkage estimator from Bayes and empirical Bayes points of view. Consequently, it would be of interest to examine the shrinkage patterns of the proposed estimates in the case of a non-invariant loss function and assess how well the invariant loss works for $p > n$ applications.

One can imagine an extension of the results of this chapter beyond the normal distribution setting. Consider a model with the joint density for (X, S) having the form

$$f\Big(\text{tr } \Sigma^{-1}[(X-\theta)(X-\theta)' + S]\Big) \tag{2.33}$$

where the $p \times 1$ location vector θ and the $p \times p$ scale matrix Σ are unknown. In the setting of $p \leq n$, Fourdrinier et al. [2003] and Kubokawa and Srivastava [2001] give some results on improved location estimation for elliptically symmetric distributions. For more on elliptical symmetry and the various choices of $f(\cdot)$ in (2.33), see Fang et al. [1990]; the class in (2.33) contains models such as the multivariate normal, $t-$, and Kotz-type distributions.

Finally, simulation study reveals that, when p is much larger than n, the estimate of Σ and Σ^{-1} are quite poor. This observation agrees with Kubokawa and Srivastava [2008], where Haff [1979a]-type improved estimates of Σ are proposed. It would be of interest to use an improved estimator of Σ in $\delta_r(X, S)$ in (2.3). As pointed out in the testing context by Srivastava and Fujikoshi [2006] and Srivastava [2007], a shortcoming of S^+ is that the associated estimator is only orthogonally invariant, while the sample mean vector is invariant.

CHAPTER 3

SECOND ORDER ESTIMATION IN THE SINGULAR MULTIVARIATE NORMAL MODEL

3.1 Introduction

Classical statistics is often confined to the setting where the sample size of the data is greater than the number of covariates under consideration. With the recent explosion of available data, much interest has arisen in degenerate situations where the number of covariates is greater than the sample size. In this situation, it is typically assumed that, despite their number, the underlying covariates are linearly independent, or in other words that their covariance matrix has full rank. However, little attention has been shown to the situation where linear dependence would hold between the covariates, that is, where the covariance matrix would be singular.

Recently, Tsukuma and Kubokawa [2014] investigated the problem of estimating the mean vector of a multivariate normal distribution when the unknown covariance matrix is singular. By deriving an unbiased risk estimator for the quadratic loss, they were able to express sufficient conditions for an estimator to dominate the maximum likelihood estimator.

This article is concerned with the same model but three different tasks. Unlike the mean estimation problem of Tsukuma and Kubokawa [2014], all three concern second order moments of the distribution. In each case we aim to provide decision-theoretic results that lead to improved inference. The first task is the estimation of the singular covariance matrix itself, under an invariant squared loss. This problem was first considered in the full rank case by Haff [1980], and in the high-dimensional setting by Konno [2009]. The second task is the estimation of the Moore-Penrose pseudo-inverse of the covariance matrix, also known as the precision matrix, under

the Frobenius loss. This problem was first considered in the full rank case by Haff [1977, 1979b] and in the high-dimensional setting by Kubokawa and Srivastava [2008]. Finally, we consider the task of estimating the discriminant coefficients that arise in Linear Discriminant Analysis (LDA), a popular linear classifier, under the squared loss. This problem was first considered in the full rank case by Haff [1986] and Dey and Srinivasan [1991]. As far as we know, no work has been done on discriminant coefficients in a high-dimensional context where the number of covariates is greater than the sample size.

The presentation of our approach to these problems is divided as follows. The decision-theoretic results are described in Section 3.2. For each of the three problems, we construct an appropriate unbiased estimator of the risk (URE) using Stein's and Haff's lemmas [Stein, 1986, Haff, 1979a], and the approach of [Tsukuma and Kubokawa, 2014]. We then consider the class of estimator given by constant multiples of a naive estimator, and minimize an upper bound on the difference in risk to obtain estimators that dominate the naive estimator. Finally, we consider a larger class given by the sum of this estimator and an appropriate trace, and again minimize an upper bound on the risk to obtain a dominating estimator.

In Section 3.3, we investigate the amount of improvement provided by the proposed estimators through Monte Carlo simulations. Finally, proofs of the statements of Section 3.2 are provided in Section 3.5.

3.2 Estimation

3.2.1 Model

Our setting is almost identical to the one of Tsukuma and Kubokawa [2014]. We observe an n-sample $X_1, ..., X_n$ identically and independently distributed from a

p-dimensional multivariate normal distribution $N_p(\mu, \Sigma)$, where μ and Σ are unknown. However, the p-dimensional covariance matrix Σ is rank-deficient with respect to the dimension and the sample size, in the sense that

$$r = \mathrm{rk}(\Sigma) < \min(n, p). \tag{3.1}$$

The resulting singular multivariate normal distribution does not have a density with respect to the Lebesgue measure on $\mathbb{R}^p$, but lives in the r-dimensional linear subspace spanned by the columns of Σ. More details can be found, for example, in Srivastava and Khatri [1979, Section 2.1].

Define the $n \times p$ data matrix $X = (X_1, ..., X_p)^t$. The sample covariance matrix $S = (X - 1_n \bar{X}^t)^t (X - 1_n \bar{X}^t)/n$ then follows a Wishart distribution $W_p(n-1, \Sigma/n)$ with $n - 1$ degrees of freedom. Since Σ is rank-deficient, it is singular in the terminology of Srivastava and Khatri [1979, Section 3.1]. We warn the reader that the expression "singular Wishart" has also been used in the literature to describe the different situation where the covariance is positive-definite and the dimension exceeds the degrees of freedom, as in Srivastava [2003]. Let $S = O_1 L O_1^t$ denote the reduced spectral decomposition of S, where $L = \mathrm{diag}(l_1, ..., l_r)$ denote the r non-zero eigenvalues and O_1 is $p \times r$ semi-orthogonal.

In this situation, neither S nor Σ are invertible. Since inverses of covariance matrix are of considerable interest in multivariate statistical analysis, some generalized inverse of these quantities is desirable. In this article, we will focus on the Moore-Penrose pseudoinverse, which will be denoted A^+ for a matrix A. Definitions and theoretical properties can be found in Harville [1997, Chapter 20].

The singular multivariate normal model is amenable to decision-theoretic analysis through a key insight of Tsukuma and Kubokawa [2014, Section 2.2]. The authors proved that when (3.1) holds, the subspace spanned by the sample covariance matrix is almost surely constant and matches the subspace spanned the true

covariance matrix, in the sense that

$$SS^+ = \Sigma\Sigma^+. \tag{3.2}$$

This fact will be repeatedly used in Section 3.5, and is essential to our derivations.

Let us now turn our attention to the three problems we wish to solve. In terms of the notation introduced above, these are:

Covariance matrix estimation. The estimation of Σ under the invariant squared loss $L(\hat{\Sigma}, \Sigma) = \mathrm{tr}[(\hat{\Sigma}\Sigma^+ - I_p)^2]$.

Precision matrix estimation. The estimation of Σ^+ under the Frobenius loss $L(\hat{\Sigma}^+, \Sigma^+) = \|\hat{\Sigma}^+ - \Sigma^+\|_F^2$.

Discriminant coefficient estimation. The estimation of $\eta = \Sigma^+\mu$ under the square loss $L(\hat{\eta}, \eta) = \|\hat{\eta} - \eta\|_2^2$.

Traditional estimators for μ and Σ are the empirical mean $\bar{X}$ and the sample covariance matrix S, which suggests the naive estimators S, S^+ and $S^+\bar{X}$ for each respective problem. We will see they are not admissible.

3.2.2 Covariance matrix estimation

The standard estimator for a covariance matrix is the sample covariance matrix S. An alternative is the unbiased estimator $\frac{n}{n-1}S$, which corrects for the loss in degrees of freedom from not knowing μ. We will look for estimators that improve over these benchmarks and study their performance.

We first show that an unbiased estimator of the risk holds for orthogonally invariant estimators, that is, estimators of the form $\hat{\Sigma} = O_1\Psi O_1^t$ with $\Psi = \mathrm{diag}(\psi_1, ..., \psi_r)$ twice-differentiable functions of $L = \mathrm{diag}(l_1, ..., l_r)$.

Theorem 3 (Unbiased risk estimation for singular covariance matrices)**.** *Let* $1 \leq r \leq n-1$ *and define*

$$\psi_k^* = \left[\frac{n-r-2}{n} \frac{\psi_k}{l_k} + \frac{4}{n} \frac{\partial \psi_k}{\partial l_k} + \frac{2}{n} \sum_{b \neq k}^{r} \frac{\psi_k - \psi_b}{l_k - l_b} - 2 \right] \psi_k.$$

Assume the regularity conditions

$$\begin{aligned}
&\mathrm{E}\left[\left| p + \sum_{k=1}^{r} \frac{n-r-2}{n} \frac{\psi_k^*}{l_k} + \frac{2}{n} \sum_{k=1}^{r} \frac{\partial \psi_k^*}{\partial l_k} + \frac{1}{n} \sum_{k \neq b}^{r} \frac{\psi_k^* - \psi_b^*}{l_k - l_b} \right| \right] < \infty, \\
&\mathrm{E}\left[\left| p + \sum_{k=1}^{r} \frac{n-r-2}{n} \frac{\psi_k}{l_k} + \frac{2}{n} \sum_{k=1}^{r} \frac{\partial \psi_k}{\partial l_k} + \frac{1}{n} \sum_{k \neq b}^{r} \frac{\psi_k - \psi_b}{l_k - l_b} \right| \right] < \infty, \\
&\mathrm{E}\left[\sum_{k=1}^{r} \left| \frac{\psi_k^*}{l_k} \right|^2 \right] < \infty \text{ and } \mathrm{E}\left[\sum_{k=1}^{r} \left| \frac{\psi_k}{l_k} \right|^2 \right] < \infty.
\end{aligned} \tag{3.3}$$

We then have

$$\begin{aligned}
&\mathrm{E}\Big[tr\Big(\big[\hat{\Sigma} \Sigma^+ - I_p \big]^2 \Big) \Big] \\
&\quad = \mathrm{E}\left[p + \frac{n-r-2}{n} \sum_{k=1}^{r} \frac{\psi_k^*}{l_k} + \frac{2}{n} \sum_{k=1}^{r} \frac{\partial \psi_k^*}{\partial l_k} + \frac{1}{n} \sum_{k \neq b}^{r} \frac{\psi_k^* - \psi_b^*}{l_k - l_b} \right].
\end{aligned} \tag{3.4}$$

Let us now consider estimators that are proportional to the sample covariance matrix, that is, of the form aS for a constant. The following result provides the optimal proportionality factor.

Proposition 2. *Let* $1 \leq r \leq n-1$*. The optimal estimator of* Σ *of the form* aS *for* $a \in \mathbb{R}$ *a deterministic constant is* $\hat{\Sigma}_{HF1} = \frac{n}{n+r} S$*, with risk*

$$\mathrm{E}\Big[tr\Big(\big[\hat{\Sigma}_{HF1} \Sigma^+ - I_p \big]^2 \Big) \Big] = p - \frac{(n-1)r}{n+r}.$$

In particular $\hat{\Sigma}_{HF1}$ *dominates* S*, which itself dominates* $\frac{n}{n-1} S$*.*

Thus $\frac{n}{n-1} S$ and S are inadmissible. We can further extend this result by considering a larger class of estimators of the form $\frac{n}{n+r} \left[S + t S S^+ \operatorname{tr}^{-1}(S^+) \right]$ for t constant. Estimators of this shape were first considered by Haff [1980]. Although computing the exact risk of these estimators is difficult, it is possible to bound the difference in risk with the one of $\hat{\Sigma}_{\mathrm{HF1}}$ as follows.

Proposition 3. *Let* $1 \leq r \leq n-4$. *Then the risk of estimators of the form* $\hat{\Sigma}_t = \frac{n}{n+r}\left[S + tSS^+\, tr^{-1}(S^+)\right]$ *for* $t \in \mathbb{R}$ *can be bounded by*

$$\mathrm{E}\Big[tr\Big(\big[\hat{\Sigma}_t\Sigma^+ - I_p\big]^2\Big)\Big] \leq \mathrm{E}\Big[tr\Big(\big[\hat{\Sigma}_{HF1}\Sigma^+ - I_p\big]^2\Big)\Big] + \left[\frac{(n-r)(n-r+2)}{(n+r)^2}t^2 - 2\frac{(n-r)(r-1)}{(n+r)^2}t\right]\mathrm{E}\left[\frac{tr(S^{+2})}{tr^2(S^+)}\right]. \tag{3.5}$$

The constant that minimizes this upper bound is $t = \frac{r-1}{n-r+2}$. *When* $r = 1$, *the corresponding estimator* $\hat{\Sigma}_{HF2} = \frac{n}{n+r}\left[S + \frac{r-1}{n-r+2}SS^+tr^{-1}(S^+)\right]$ *equals* $\hat{\Sigma}_{HF1}$, *while for* $r \geq 2$ *it dominates* $\hat{\Sigma}_{HF1}$.

Thus $\hat{\Sigma}_{\mathrm{HF1}}$ is itself inadmissible for $r > 1$. Although this result does not show $\hat{\Sigma}_{\mathrm{HF2}}$ optimal within the class, it might be a good approximation.

3.2.3 Precision matrix estimation

A standard estimator for a singular precision matrix is the Moore-Penrose pseudoinverse of the sample covariance matrix S^+. Note that by Muirhead [1982, Page 97, Equation (12)] we have

$$\mathrm{E}\big[S^+\big] = \frac{n}{n-r-2}\Sigma^+.$$

for $n-r-2 > 0$. Thus in this case an alternative could be the unbiased estimator $\frac{n-r-2}{n}S^+$. We will look for estimators that improve over these benchmarks and study their performance.

We first show that an unbiased estimator of the risk holds for orthogonally invariant estimators, that is, estimators of the form $\hat{\Sigma}^+ = O_1\Psi O_1^t$ with $\Psi = \mathrm{diag}(\psi_1, ..., \psi_r)$ twice-differentiable functions of $L = \mathrm{diag}(l_1, ..., l_r)$.

Theorem 4 (Unbiased risk estimation for singular precision matrices)**.** *Let* $1 \leq r \leq n-1$. *Assume the regularity condition*

$$\mathrm{E}\left[\left|\frac{n-r-2}{n}\sum_{k=1}^{p}\frac{\psi_k}{l_k} + \frac{2}{n}\sum_{k=1}^{r}\frac{\partial\psi_k}{\partial l_k} + \frac{1}{n}\sum_{k\neq b}\frac{\psi_k - \psi_b}{l_k - l_b}\right|\right] < \infty.$$

Then

$$\mathrm{E}\Big[\|\hat{\Sigma}^+ - \Sigma^+\|_F^2\Big] = \mathrm{E}\left[\sum_{k=1}^{r}\psi_k^2 - 2\frac{n-r-2}{n}\sum_{k=1}^{r}\frac{\psi_k}{l_k} - \frac{4}{n}\sum_{k=1}^{r}\frac{\partial\psi_k}{\partial l_k} - \frac{2}{n}\sum_{k\neq b}^{r}\frac{\psi_k-\psi_b}{l_k-l_b}\right] + \mathrm{tr}\big(\Sigma^{-2}\big)\,.$$

Let us now consider estimators that are proportional to the Moore-Penrose inverse of the sample covariance matrix, that is, of the form aS^+ for a constant. The following optimality result holds over this class.

Proposition 4. *Let $1 \leq r \leq n-5$. The risk of estimators of the form aS^+ for $a \leq \frac{n-r-2}{n}$ can be bounded in terms of the risk of $\frac{n-r-2}{n}S^+$ by*

$$\mathrm{E}\big[\|aS^+ - \Sigma^+\|_F^2\big] \leq \mathrm{E}\left[\left\|\frac{n-r-2}{n}S^+ - \Sigma^+\right\|_F^2\right] + \left(a - \frac{n-r-2}{n}\right)\left(a - \frac{n-r-6}{n}\right)\mathrm{E}\big[tr(S^{+2})\big]\,. \quad (3.6)$$

The constant that minimizes this upper bound is $a = \frac{n-r-4}{n}$, and the corresponding estimator $\hat{\Sigma}^+_{EM1} = \frac{n-r-4}{n}S^+$ dominates $\frac{n-r-2}{n}S^+$, which itself dominates S^+.

Thus $\frac{n-r-2}{n}S^+$ and S^+ are inadmissible. Note that our bound on the risk only holds for $a \leq \frac{n-r-2}{n}$: presumably, estimators aS^+ with $a > \frac{n-r-2}{n}$ do not dominate $\frac{n-r-2}{n}S^+$, but we have not been able to prove this hypothesis.

In any case, we can further extend this result by considering a larger class of estimators of the form $\frac{n-r-4}{n}\left[S^+ + t\,SS^+\mathrm{tr}^{-1}(S)\right]$ for t constant. Estimators of this shape were first considered by Efron and Morris [1976]. It is possible to bound the difference in risk with the one of $\hat{\Sigma}^+_{\mathrm{EM1}}$ as follows.

Proposition 5. *Let $1 \leq r \leq n-5$. The risk of estimators of the form $\hat{\Sigma}^+_t = \frac{n-r-4}{n}\left[S^+ + t\,SS^+tr^{-1}(S)\right]$ for $t \in \mathbb{R}$ can be bounded in terms of the risk of $\hat{\Sigma}^+_{EM1} =$*

$\frac{n-r-4}{n}S^+$ *through*

$$\mathrm{E}\Big[\|\hat{\Sigma}_t^+ - \Sigma^+\|_F^2\Big] \leq \mathrm{E}\left[\left\|\hat{\Sigma}_{EM1}^+ - \Sigma^+\right\|_F^2\right] + \frac{(n-r-4)r}{n^2}\Big[(n-r-4)t^2 - 4(r-1)t\Big]\,\mathrm{E}\left[\frac{1}{tr^2(S)}\right]. \quad (3.7)$$

The constant that minimizes this upper bound is $t = 2\frac{r-1}{n-r-4}$*, and the corresponding estimator* $\hat{\Sigma}_{EM2}^+ = \frac{n-r-4}{n}\left[S^+ + 2\frac{r-1}{n-r-4}SS^+tr^{-1}(S)\right]$ *dominates* $\hat{\Sigma}_{EM1}^+$.

Thus $\hat{\Sigma}^+_{\mathrm{EM1}}$ is itself inadmissible. Although these results does not show $\hat{\Sigma}^+_{\mathrm{EM1}}$ and $\hat{\Sigma}^+_{\mathrm{EM2}}$ optimal within their classes, they might be good approximations.

3.2.4 Discriminant coefficients estimation

A standard estimator for a singular discriminant coefficient is $S^+\bar{X}$. Note that since $\bar{X}$ and S are independent, we have

$$\mathrm{E}\big[S^+\bar{X}\big] = \frac{n}{n-r-2}\Sigma^+\mu$$

for $n-r-2 > 0$. Thus in this case an alternative could be the unbiased estimator $\frac{n-r-2}{n}S^+\bar{X}$. We will look for estimators that improve over these benchmarks and study their performance.

We first show that an unbiased estimator of the risk holds for estimators of the form $\hat{\eta} = O_1\Psi O_1^t\bar{X}$ with $\Psi = \mathrm{diag}(\psi_1, ..., \psi_r)$ twice-differentiable functions of $L = \mathrm{diag}(l_1, ..., l_r)$.

Theorem 5 (Unbiased risk estimation for singular discriminant coefficients)**.** *Let* $\Psi^* = diag(\psi_1^*, ..., \psi_r^*)$ *with*

$$\psi_k^* = \frac{n-r-2}{n}\frac{\psi_k}{l_k} + \frac{2}{n}\frac{\partial\psi_k}{\partial l_k} + \frac{1}{n}\sum_{b\neq k}^{r}\frac{\psi_k - \psi_b}{l_k - l_b}.$$

Assume the regularity conditions

$$\mathrm{E}\left[\left|\sum_{k=1}^{r}\psi_k\right|\right] < \infty \text{ and } \mathrm{E}\left[\sum_{k=1}^{r}\left|\psi_k^*\right|\right] < \infty.$$

Then

$$\mathrm{E}\left[\left\|\hat{\eta}-\eta\right\|_2^2\right] = \mathrm{E}\left[\frac{2}{n}tr\hat{\Sigma}^{+} + \bar{X}^t O_1(\Psi^2 - 2\Psi^*)O_1^t\bar{X}\right] - \mathrm{E}\left[(\bar{X}-\mu)^t\Sigma^{+2}(\bar{X}+\mu)\right].$$

Let us now consider estimators that are proportional to the naive estimator, that is, of the form $aS^{+}\bar{X}$ for a constant. The following optimality result holds over this class.

Proposition 6. *Let $1 \le r \le n-5$. The risk of estimators of the form $aS^{+}\bar{X}$ for $a \le \frac{n-r-2}{n}$ can be bounded in terms of the risk of $\frac{n-r-2}{n}S^{+}\bar{X}$ by*

$$\mathrm{E}\left[\left\|aS^{+}\bar{X}-\eta\right\|_2^2\right] \le \mathrm{E}\left[\left\|\frac{n-r-2}{n}S^{+}\bar{X}-\eta\right\|_2^2\right] + \left(a-\frac{n-r-2}{n}\right)\left(a-\frac{n-r-4}{n}\right)E\Big(\bar{X}^t S^{+2}\bar{X}\Big). \tag{3.8}$$

The constant that minimizes this upper bound is $a=\frac{n-r-3}{n}$, and the corresponding estimator $\hat{\eta}_{TK1}=\frac{n-r-3}{n}S^{+}\bar{X}$ dominates $\frac{n-r-2}{n}S^{+}\bar{X}$, which itself dominates $S^{+}\bar{X}$.

Thus $\frac{n-r-2}{n}S^{+}$ and S^{+} are inadmissible. Again, note that our bound on the risk only holds on the subset $a \le \frac{n-r-2}{n}$. Presumably, estimators aS^{+} with $a > \frac{n-r-2}{n}$ do not dominate $\frac{n-r-2}{n}S^{+}\bar{X}$, but we have not been able to prove this result.

We can further extend this result by considering a larger class of estimators of the form $\frac{n-r-3}{n}\left[S^{+} + t\,SS^{+}\mathrm{tr}^{-1}(S)\right]\bar{X}$ for t constant. Estimators of this shape were first considered by Dey and Srinivasan [1991]. It is possible to bound the difference in risk with the one of $\hat{\eta}_{\mathrm{TK1}}$ as follows.

Proposition 7. *Let* $1 \leq r \leq n-5$. *The risk of estimators of the form* $\hat{\eta}_t = \frac{n-r-3}{n}\left[S^+ + t\, SS^+ tr^{-1}(S)\right]\bar{X}$ *for* $t \in \mathbb{R}$ *can be bounded in terms of the risk of* $\eta_{TK1} = \frac{n-r-3}{n} S^+ \bar{X}$ *through*

$$\begin{aligned}\mathrm{E}\left[\|\hat{\eta}_t - \eta\|_2^2\right] \leq\ & \mathrm{E}\left[\left\|\hat{\eta}_{TK1} - \eta\right\|_2^2\right] \\ & + \frac{(n-r-3)}{n^2}\left[2(r+1)t + (n-r-3)t^2\right]\mathrm{E}\left[\frac{1}{tr(S)}\right]. \qquad (3.9)\end{aligned}$$

The constant that minimizes this upper bound is $t = -\frac{r+1}{n-r-3}$, *and the corresponding estimator* $\hat{\eta}_{TK2} = \frac{n-r-3}{n}\left[S^+ - \frac{r+1}{n-r-3} SS^+ tr^{-1}(S)\right]\bar{X}$ *dominates* $\hat{\eta}_{TK1}$.

Thus $\hat{\eta}_{\mathrm{TK1}}$ is itself inadmissible. Although these results does not show $\hat{\eta}_{\mathrm{TK1}}$ and $\hat{\eta}_{\mathrm{TK2}}$ optimal within their classes, they are hopefully good approximations.

3.3 Numerical study

We investigated the risk performance of the proposed estimator for covariance, precision and discriminant coefficient estimation through two Monte Carlo simulations.

3.3.1 Autoregressive simulation

We let (n, p) be $(150, 100)$, $(200, 100)$, $(200, 150)$ and $(250, 150)$. For each r from 1 to $(n-4) \wedge p$, we constructed the true covariance matrix Σ from an autoregressive structure with coefficient 0.9 and set its $p-r$ smallest eigenvalues to zero to create a rank r matrix, as described in Algorithm 1. We then randomly generated $1,000$ replications from a multivariate normal distribution with mean $\mu = (1, \ldots, 1)$ and singularized autoregressive covariance Σ, and computed the resulting sample covariance matrix $S = X^t X/n$.

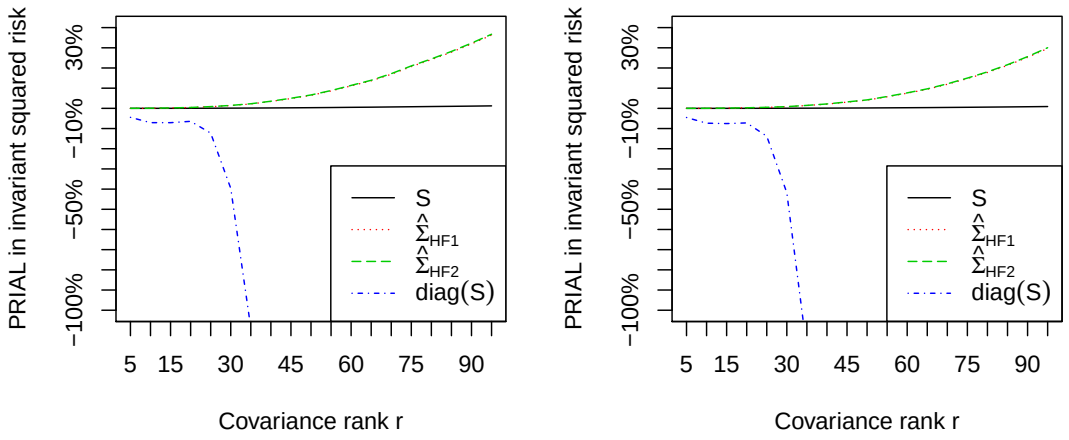
Algorithm 1: Algorithm for generating Σ

Data: p, r
Result: Σ
for $i, j \in \{1, ..., p\}$ **do**
  $\Sigma_{ij} = 0.5^{|i-j|}$
end
for $k \in \{r+1, ..., p\}$ **do**
  $\lambda_k(\Sigma) = 0$
end

For the covariance matrix estimation problem, we computed the Percentage Reduction In Average Loss (PRIAL) with respect to $\frac{n}{n-1}S$ in invariant squared loss $L(\hat{\Sigma}, \Sigma) = \text{tr}[(\hat{\Sigma}\Sigma^+ - I_p)^2]$ for four estimators. The first three are the estimators S, $\hat{\Sigma}_{\text{HF1}} = \frac{n}{n+r}S$ and $\hat{\Sigma}_{\text{HF2}} = \frac{n}{n+r}\left[S + \frac{r-1}{n-r+2}SS^+\text{tr}^{-1}(S^+)\right]$ considered in Subsection 3.2.2. We also included as fourth estimator the diagonal of the sample covariance matrix $\text{diag}(S)$. The simulation results are given in Figure 3.1. We notice that $\hat{\Sigma}_{\text{HF1}}$ and $\hat{\Sigma}_{\text{HF2}}$ behave similarly, and both improve substantially on S, while the diagonal estimator does much worse.

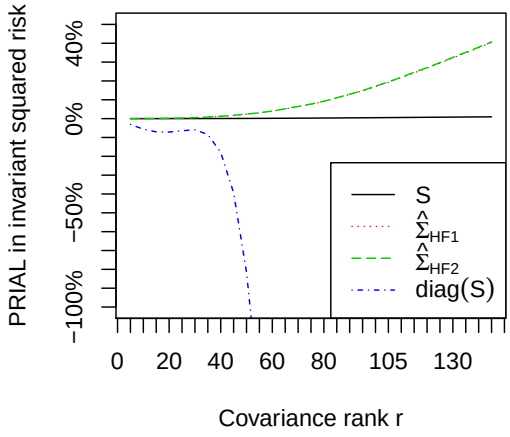
Similarly, for the precision matrix estimation problem, we estimated the PRIAL with respect to S^+ in the Frobenius loss $L(\hat{\Sigma}^+, \Sigma^+) = \|\hat{\Sigma}^+ - \Sigma^+\|_F^2$ for four estimators. The first three are the estimators $\frac{n-r-2}{n}S^+$, $\hat{\Sigma}_{\text{EM1}} = \frac{n-r-4}{n}S^+$ and $\hat{\Sigma}_{\text{EM2}} = \frac{n-r-4}{n}\left[S^+ + 2\frac{r-1}{n-r-4}SS^+\text{tr}^{-1}(S)\right]$ from Subsection 3.2.3. The fourth one is the inverse of the diagonal of the sample covariance matrix, $\text{diag}(S)^{-1}$. The simulation results are given in Figure 3.2. We can see that all first three estimators improve substantially over S^+, but do not differ significantly in risk. In contrast, the diagonal estimator performs well when the true matrix is almost full rank, but becomes worse and worse for smaller covariance ranks.

Finally, for the discriminant coefficient estimation problem, we estimated the PRIAL with respect to $S^+\bar{X}$ in the square loss $L(\hat{\eta}, \eta) = \|\hat{\eta} - \eta\|_2^2$ for four estimators. The first three estimators are $\frac{n-r-2}{n}S^+\bar{X}$, $\hat{\eta}_{\text{TK1}} = \frac{n-r-3}{n}S^+\bar{X}$ and

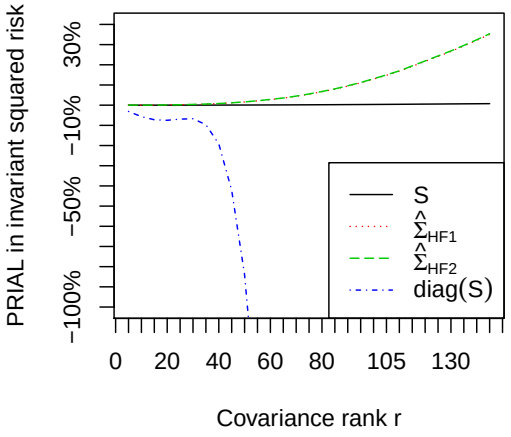


(a) n=150 and p=100

(b) n=200 and p=100



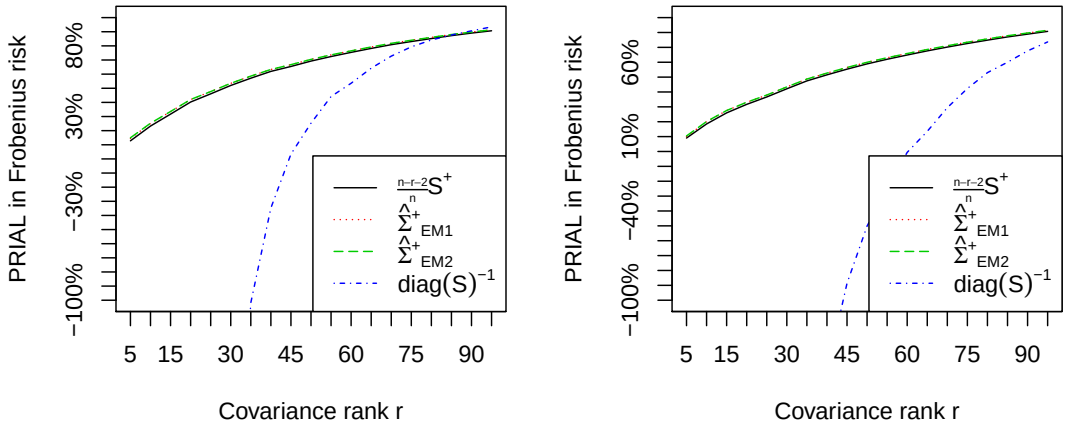
(c) n=200 and p=150



(d) n=250 and p=150

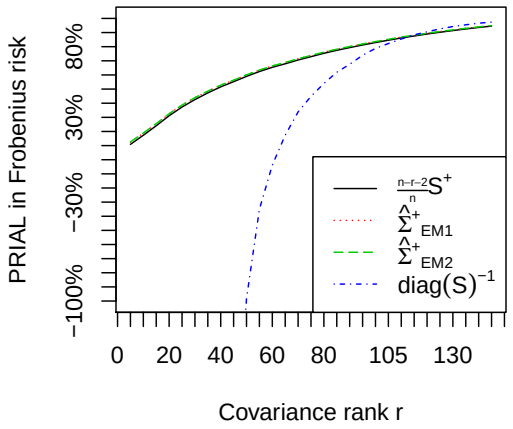
Figure 3.1: PRIAL of S, $\hat{\Sigma}_{\text{HF1}}$, $\hat{\Sigma}_{\text{HF2}}$ and $\text{diag}(S)$ with respect to $\frac{n-1}{n}S$ for estimating Σ in invariant squared loss.

$\hat{\eta}_{\text{TK2}} = \frac{n-r-3}{n}\big[S^{+} - \frac{r+1}{n-r-3}\text{tr}^{-1}(S)\big]\bar{X}$, which were considered in Subsection 3.2.4. The fourth one is the estimator $\text{diag}(S)^{-1}\bar{X}$, which has been considered in linear discriminant analysis when $p > n$. The simulation results are given in Figure 3.3. In this case again, all first three estimators have similar risk and substantially improve on the naive estimator, $S^{+}\bar{X}$, while the diagonal estimator is acceptable only when the true covariance matrix is almost full rank and quite bad otherwise.

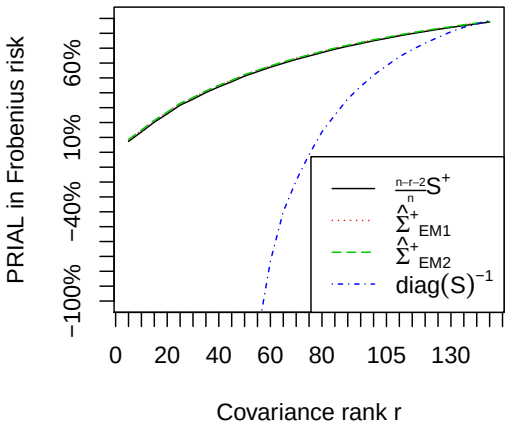


(a) n=150 and p=100

(b) n=200 and p=100



(c) n=200 and p=150



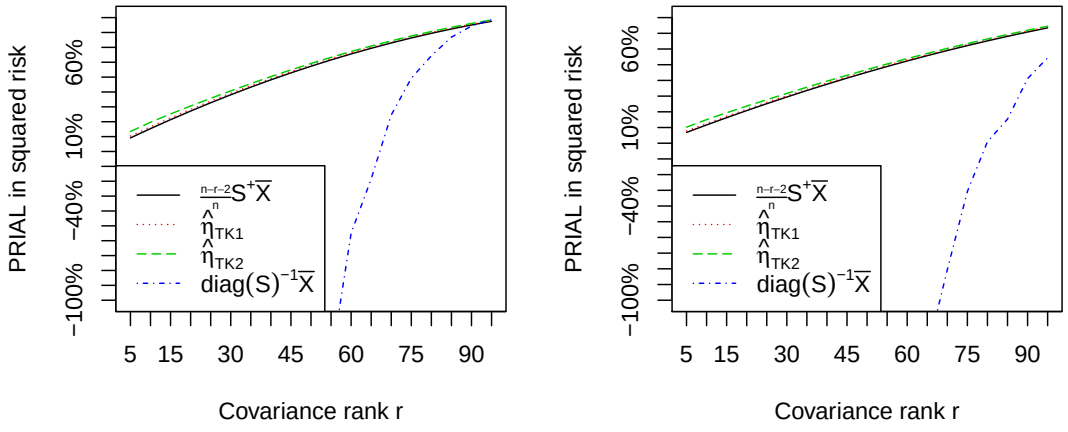
(d) n=250 and p=150

Figure 3.2: PRIAL of $\frac{n-r-2}{n}S^+$, $\hat{\Sigma}^+_{\text{EM1}}$, $\hat{\Sigma}^+_{\text{EM2}}$ and $\text{diag}(S)^{-1}$ with respect to S^+ for estimating Σ^+ in Frobenius loss.

3.3.2 NASDAQ-100 simulation

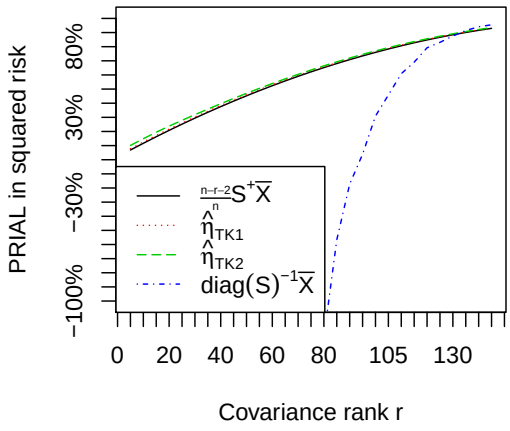
To explore more realistic designs than an autoregressive covariance matrix, we also considered a setting where the true covariance matrix was constructed from real data.

The NASDAQ-100 is a stock market index composed of the hundred largest non-financial companies on the NASDAQ. As of 2015, this is composed of 107 securities, since some companies offer several classes of stock. We computed the net daily returns of these assets up to March 6, 2015. The newest security is Liberty

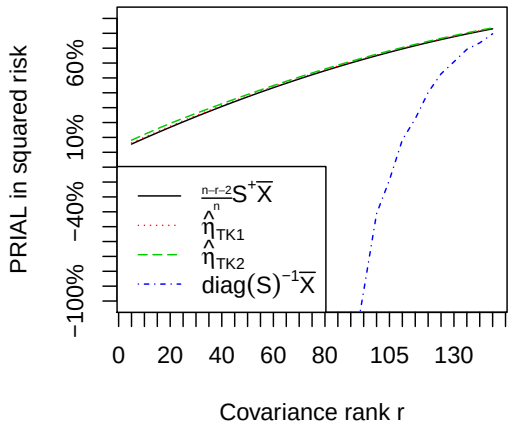


(a) n=150 and p=100

(b) n=200 and p=100



(c) n=200 and p=150



(d) n=250 and p=150

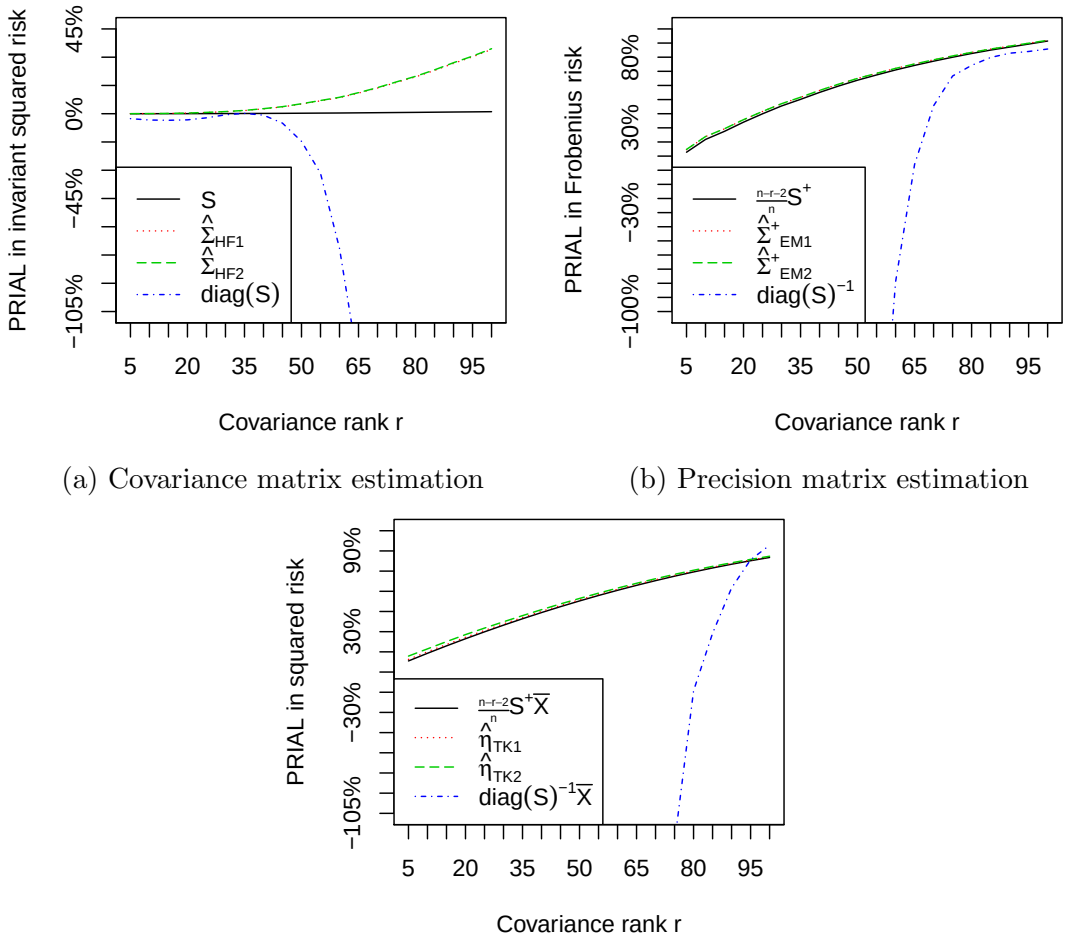
Figure 3.3: PRIAL of $\frac{n-r-2}{n}S^{+}\bar{X}$, $\hat{\eta}^{+}_{\mathrm{TK1}}$, $\hat{\eta}^{+}_{\mathrm{TK2}}$ and $\mathrm{diag}(S)^{-1}\bar{X}$ with respect to $S^{+}\bar{X}$ for estimating $\eta = \Sigma^{+}\mu$ in squared loss.

Media Corp Series C (LMCK), which was issued to series A and B shareholders as dividend on July 7, 2014. To avoid missing data issues, we took this date as the initial time point. This yielded a sample size of 167 trading days. From this data we computed a 107×107 sample covariance matrix of the NASDAQ-100 returns.

We then proceeded with the risk simulation as follows. For every r from 1 to $(n-4) \wedge p$, the true covariance matrix Σ was defined as the NASDAQ-100 sample covariance matrix with its $p-r$ smallest eigenvalues set to zero. We then randomly generated $1,000$ replications from a multivariate normal distribution

with mean $\mu = (1, \ldots, 1)$ and singular covariance Σ, and computed the resulting sample covariance matrix $S = X^t X/n$.

For each of the three estimation problems, we computed the PRIAL as in Subsection 3.3.1. The simulation results are given in Figure 3.4. The results appear similar to the singularized autoregressive setting.



(a) Covariance matrix estimation

(b) Precision matrix estimation

(c) Discriminant coefficient estimation

Figure 3.4: PRIAL for the singularized NASDAQ-100 covariance matrix in the three estimation tasks.

3.4 Discussion

The Tsukuma and Kubokawa technique exposed in Subsection 3.2.1 allows in essence to change the dimension from p to r. Since $r < \min(n, p)$, this in effect turns the problem into a classical setting where the sample size is greater than the dimension, and allows the usual theory to be applied.

An interesting extension is the setting where $n \leq r < p$. In that case, an adaptation of the method would yield a high-dimensional context where the true covariance matrix is full rank, but the sample size n is still smaller than the dimension p. Recent work, for example by Konno [2009], could allow the construction of improved estimators analogous to the ones presented in this article.

Recent attention has been given to the notion of the effective rank of a matrix $r(A) = \mathrm{tr}(S)/\|A\|_2$, first suggested by Vershynin [2010], in the study of spiked covariance matrices [Bunea and Xiao, 2012]. Singular covariance matrices can be regarded as a boundary case of spiked matrices where the noise equals zero. In that regard, it is interesting to notice that the quantity $\mathrm{tr}(S^{+2})/\mathrm{tr}^2(S^+)$ that appear in Inequality (3.5) is related to the effective rank of S^+ through the inequality

$$\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)} \leq r(S^+) \leq r^2 \frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)}.$$

The presence of this quantity is likely connected to the orthogonal invariance of the loss function.

Finally, in applications where a singular covariance matrix is unlikely but a low-dimensional approximation is desired, it might be beneficial to use one of the estimators proposed in this article and cross-validate the rank r on the task to accomplish. For example, a mean-variance portfolio optimization problem could use $\hat{\Sigma}^+_{\mathrm{EM2}}$ as precision matrix estimate, with rank r cross-validated on some validation set. To the best of our knowledge, this methodology has no theoretical grounding but might nevertheless prove useful in some high-dimensional problems.

3.5 Proofs

3.5.1 Preliminaries

Before presenting the proofs of the statements from Section 3.2, we explain the techniques employed by Tsukuma and Kubokawa [2014] to work around the singularity of the covariates in the model. Define the sample mean and covariance matrix to be

$$\begin{aligned} \bar{X} &= X'1_n/n & &\sim \mathrm{N}_p(\mu, \Sigma/n), \\ S &= [X - 1_n\bar{X}']'[X - 1_n\bar{X}']/n & &\sim \mathrm{W}_p(n-1, \Sigma/n). \end{aligned}$$

Since Σ has rank r, we can factorize it as $\Sigma = BB^t$ for some full rank $p \times r$ matrix B. Let $H = B(B^tB)^{-1/2}$ and $\Omega = B^tB$ - then H is $p \times r$ semi-orthogonal $H^tH = I_r$ and $HH^t = \Sigma\Sigma^+$, Ω is $r \times r$ invertible, $\Sigma = H\Omega H^t$ and $\Sigma^+ = H\Omega^{-1}H^t$. Since Σ is rank deficient, there must be a $Z \sim \mathrm{N}_{n,r}(0, I_r)$ such that $X = 1_n\mu^t + ZB^t$, and therefore we can write $X = 1_n\mu^t + Z(B^tB)^{1/2}(B^tB)^{-1/2}B^t = 1_n\mu^t + YH^t$ for $Y = Z\Omega^{-1/2} \sim N_{n,r}(0, \Omega)$. Define then

$$\begin{aligned} \bar{Y} &= Y^t1_n/n & &\sim \mathrm{N}_r(0, \Omega/n), \\ T &= [Y - 1_n\bar{Y}^t]'[Y - 1_n\bar{Y}^t]/n & &\sim \mathrm{W}_r(n-1, \Omega/n). \end{aligned}$$

Notice how T is full rank, since $r \leq n-1$, in contrast with S. Using $X = 1_n\mu^t + YH^t$, we can see that these constructions are related to $\bar{X}$ and S through

$$\bar{X} = \mu + H\bar{Y}, \qquad S = HTH^t.$$

Recall that $SS^+ = \Sigma\Sigma^+$ almost surely, from Equation (3.2). Since S has rank $r < p$, there must be a $p \times r$ semi-orthogonal matrix O_1 such that $O_1^tO_1 = I_r$, $O_1O_1^t = \Sigma\Sigma^+$ almost surely and $S = O_1LO_1^t$ for $L = \mathrm{diag}(\lambda_1(S), ..., \lambda_r(S))$. The $r \times r$ matrix $U = H^tO_1$ is easily seen to be orthogonal, and so by $T = H^tSH =$

$H^tO_1LO_1H^t = ULU^t$, we see that T and S must share the same r non-zero eigenvalues, i.e. $\lambda_i(S) = \lambda_i(T)$.

These constructions and facts form the basis of our risk estimation procedures and the notation will be repeatedly used in the following subsections.

3.5.2 Proofs of Subsection 3.2.2

Proof of Theorem 3. Since T and S share the same non-zero eigenvalues, we can regard Ψ as a function of $T \sim \mathrm{W}_r(n-1, \Omega/n)$ only. Since $r \leq n-1$ and Ω is full rank, we can apply Lemma 1 and 2 of Chételat and Wells [2014] to $H^t\hat{\Sigma}H = U\Psi U^t$. On that result, one can also consult Sheena [1995, Theorem 4.1], and in the singular case Kubokawa and Srivastava [2008, Proposition 2.1] and Konno [2009, Theorem 2.4]. In any case, we get

$$\begin{aligned}
\mathrm{E}\Big[\mathrm{tr}\Big(\big[\hat{\Sigma}\Sigma^+ - I_p\big]^2\Big)\Big] &= \mathrm{E}\Big[p - 2\,\mathrm{tr}\Big(\Sigma^+\hat{\Sigma}\Big) + \mathrm{tr}\Big(\Sigma^+\hat{\Sigma}\Sigma^+\hat{\Sigma}\Big)\Big] \\
&= \mathrm{E}\big[(p-r) + r - 2\,\mathrm{tr}\big(\Omega^{-1}U\Psi U'\big) + \mathrm{tr}\big(\Omega^{-1}U\Psi U'\Omega^{-1}U\Psi U'\big)\big] \\
&= \mathrm{E}\Big[p - r + \mathrm{tr}\Big(\big[U\Psi U'\Omega^{-1} - I_r\big]^2\Big)\Big] \\
&= \mathrm{E}\left[(p-r) + r + \frac{n-r-2}{n}\sum_{k=1}^{r}\frac{\psi_k^*}{l_k} + \frac{2}{n}\sum_{k=1}^{r}\frac{\partial\psi_k^*}{\partial l_k} + \frac{1}{n}\sum_{k\neq b}^{r}\frac{\psi_k^* - \psi_b^*}{l_k - l_b}\right],
\end{aligned}$$

under the regularity conditions

$$\begin{aligned}
&\mathrm{E}\left[\left|\sum_{k=1}^{r}\frac{n-r-2}{n}\frac{\psi_k}{l_k} + \frac{2}{n}\sum_{k=1}^{r}\frac{\partial\psi_k}{\partial l_k} + \frac{1}{n}\sum_{k\neq b}^{r}\frac{\psi_k - \psi_b}{l_k - l_b}\right|\right] < \infty, \\
&\mathrm{E}\left[\left|\sum_{k=1}^{r}\frac{n-r-2}{n}\frac{\psi_k^* + 2\psi_k}{l_k} + \frac{2}{n}\sum_{k=1}^{r}\frac{\partial\psi_k^* + 2\psi_k}{\partial l_k}\right.\right. \\
&\qquad \left.\left. + \frac{1}{n}\sum_{k\neq b}^{r}\frac{\psi_k^* + 2\psi_k - \psi_b^* - 2\psi_b}{l_k - l_b}\right|\right] < \infty \text{ and } \mathrm{E}\left[\sum_{k=1}^{r}\left|\frac{\psi_k^* + 2\psi_k}{l_k}\right|^2\right] < \infty.
\end{aligned}$$

But these are satisfied by Inequalities (3.3). This concludes the proof. □

Proof of Proposition 2. Let us apply the results of Theorem 3. We have $\psi_k = al_k$, so

$$\begin{aligned}\psi_k^* &= \left[\frac{n-r-2}{n}a + \frac{4}{n}a + \frac{2}{n}a(r-1) - 2\right] al_k \\ &= \left[\frac{n+r}{n}a - 2\right] al_k.\end{aligned}$$

Then the unbiased risk estimator (3.4) equals

$$\begin{aligned}U &= p + \frac{n-r-2}{n}\sum_{k=1}^{r}\frac{\psi_k^*}{l_k} + \frac{2}{n}\sum_{k=1}^{r}\frac{\partial\psi_k^*}{\partial l_k} + \frac{1}{n}\sum_{k\neq b}^{r}\frac{\psi_k^* - \psi_b^*}{l_k - l_b} \\ &= p + \frac{n-r-2}{n}\left[\frac{n+r}{n}a - 2\right] ar + \frac{2}{n}\left[\frac{n+r}{n}a - 2\right] ar \\ &\qquad + \frac{1}{n}\left[\frac{n+r}{n}a - 2\right] ar(r-1) \\ &= p - 2\frac{(n-1)r}{n}a + \frac{(n-1)(n+r)r}{n^2}a^2.\end{aligned}$$

Clearly, $\mathrm{E}\Big[\Big|U\Big|\Big] = \Big|p - 2\frac{(n-1)r}{n}a + \frac{(n-1)(n+r)r}{n^2}a^2\Big| < \infty$. Similarly,

$$\begin{aligned}&\mathrm{E}\left[\left|p + \sum_{k=1}^{r}\frac{n-r-2}{n}\frac{\psi_k}{l_k} + \frac{2}{n}\sum_{k=1}^{r}\frac{\partial\psi_k}{\partial l_k} + \frac{1}{n}\sum_{k\neq b}^{r}\frac{\psi_k - \psi_b}{l_k - l_b}\right|\right] \\ &\quad = \mathrm{E}\left[\left|p + \frac{(n-r-2)r}{n}a + \frac{2r}{n}a + \frac{r(r-1)}{n}a\right|\right] \\ &\quad = \left|p + \frac{(n-1)r}{n}a\right| < \infty, \\ &\mathrm{E}\left[\sum_{k=1}^{r}\left|\frac{\psi_k^*}{l_k}\right|^2\right] = r\left[\frac{n+r}{n}a - 2\right]^2 a^2 < \infty, \quad \mathrm{E}\left[\sum_{k=1}^{r}\left|\frac{\psi_k}{l_k}\right|^2\right] = ra^2 < \infty.\end{aligned}$$

Thus the regularity conditions of Theorem 3 are satisfied and

$$\mathrm{E}\Big[\mathrm{tr}\Big(\big[\hat{\Sigma}\Sigma^+ - I_p\big]^2\Big)\Big] = \mathrm{E}[U] = p - 2\frac{(n-1)r}{n}a + \frac{(n-1)(n+r)r}{n^2}a^2.$$

But this is minimized when $a = \frac{n}{n+r}$. In particular, notice that since $n \geq r+1 = 2$,

$$\begin{aligned}\mathrm{E}\Big[\mathrm{tr}\Big(\big[\hat{\Sigma}_{\mathrm{HF1}}\Sigma^+ - I_p\big]^2\Big)\Big] &= p - \frac{(n-1)r}{n+r} \\ &< p - \frac{(n-r)(n-1)r}{n^2} = \mathrm{E}\Big[\mathrm{tr}\Big(\big[S\Sigma^+ - I_p\big]^2\Big)\Big] \\ &< p - \frac{(n-r-2)r}{n-1} = \mathrm{E}\left[\mathrm{tr}\Big(\big[\frac{n}{n-1}S\Sigma^+ - I_p\big]^2\Big)\right],\end{aligned}$$

so $\hat{\Sigma}_{\text{HF1}}$ dominates S, which dominates $\frac{n}{n-1}S$, as desired. $\square$

Proof of Proposition 3. Again, let us apply the results of Theorem 3. Here $\psi_k = \frac{n}{n+r}[l_k + t/\text{tr}(S^+)]$, so using that $\frac{\partial}{\partial l_k}\frac{1}{\text{tr}(S^+)} = \frac{1}{l_k^2\text{tr}^2(S^+)}$ we find

$$\begin{aligned}
\psi_k^* &= \left[\frac{n-r-2}{n}\frac{\psi_k}{l_k} + \frac{4}{n}\frac{\partial\psi_k}{\partial l_k} + \frac{2}{n}\sum_{b\neq k}^{r}\frac{\psi_k-\psi_b}{l_k-l_b} - 2\right]\psi_k \\
&= \frac{n}{n+r}\left[\frac{n-r-2}{n+r}\left(1+\frac{t}{l_k\text{tr}\,(S^+)}\right) + \frac{4}{n+r}\left(1+\frac{t}{l_k^2\text{tr}^2(S^+)}\right)\right. \\
&\qquad\qquad \left. + \frac{2(r-1)}{n+r} - 2\right]\cdot\left[l_k + \frac{t}{\text{tr}(S^+)}\right] \\
&= \frac{n}{n+r}\left[1 + \frac{n-r-2}{n+r}\frac{t}{l_k\text{tr}\,(S^+)} + \frac{4}{n+r}\frac{t}{l_k^2\text{tr}^2(S^+)} - 2\right]\left[l_k + \frac{t}{\text{tr}(S^+)}\right] \\
&= -\frac{n}{n+r}l_k + \left[-2\frac{r+1}{n+r}\frac{1}{\text{tr}\,(S^+)} + \frac{4}{n+r}\frac{1}{l_k\text{tr}^2(S^+)}\right]\frac{nt}{n+r} \\
&\quad + \left[\frac{n-r-2}{n+r}\frac{1}{l_k\text{tr}^2(S^+)} + \frac{4}{n+r}\frac{1}{l_k^2\text{tr}^3(S^+)}\right]\frac{nt^2}{n+r}.
\end{aligned}$$

Let us now compute the terms in the URE. We find for the first term:

$$\begin{aligned}
\frac{n-r-2}{n}\sum_{k=1}^{r}\frac{\psi_k^*}{l_k} &= -\frac{n-r-2}{n}\sum_{k=1}^{r}\frac{n}{n+r} \\
&\quad + \frac{n-r-2}{n}\sum_{k=1}^{r}\left[-2\frac{r+1}{n+r}\frac{1}{l_k\text{tr}\,(S^+)} + \frac{4}{n+r}\frac{1}{l_k^2\text{tr}^2(S^+)}\right]\frac{nt}{n+r} \\
&\quad + \frac{n-r-2}{n}\sum_{k=1}^{r}\left[\frac{n-r-2}{n+r}\frac{1}{l_k^2\text{tr}^2(S^+)} + \frac{4}{n+r}\frac{1}{l_k^3\text{tr}^3(S^+)}\right]\frac{nt^2}{n+r} \\
&= -\frac{n-r-2}{n+r}r + \frac{n-r-2}{n+r}\left[-2\frac{r+1}{n+r} + \frac{4}{n+r}\frac{\text{tr}(S^{+2})}{\text{tr}^2(S^+)}\right]t \\
&\quad + \frac{n-r-2}{n+r}\left[\frac{n-r-2}{n+r}\frac{\text{tr}(S^{+2})}{\text{tr}^2(S^+)} + \frac{4}{n+r}\frac{\text{tr}(S^{+3})}{\text{tr}^3(S^+)}\right]t^2.
\end{aligned}$$

Next, using the fact that $\frac{\partial}{\partial l_k}\frac{1}{l_k\text{tr}^2(S^+)} = -\frac{1}{l_k^2\text{tr}^2(S^+)} + \frac{2}{l_k^3\text{tr}^3(S^+)}$ and that $\frac{\partial}{\partial l_k}\frac{1}{l_k^2\text{tr}^3(S^+)} = -\frac{2}{l_k^3\text{tr}^3(S^+)} + \frac{3}{l_k^4\text{tr}^4(S^+)}$, we find

$$\begin{aligned}
\frac{2}{n}\sum_{k=1}^{r}\frac{\partial\psi_k^*}{\partial l_k} &= -\frac{2}{n}\sum_{k=1}^{r}\frac{\partial}{\partial l_k}\frac{n}{n+r}l_k \\
&\quad + \frac{2}{n}\sum_{k=1}^{r}\frac{\partial}{\partial l_k}\left[-2\frac{r+1}{n+r}\frac{1}{\text{tr}\,(S^+)} + \frac{4}{n+r}\frac{1}{l_k\text{tr}^2(S^+)}\right]\frac{nt}{n+r}
\end{aligned}$$

$$
\begin{aligned}
&+\frac{2}{n}\sum_{k=1}^{r}\frac{\partial}{\partial l_k}\left[\frac{n-r-2}{n+r}\frac{1}{l_k\mathrm{tr}^2(S^+)}+\frac{4}{n+r}\frac{1}{l_k^2\mathrm{tr}^3(S^+)}\right]\frac{nt^2}{n+r}\\
&=-\frac{2}{n}\sum_{k=1}^{r}\frac{n}{n+r}+\frac{2}{n}\sum_{k=1}^{r}\Bigg[-2\frac{r+1}{n+r}\frac{1}{l_k^2\mathrm{tr}^2\left(S^+\right)}-\frac{4}{n+r}\frac{1}{l_k^2\mathrm{tr}^2(S^+)}\\
&\quad+\frac{4}{n+r}\frac{2}{l_k^3\mathrm{tr}^3(S^+)}\Bigg]\frac{nt}{n+r}+\frac{2}{n}\sum_{k=1}^{r}\Bigg[-\frac{n-r-2}{n+r}\frac{1}{l_k^2\mathrm{tr}^2(S^+)}\\
&\quad+\frac{n-r-2}{n+r}\frac{2}{l_k^3\mathrm{tr}^3(S^+)}-\frac{4}{n+r}\frac{2}{l_k^3\mathrm{tr}^3(S^+)}+\frac{4}{n+r}\frac{3}{l_k^4\mathrm{tr}^4(S^+)}\Bigg]\frac{nt^2}{n+r}\\
&=-\frac{2r}{n+r}+\frac{2}{n+r}\left[-2\frac{r+3}{n+r}\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2\left(S^+\right)}+\frac{8}{n+r}\frac{\mathrm{tr}(S^{+3})}{\mathrm{tr}^3\left(S^+\right)}\right]t\\
&\quad+\frac{2}{n+r}\left[-\frac{n-r-2}{n+r}\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2\left(S^+\right)}+2\frac{n-r-6}{n+r}\frac{\mathrm{tr}(S^{+3})}{\mathrm{tr}^3\left(S^+\right)}+\frac{12}{n+r}\frac{\mathrm{tr}(S^{+4})}{\mathrm{tr}^4\left(S^+\right)}\right]t^2.
\end{aligned}
$$

Finally, using that $\sum_{k\neq b}^{r}\frac{l_k^{-1}-l_b^{-1}}{l_k-l_b}\leq 0$ and $\sum_{k\neq b}^{r}\frac{l_k^{-2}-l_b^{-2}}{l_k-l_b}\leq 0$ we can bound

$$
\begin{aligned}
\frac{1}{n}\sum_{k\neq b}^{r}\frac{\psi_k^*-\psi_b^*}{l_k-l_b}&=-\frac{1}{n}\sum_{k\neq b}^{r}\frac{n}{n+r}+\frac{1}{n}\left[\frac{4}{n+r}\frac{1}{\mathrm{tr}^2(S^+)}\sum_{k\neq b}^{r}\frac{l_k^{-1}-l_b^{-1}}{l_k-l_b}\right]\frac{nt}{n+r}\\
&\quad+\frac{1}{n}\left[\frac{n-r-2}{n+r}\frac{1}{\mathrm{tr}^2(S^+)}\sum_{k\neq b}^{r}\frac{l_k^{-1}-l_b^{-1}}{l_k-l_b}+\frac{4}{n+r}\frac{1}{\mathrm{tr}^3(S^+)}\sum_{k\neq b}^{r}\frac{l_k^{-2}-l_b^{-2}}{l_k-l_b}\right]\frac{nt^2}{n+r}\\
&\leq\ -\frac{r(r-1)}{n+r}.
\end{aligned}
$$

Hence the URE (3.4) equals

$$
\begin{aligned}
U=&\ p+\frac{n-r-2}{n}\sum_{k=1}^{r}\frac{\psi_k^*}{l_k}+\frac{2}{n}\sum_{k=1}^{r}\frac{\partial\psi_k^*}{\partial l_k}+\frac{1}{n}\sum_{k\neq b}^{r}\frac{\psi_k^*-\psi_b^*}{l_k-l_b}\\
\leq&\ p-\frac{n-r-2}{n+r}r-\frac{2}{n+r}r-\frac{r-1}{n+r}r\\
&+\frac{n-r-2}{n+r}\left[-2\frac{r+1}{n+r}+\frac{4}{n+r}\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)}\right]t\\
&+\frac{2}{n+r}\left[-2\frac{r+3}{n+r}\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2\left(S^+\right)}+\frac{8}{n+r}\frac{\mathrm{tr}(S^{+3})}{\mathrm{tr}^3\left(S^+\right)}\right]t\\
&+\frac{n-r-2}{n+r}\left[\frac{n-r-2}{n+r}\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)}+\frac{4}{n+r}\frac{\mathrm{tr}(S^{+3})}{\mathrm{tr}^3(S^+)}\right]t^2\\
&+\frac{2}{n+r}\left[-\frac{n-r-2}{n+r}\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2\left(S^+\right)}+2\frac{n-r-6}{n+r}\frac{\mathrm{tr}(S^{+3})}{\mathrm{tr}^3\left(S^+\right)}+\frac{12}{n+r}\frac{\mathrm{tr}(S^{+4})}{\mathrm{tr}^4\left(S^+\right)}\right]t^2\\
=&\ p-\frac{(n-1)r}{n+r}+\left[-2\frac{(n-r-2)(r+1)}{(n+r)^2}+4\frac{n-2r-5}{(n+r)^2}\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)}\right.
\end{aligned}
$$

$$+\frac{16}{(n+r)^2}\frac{\mathrm{tr}(S^{+3})}{\mathrm{tr}^3(S^+)}\Big]t+\Big[\frac{(n-r-2)(n-r-4)}{(n+r)^2}\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)}$$
$$+8\frac{n-r-4}{(n+r)^2}\frac{\mathrm{tr}(S^{+3})}{\mathrm{tr}^3(S^+)}+\frac{24}{(n+r)^2}\frac{\mathrm{tr}(S^{+4})}{\mathrm{tr}^4(S^+)}\Big]t^2.$$

Now note that $\mathrm{tr}(S^{+3}) \le \mathrm{tr}^{\frac{1}{2}}(S^{+4})\mathrm{tr}^{\frac{1}{2}}(S^{+2}) \le \mathrm{tr}(S^{+2})\mathrm{tr}(S^+)$ and $\mathrm{tr}(S^{+4}) \le \mathrm{tr}^{\frac{1}{2}}(S^{+6})\mathrm{tr}^{\frac{1}{2}}(S^{+2}) \le \mathrm{tr}(S^{+3})\mathrm{tr}(S^+) \le \mathrm{tr}(S^{+2})\mathrm{tr}^2(S^+)$. Then since $r \le n-4$ and $-1 \le -\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)}$ we can write

$$\begin{aligned}
U \le\ & p-\frac{(n-1)r}{n+r}+\Big[-2\frac{(n-r-2)(r+1)}{(n+r)^2}\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)}+4\frac{n-2r-5}{(n+r)^2}\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)}\\
&+\frac{16}{(n+r)^2}\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)}\Big]t+\Big[\frac{(n-r-2)(n-r-4)}{(n+r)^2}\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)}\\
&+8\frac{n-r-4}{(n+r)^2}\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)}+\frac{24}{(n+r)^2}\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)}\Big]t^2\\
\le\ & p-\frac{(n-1)r}{n+r}+\Big[\frac{(n-r)(n-r+2)}{(n+r)^2}t^2-2\frac{(n-r)(r-1)}{(n+r)^2}t\Big]\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)}. \quad (3.10)
\end{aligned}$$

Now, using that $\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)}, \frac{\mathrm{tr}(S^{+3})}{\mathrm{tr}^3(S^+)}, \frac{\mathrm{tr}(S^{+4})}{\mathrm{tr}^4(S^+)} \le 1$ we find

$$\begin{aligned}
&\mathrm{E}\Big[\Big|p+\sum_{k=1}^{r}\frac{n-r-2}{n}\frac{\psi_k}{l_k}+\frac{2}{n}\sum_{k=1}^{r}\frac{\partial\psi_k}{\partial l_k}+\frac{1}{n}\sum_{k\neq b}^{r}\frac{\psi_k-\psi_b}{l_k-l_b}\Big|\Big]\\
&\quad=\mathrm{E}\Big[\Big|p+\frac{n-r-2}{n+r}\sum_{k=1}^{r}\Big[1+\frac{t}{l_k\mathrm{tr}(S^+)}\Big]\\
&\qquad\qquad+\frac{2}{n+r}\sum_{k=1}^{r}\Big[1+\frac{t}{l_k^2\mathrm{tr}^2(S^+)}\Big]+\frac{1}{n+r}\sum_{k\neq b}^{r}1\Big|\Big]\\
&\quad=\Big|p+\frac{n-r-2}{n+r}(r+t)+\frac{2}{n+r}(r+t)+\frac{r(r-1)}{n+r}\Big|<\infty,\\
&\mathrm{E}\Big[\Big|\sum_{k=1}^{r}\frac{\psi_k}{l_k}\Big|\Big]=\mathrm{E}\Big[\Big|\frac{n}{n+r}\sum_{k=1}^{r}\Big[1+\frac{t}{l_k\mathrm{tr}(S^+)}\Big]\Big|\Big]=\frac{n}{n+r}|r+t|<\infty,\\
&\mathrm{E}\Big[\Big|\sum_{k=1}^{r}\frac{\psi_k^*}{l_k}\Big|\Big]=\frac{n}{n+r}\,\mathrm{E}\Big[\Big|-r+\Big[-2\frac{r+1}{n+r}+\frac{4}{n+r}\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)}\Big]t\\
&\qquad\qquad+\Big[\frac{n}{n+r}\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)}+\frac{4}{n+r}\frac{\mathrm{tr}(S^{+3})}{\mathrm{tr}^3(S^+)}\Big]t^2\Big|\Big]\\
&\qquad\le\ \frac{n}{n+r}\Big[r+\Big(2\frac{r+1}{n+r}+\frac{4}{n+r}\Big)|t|+\Big(\frac{n}{n+r}+\frac{4}{n+r}\Big)t^2\Big]<\infty
\end{aligned}$$

and by (3.10)

$$\mathrm{E}\Big[\Big|U\Big|\Big] \leq p + \frac{(n-1)r}{n+r} + \left[\frac{(n-r)(n-r+2)}{(n+r)^2}t^2 + 2\frac{(n-r)(r-1)}{(n+r)^2}t\right] < \infty.$$

Thus all the regularity conditions of Theorem 3 are satisfied, and we find

$$\begin{aligned}\mathrm{E}\Big[\mathrm{tr}\Big(\big[\hat{\Sigma}_t\Sigma^+ - I_p\big]^2\Big)\Big] &= \mathrm{E}[U] \\ &\leq\ p - \frac{(n-1)r}{n+r} + \left[\frac{(n-r)(n-r+2)}{(n+r)^2}t^2 - 2\frac{(n-r)(r-1)}{(n+r)^2}t\right]\mathrm{E}\left[\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)}\right],\end{aligned}$$

which proves inequality (3.5). To minimize this upper bound, notice that since $\mathrm{E}\left[\frac{\mathrm{tr}(S^{+2})}{\mathrm{tr}^2(S^+)}\right] \geq 0$, it is enough to minimize the quadratic coefficient $\frac{(n-r)(n-r+2)}{(n+r)^2}t^2 - 2\frac{(n-r)(r-1)}{(n+r)^2}t$. This is achieved precisely when $t = \frac{r-1}{n-r+2}$. When $r > 1$, this makes this quadratic coefficient strictly negative, which in view of Proposition 2 guarantees

$$\mathrm{E}\Big[\mathrm{tr}\Big(\big[\hat{\Sigma}_{\mathrm{HF2}}\Sigma^+ - I_p\big]^2\Big)\Big] < p - \frac{(n-1)r}{n+r} = \mathrm{E}\Big[\mathrm{tr}\Big(\big[\hat{\Sigma}_{\mathrm{HF1}}\Sigma^+ - I_p\big]^2\Big)\Big].$$

Thus in this case $\hat{\Sigma}_{\mathrm{HF2}}$ dominates $\hat{\Sigma}_{\mathrm{HF1}}$, as desired. □

3.5.3 Proofs of Subsection 3.2.3

Proof of Theorem 4. Since T and S share the same non-zero eigenvalues, we can regard Ψ as a function of $T \sim \mathrm{W}_r(n-1, \Omega/n)$ only. Since $r \leq n-1$ and Ω is full rank we can apply Lemma 2.1 from Dey [1987]. However, the proposition is given without proof and, more importantly, without the implied regularity conditions that inevitably come from using Stein's and Haff's lemmas. For completeness, we therefore derive again this result in our context. First, we can write

$$\begin{aligned}\mathrm{E}\big[\|O_1\Psi O_1^t - H\Omega^{-1}H^t\|_F^2\big] &= \mathrm{E}\big[\|U\Psi U^t - \Omega^{-1}\|_F^2\big] \\ &= \mathrm{E}\big[\mathrm{tr}\big(U\Psi^2U^t\big) - 2\,\mathrm{tr}\big(\Omega^{-1}U\Psi U^t\big)\big] + \mathrm{tr}\big(\Omega^{-2}\big)\end{aligned}$$

By Lemma 3 of Chételat and Wells [2014], this equals

$$= \mathrm{E}\left[\sum_{k=1}^{r}\psi_k^2 - 2\left(\frac{n-r-2}{n}\sum_{k=1}^{p}+\frac{2}{n}\sum_{k=1}^{r}\frac{\partial\psi_k}{\partial l_k}+\frac{1}{n}\sum_{k\neq b}\frac{\psi_k-\psi_b}{l_k-l_b}\right)\right] + \mathrm{tr}\big(\Omega^{-2}\big)$$

$$= \mathrm{E}\left[\sum_{k=1}^{r}\psi_k^2 - 2\frac{n-r-2}{n}\sum_{k=1}^{r}\frac{\psi_k}{l_k} - \frac{4}{n}\sum_{k=1}^{r}\frac{\partial\psi_k}{\partial l_k} - \frac{2}{n}\sum_{k\neq b}^{r}\frac{\psi_k-\psi_b}{l_k-l_b}\right] + \mathrm{tr}\big(\Omega^{-2}\big)$$

under the regularity condition

$$\mathrm{E}\left[\left|\frac{n-r-2}{n}\sum_{k=1}^{p}\frac{\psi_k}{l_k}+\frac{2}{n}\sum_{k=1}^{r}\frac{\partial\psi_k}{\partial l_k}+\frac{1}{n}\sum_{k\neq b}\frac{\psi_k-\psi_b}{l_k-l_b}\right|\right] < \infty.$$

The result follows from the fact that $\mathrm{tr}(\Omega^{-2}) = \mathrm{tr}(H^tH\Omega^{-1}H^tH\Omega^{-1}) = \mathrm{tr}(\Sigma^{-2})$.

□

Proof of Proposition 4. We have $\psi_k = a/l_k$, so

$$\sum_{k=1}^{r}\psi_k^2 = a^2\sum_{k=1}^{r}\frac{1}{l_k^2} = a^2\mathrm{tr}(S^{+2})$$

$$-2\frac{n-r-2}{n}\sum_{k=1}^{r}\frac{\psi_k}{l_k} = -2\frac{n-r-2}{n}a\sum_{k=1}^{r}\frac{1}{l_k^2} = -2\frac{n-r-2}{n}a\mathrm{tr}(S^{+2})$$

$$-\frac{4}{n}\sum_{k=1}^{r}\frac{\partial\psi_k}{\partial l_k} = -\frac{4}{n}a\sum_{k=1}^{r}-\frac{1}{l_k^2} = \frac{4}{n}a\mathrm{tr}(S^{+2})$$

$$-\frac{2}{n}\sum_{k\neq b}^{r}\frac{\psi_k-\psi_b}{l_k-l_b} = -\frac{2}{n}a\sum_{k\neq b}^{r}\frac{l_k^{-1}-l_b^{-1}}{l_k-l_b} = \frac{2}{n}a\mathrm{tr}^2(S^{+}) - \frac{2}{n}a\mathrm{tr}(S^{+2}).$$

Summing everything, we get the URE

$$U = \frac{2}{n}a\mathrm{tr}^2(S^{+}) + \left(a^2 - 2\frac{n-r-3}{n}a\right)\mathrm{tr}(S^{+2}).$$

Now notice that

$$\mathrm{E}\left[\left|\frac{n-r-2}{n}\sum_{k=1}^{p}\frac{\psi_k}{l_k}+\frac{2}{n}\sum_{k=1}^{r}\frac{\partial\psi_k}{\partial l_k}+\frac{1}{n}\sum_{k\neq b}\frac{\psi_k-\psi_b}{l_k-l_b}\right|\right]$$

$$= |a|\,\mathrm{E}\left[\left|\frac{n-r-3}{n}\mathrm{tr}(S^{+2}) - \frac{1}{n}\mathrm{tr}^2(S^{+})\right|\right]$$

$$\leq \frac{n-r-3}{n}\,|a|\,\mathrm{E}\big[\mathrm{tr}(S^{+2})\big] + \frac{1}{n}|a|\,\mathrm{E}\big[\mathrm{tr}^2(S^{+})\big]\,.$$

Since $T \sim \mathrm{W}_r(n-1, \Omega/n)$, by Theorem 2.4.14 (viii) from Kollo and von Rosen [2006] we have the bound

$$\mathrm{E}\big[\mathrm{tr}(S^{+2})\big] \leq \mathrm{E}\big[\mathrm{tr}^2(S^{+})\big] = \mathrm{E}\big[\mathrm{tr}^2(T^{-1})\big] < \infty \tag{3.11}$$

when $n-r-4 > 0$, which holds since $r \leq n-5$. Therefore, the regularity condition hold and we can apply Theorem 4 to conclude that

$$\begin{aligned}
&\mathrm{E}\big[\|aS^{+} - \Sigma^{+}\|_F^2\big] = \mathrm{E}[U] \\
&\quad = \frac{2}{n} a\, \mathrm{E}\big[\mathrm{tr}^2(S^{+})\big] + \left(a^2 - 2\frac{n-r-3}{n}a\right) \mathrm{E}\big[\mathrm{tr}(S^{+2})\big]
\end{aligned}$$

for any $a \in \mathbb{R}$. Thus, in particular, the risk of the unbiased estimator $\frac{n-r-2}{n}S$ must equal $\frac{2(n-r-2)}{n^2}\,\mathrm{E}[\mathrm{tr}^2(S^{+})] - \frac{(n-r-2)(n-r-4)}{n^2}\,\mathrm{E}[\mathrm{tr}(S^{+2})]$. When $a \leq \frac{n-r-2}{n}$ we can bound

$$\begin{aligned}
&\mathrm{E}\big[\|aS^{+} - \Sigma^{+}\|_F^2\big] - \mathrm{E}\left[\left\|\frac{n-r-2}{n}S^{+} - \Sigma^{+}\right\|_F^2\right] \\
&\quad = \frac{2}{n}\left(a - \frac{n-r-2}{n}\right) \mathrm{E}\big[\mathrm{tr}^2(S^{+})\big] \\
&\qquad + \left(a^2 - 2\frac{n-r-3}{n}a + \frac{(n-r-2)(n-r-4)}{n^2}\right) \mathrm{E}\big[\mathrm{tr}(S^{+2})\big] \\
&\quad = \frac{2}{n}\left(a - \frac{n-r-2}{n}\right) \mathrm{E}\big[\mathrm{tr}^2(S^{+})\big] \\
&\qquad + \left(a - \frac{n-r-2}{n}\right)\left(a - \frac{n-r-4}{n}\right) \mathrm{E}\big[\mathrm{tr}(S^{+2})\big] \\
&\quad \leq \left(a - \frac{n-r-2}{n}\right)\left(a - \frac{n-r-6}{n}\right) \mathrm{E}\big[\mathrm{tr}(S^{+2})\big]\,,
\end{aligned}$$

which shows inequality 3.6. This upper bound has a minimum at $a = \frac{n-r-4}{n}$, which yields

$$\mathrm{E}\Big[\big\|aS^{+} - \Sigma^{+}\big\|_F^2\Big] - \mathrm{E}\left[\left\|\frac{n-r-2}{n}S^{+} - \Sigma^{+}\right\|_F^2\right] \leq\ -\frac{4}{n^2}\,\mathrm{E}\big[\mathrm{tr}(S^{+2})\big] <\ 0.$$

Thus $\frac{n-r-4}{n}S^{+}$ dominates $\frac{n-r-2}{n}S^{+}$, as desired. Moreover, the URE of S^{+} is

$\frac{2}{n}\mathrm{tr}^2(S^+) - \frac{n-2r-6}{n}\mathrm{tr}(S^{+2})$ and so

$$\begin{aligned}
&\mathrm{E}\left[\|\frac{n-r-2}{n}S^+ - \Sigma^+\|_F^2\right] - \mathrm{E}\left[\left\|S^+ - \Sigma^+\right\|_F^2\right] \\
&\quad = \mathrm{E}\left[-2\frac{r+2}{n^2}\mathrm{tr}^2(S^+) - \frac{(r+2)(r+4)}{n^2}\mathrm{tr}(S^{+2})\right] \qquad \leq\ 0,
\end{aligned}$$

so $\frac{n-r-2}{n}S^+$ dominates S^+, as claimed. □

Proof of Proposition 5. We have $\psi_k = a[1/l_k + t\mathrm{tr}^{-1}(S)]$, so

$$\begin{aligned}
\sum_{k=1}^r \psi_k^2 &= \frac{(n-r-4)^2}{n^2}\sum_{k=1}^r\left[\frac{1}{l_k^2} + \frac{2t}{l_k\mathrm{tr}(S)} + \frac{t^2}{\mathrm{tr}(S)}\right] \\
&= \frac{(n-r-4)^2}{n^2}\mathrm{tr}(S^{+2}) + 2\frac{(n-r-4)^2}{n^2}t\frac{\mathrm{tr}(S^+)}{\mathrm{tr}(S)} + \frac{(n-r-4)^2 r}{n^2}t^2\frac{1}{\mathrm{tr}^2(S)} \\
-2\frac{n-r-2}{n}\sum_{k=1}^r\frac{\psi_k}{l_k} &= -2\frac{(n-r-2)(n-r-4)}{n^2}\sum_{k=1}^r\left[\frac{1}{l_k^2} + \frac{t}{l_k\mathrm{tr}(S)}\right] \\
&= -2\frac{(n-r-2)(n-r-4)}{n^2}\,\mathrm{tr}(S^{+2}) - 2\frac{(n-r-2)(n-r-4)}{n^2}t\frac{\mathrm{tr}(S^+)}{\mathrm{tr}(S)} \\
-\frac{4}{n}\sum_{k=1}^r\frac{\partial\psi_k}{\partial l_k} &= -4\frac{n-r-4}{n^2}\sum_{k=1}^r\left[-\frac{1}{l_k^2} - \frac{t}{\mathrm{tr}^2(S)}\right] \\
&= 4\frac{n-r-4}{n^2}\,\mathrm{tr}(S^{+2}) + 4\frac{(n-r-4)r}{n^2}t\frac{1}{\mathrm{tr}^2(S)} \\
-\frac{2}{n}\sum_{k\neq b}^r\frac{\psi_k - \psi_b}{l_k - l_b} &= -2\frac{n-r-4}{n^2}\sum_{k\neq b}^r\frac{l_k^{-1} - l_b^{-1}}{l_k - l_b} \\
&= 2\frac{n-r-4}{n^2}\mathrm{tr}^2(S^+) - 2\frac{n-r-4}{n^2}\mathrm{tr}(S^{+2}).
\end{aligned}$$

Summing everything, we get the URE

$$\begin{aligned}
U = &\,2\frac{n-r-4}{n^2}\mathrm{tr}^2(S^+) - \frac{(n-r-4)(n-r-2)}{n^2}\mathrm{tr}(S^{+2}) \\
&+ 4\frac{n-r-4}{n^2}\left[r\frac{1}{\mathrm{tr}^2(S)} - \frac{\mathrm{tr}(S^+)}{\mathrm{tr}(S)}\right]t + \frac{(n-r-4)^2 r}{n^2}t^2\frac{1}{\mathrm{tr}^2(S)}.
\end{aligned}$$

Now note, using $\mathrm{tr}^{-1}(S) \leq \mathrm{tr}(S^+)/r^2$ and $\mathrm{tr}(S^{+2}) \leq \mathrm{tr}^2(S^+)$ that

$$\begin{aligned}
&\mathrm{E}\left[\left|\frac{n-r-2}{n}\sum_{k=1}^{p}\frac{\psi_k}{l_k} + \frac{2}{n}\sum_{k=1}^{r}\frac{\partial\psi_k}{\partial l_k} + \frac{1}{n}\sum_{k\neq b}\frac{\psi_k-\psi_b}{l_k-l_b}\right|\right] \\
&\quad= \frac{(n-r-3)(n-r-4)}{n^2}\,\mathrm{E}\big[\mathrm{tr}(S^{+2})\big] + \frac{n-r-4}{n^2}\,\mathrm{E}\big[\mathrm{tr}^2(S^+)\big] \\
&\qquad+ \frac{(n-r-2)(n-r-4)}{n^2}t\,\mathrm{E}\left[\frac{\mathrm{tr}(S^+)}{\mathrm{tr}(S)}\right] - 2\frac{(n-r-4)r}{n^2}t\,\mathrm{E}\left[\frac{1}{\mathrm{tr}^2(S)}\right] \\
&\quad\leq \Big(\frac{(n-r-1)(n-r-4)}{n^2} + \frac{(n-r-2)(n-r-4)}{r^2n^2}|t| \\
&\qquad\qquad + 2\frac{n-r-4}{r^3n^2}|t|\Big)\,\mathrm{E}\big[\mathrm{tr}^2(S^+)\big] \qquad < \infty,
\end{aligned}$$

since $\mathrm{E}[\mathrm{tr}^2(S^+)] < \infty$ by equation (3.11). Therefore, we can apply Theorem 4 to obtain

$$\begin{aligned}
&\mathrm{E}\left[\left\|\hat{\Sigma}_t^+ - \Sigma^+\right\|_F^2\right] = \mathrm{E}[U] \\
&\quad= 2\frac{n-r-4}{n^2}\,\mathrm{E}\big[\mathrm{tr}^2(S^+)\big] - \frac{(n-r-4)(n-r-2)}{n^2}\,\mathrm{E}\big[\mathrm{tr}(S^{+2})\big] \\
&\qquad+ 4\frac{n-r-4}{n^2}t\,\mathrm{E}\left[r\frac{1}{\mathrm{tr}^2(S)} - \frac{\mathrm{tr}(S^+)}{\mathrm{tr}(S)}\right] + \frac{(n-r-4)^2r}{n^2}t^2\,\mathrm{E}\left[\frac{1}{\mathrm{tr}^2(S)}\right]
\end{aligned}$$

for all $t \in \mathbb{R}$. Using that $n-r-4 > 0$ and $r^2\mathrm{tr}^{-1}(S) \leq \mathrm{tr}(S^+)$ again, we can bound the difference in risk as

$$\begin{aligned}
&\mathrm{E}\left[\left\|\hat{\Sigma}_t^+ - \Sigma^+\right\|_F^2\right] - \mathrm{E}\left[\left\|\hat{\Sigma}_{\mathrm{EM1}}^+ - \Sigma^+\right\|_F^2\right] \\
&\qquad\leq \frac{(n-r-4)r}{n^2}\Big[(n-r-4)t^2 - 4(r-1)t\Big]\,\mathrm{E}\left[\frac{1}{\mathrm{tr}^2(S)}\right]
\end{aligned}$$

which proves inequality (3.7). There is a minimum in t since $n-r-4 > 0$, which is $t = 2\frac{r-1}{n-r-4}$. In this case the quadratic coefficient and thus the difference in risk is strictly negative, so the corresponding estimator $\hat{\Sigma}_{\mathrm{EM2}} = \frac{n-r-4}{n}\left[S^+ + 2\frac{r-1}{n-r-4}\,\mathrm{tr}^{-1}(S)\right]$ dominates $\hat{\Sigma}_{\mathrm{EM1}}^+$, as desired. □

3.5.4 Proofs of Subsection 3.2.4

Proof of Theorem 5. Since T and S share the same non-zero eigenvalues, we can regard Ψ as a function of $T \sim \mathrm{W}_r(n-1, \Omega/n)$ only. Moreover, $\bar{X} = \mu + H\bar{Y}$. Using that $O_1 O_1^t = HH^t$ almost surely, we find

$$\begin{aligned}\mathrm{E}\left[\left\|\hat{\Sigma}^+\bar{X} - \Sigma^+\mu\right\|_2^2\right] &= \mathrm{E}\left[\left\|O_1O_1^tO_1\Psi O_1^tO_1O_1^t[\mu + H\bar{Y}] - H\Omega^{-1}H^t\mu\right\|_2^2\right] \\ &= \mathrm{E}\left[\left\|U\Psi U^t[H^t\mu + \bar{Y}] - \Omega^{-1}H^t\mu\right\|_2^2\right]\end{aligned}$$

Define $G = H^t\mu + \bar{Y} \sim \mathrm{N}_r(H^t\mu, \Omega/n)$ and notice it is independent of $U\Psi U^t$ since $\bar{X}$ and S are independent. Then

$$\begin{aligned}&= \mathrm{E}\left[\left\|U\Psi U^tW - \Omega^{-1}H^t\mu\right\|_2^2\right] \\ &= 2\,\mathrm{E}\big[(G - H^t\mu)^t\Omega^{-1}U\Psi U^tG\big] - 2\,\mathrm{E}\big[\mathrm{tr}\big(\Omega^{-1}U\Psi U^tGG^t\big)\big] \\ &\quad + \mathrm{E}\big[G^tU\Psi^2U^tG\big] - \mathrm{E}\big[(G - H^t\mu)^t\Omega^{-2}(G + H^t\mu)\big]\,.\end{aligned}$$

The first term can be handled as follows. By independence of G and $U\Psi U^t$, and Stein's lemma [Fourdrinier and Strawderman, 2003, Lemma A.1], we get

$$\begin{aligned}2\,\mathrm{E}\big[(G - H^t\mu)^t\Omega^{-1}U\Psi U^tG\big] &= \frac{2}{n}\mathrm{E}_G\left[(G - H^t\mu)^t\left[\frac{\Omega}{n}\right]^{-1}\mathrm{E}_T\left[U\Psi U^t\right]G\right] \\ &= \frac{2}{n}\mathrm{E}_G\left[\nabla_G G'\mathrm{E}_T\left[U\Psi U^t\right]\right] = \frac{2}{n}\mathrm{E}\mathrm{tr}\left[\Psi\right]\end{aligned}$$

under the condition

$$\mathrm{E}_G\left[\left|\nabla_G G'\mathrm{E}_T\left[U\Psi U^t\right]\right|\right] = \mathrm{E}_G\left[\left|\,\mathrm{tr}(\Psi)\,\right|\right] = \mathrm{E}\left[\left|\sum_{k=1}^r \psi_k\right|\right] < \infty.$$

For the second term, we will make use of the fact that

$$\mathrm{E}_T\left[\Omega^{-1}U\Psi U^t\right] = \mathrm{E}_T\left[U\Psi^*U^t\right], \tag{3.12}$$

where Ψ^* is defined as the statement. This is the result of a non-singular analogue of Theorem 2.2 from Konno [2009], or alternatively of a matrix analogue of Lemma

3 from Chételat and Wells [2014]. By appropriate modifications to the latter result and the underlying Lemma 3 from Chételat and Wells [2012] on which it depends, it can be seen that sufficient conditions for equation 3.12 to hold are

$$\mathrm{E}_T\left[\left|U\Psi^* U^t\right|_{ij}\right] < \infty \qquad\qquad \forall 1 \le i,j \le r.$$

A sufficient condition for this to happen is

$$\max_{1\le i,j\le r} \mathrm{E}_T\left[\left|U\Psi^* U^t\right|_{ij}\right] \le \mathrm{E}\left[\sum_{k=1}^{r}\left|\psi_k^*\right|\right] < \infty.$$

Then, using the independence of G and T, we can conclude

$$\begin{aligned} -2\,\mathrm{E}\big[\mathrm{tr}\big(\Omega^{-1}U\Psi U^t G G^t\big)\big] &= -2\,\mathrm{tr}\big(\mathrm{E}_T\left[\Omega^{-1}U\Psi U^t\right]\mathrm{E}_G\left[GG^t\right]\big) \\ &= -2\,\mathrm{tr}\big(\mathrm{E}_T\left[U\Psi^* U^t\right]\mathrm{E}_G\left[GG^t\right]\big) = -2\,\mathrm{E}\big[G^t U\Psi^* U^t G\big]. \end{aligned}$$

Thus

$$\begin{aligned} \mathrm{E}\left[\left\|\hat{\Sigma}^+\bar{X} - \Sigma^+\mu\right\|_F^2\right] = \frac{2}{n}\,\mathrm{E}[\mathrm{tr}(\Psi)] - 2\,\mathrm{E}\big[G^t U\Psi^* U^t G\big] + \mathrm{E}\big[G^t U\Psi^2 U^t G\big] \\ - \mathrm{E}\big[(G - H^t\mu)^t\Omega^{-2}(G + H^t\mu)\big]. \end{aligned}$$

But $U^t G = O_1^t H[H^t\mu + \bar{Y}] = O_1^t\bar{X}$ and $(G - H^t\mu)^t\Omega^{-2}(G + H^t\mu) = (G - H^t\mu)^t H^t\Sigma^{+2}H(G + H^t\mu) = (\bar{X} - \mu)^t\Sigma^{+2}(\bar{X} + \mu)$. Hence

$$\begin{aligned} \mathrm{E}\left[\left\|\hat{\Sigma}^+\bar{X} - \Sigma^+\mu\right\|_2^2\right] = \mathrm{E}\left[\frac{2}{n}\mathrm{tr}\,\hat{\Sigma}^+ + \bar{X}^t O_1(\Psi^2 - 2\Psi^*)O_1^t\bar{X}\right] \\ - \mathrm{E}\big[(\bar{X} - \mu)^t\Sigma^{+2}(\bar{X} + \mu)\big]. \end{aligned}$$

This proves the result. □

Proof of Proposition 6. We have $\psi_k = a/l_k$, so

$$\begin{aligned} \psi_k^* &= \frac{n-r-2}{n}\frac{\psi_k}{l_k} + \frac{2}{n}\frac{\partial\psi_k}{\partial l_k} + \frac{1}{n}\sum_{b\neq k}^{r}\frac{\psi_k - \psi_b}{l_k - l_b} \\ &= \frac{n-r-2}{n}\frac{1}{l_k^2}a - \frac{2}{n}\frac{1}{l_k^2}a - \frac{1}{n}\frac{\mathrm{tr}(S^+)}{l_k}a + \frac{1}{n}\frac{1}{l_k^2}a \\ &= \frac{n-r-3}{n}\frac{1}{l_k^2}a - \frac{1}{n}\frac{\mathrm{tr}(S^+)}{l_k}a. \end{aligned}$$

We can bound

$$\mathrm{E}\left[\left|\sum_{k=1}^{r}\psi_k\right|\right] = |a|\,\mathrm{E}\big[\mathrm{tr}(S^{+})\big] \leq\ |a|\,\mathrm{E}\big[\mathrm{tr}^2(S^{+})\big]^{\frac{1}{2}},$$

$$\mathrm{E}\left[\sum_{k=1}^{r}\left|\psi_k^*\right|\right] \leq \frac{n-r-3}{n}|a|\,\mathrm{E}\big[\mathrm{tr}(S^{+2})\big] + \frac{1}{n}|a|\,\mathrm{E}\big[\mathrm{tr}^2(S^{+})\big],$$

so by inequality (3.11) and the fact that $n-r-4>0$ these two expressions are finite. Therefore, we can apply the results of Theorem 5 to obtain

$$\begin{aligned}
\mathrm{E}\left[\left\|aS^{+}\bar{X}-\Sigma^{+}\mu\right\|_2^2\right] &= \frac{2}{n}a\,\mathrm{E}\big[\mathrm{tr}(S^{+})\big] \\
&\quad + \mathrm{E}\left[\sum_{k=1}^{r}\left(\frac{a}{l_k^2} - 2\frac{n-r-3}{n}\frac{1}{l_k^2} + \frac{2}{n}\frac{\mathrm{tr}(S^{+})}{l_k}\right)a\left(O_1^t\bar{X}\bar{X}^tO_1\right)_{kk}\right] \\
&\quad - \mathrm{E}\Big[(\bar{X}-\mu)^t\Sigma^{+2}(\bar{X}+\mu)\Big] \\
&= \frac{2}{n}a\,\mathrm{E}\big[\mathrm{tr}(S^{+})\big] + \left(a^2 - 2\frac{n-r-3}{n}a\right)\mathrm{E}\big[\bar{X}^tS^{+2}\bar{X}\big] \\
&\quad + \frac{2}{n}a\ \mathrm{E}\Big[\mathrm{tr}(S^{+})\bar{X}^tS^{+}\bar{X}\Big] - \mathrm{E}\Big[(\bar{X}-\mu)^t\Sigma^{+2}(\bar{X}+\mu)\Big]
\end{aligned}$$

for any $a\in\mathbb{R}$. Therefore, for $a\leq\frac{n-r-2}{n}$ we can bound the difference in risk by

$$\begin{aligned}
&\mathrm{E}\left[\left\|aS^{+}\bar{X}-\Sigma^{+}\mu\right\|_2^2\right] - \mathrm{E}\left[\left\|\frac{n-r-2}{n}S^{+}\bar{X}-\Sigma^{+}\mu\right\|_2^2\right] \\
&\quad = \frac{2}{n}\left(a-\frac{n-r-2}{n}\right)\mathrm{E}\big[\mathrm{tr}(S^{+})\big] \\
&\qquad + \left(a^2 - 2\frac{n-r-3}{n}a + \frac{(n-r-2)(n-r-4)}{n^2}\right)\mathrm{E}\big[\bar{X}^tS^{+2}\bar{X}\big] \\
&\qquad + \frac{2}{n}\left(a-\frac{n-r-2}{n}\right)\mathrm{E}\Big[\mathrm{tr}(S^{+})\bar{X}^tS^{+}\bar{X}\Big] \\
&\quad \leq\ \left(a-\frac{n-r-2}{n}\right)\left(a-\frac{n-r-4}{n}\right)\mathrm{E}\big[\bar{X}^tS^{+2}\bar{X}\big],
\end{aligned}$$

which proves inequality (3.8). The quadratic coefficient is minimized at $a=\frac{n-r-3}{n}$, at which point we have

$$\begin{aligned}
&\mathrm{E}\left[\left\|\frac{n-r-3}{n}S^{+}\bar{X}-\Sigma^{+}\mu\right\|_2^2\right] - \mathrm{E}\left[\left\|\frac{n-r-2}{n}S^{+}\bar{X}-\Sigma^{+}\mu\right\|_2^2\right] \\
&\qquad \leq\ -\frac{1}{n^2}\,\mathrm{E}\big[\bar{X}^tS^{+2}\bar{X}\big] \qquad <\ 0.
\end{aligned}$$

Thus $\frac{n-r-3}{n}S^+\bar{X}$ dominates $\frac{n-r-2}{n}S^+\bar{X}$, as desired. Moreover,

$$\begin{aligned}
&\mathrm{E}\left[\|\frac{n-r-2}{n}S^+\bar{X}-\Sigma^+\bar{X}\|_2^2\right]-\mathrm{E}\left[\left\|S^+\bar{X}-\Sigma^+\bar{X}\right\|_2^2\right]\\
&\quad=-2\frac{r+2}{n^2}\,\mathrm{E}\big[\mathrm{tr}(S^+)\big]-\frac{(r+2)(r+4)}{n^2}\,\mathrm{E}\big[\bar{X}^tS^{+2}\bar{X}\big]\\
&\qquad-2\frac{r+2}{n^2}\,\mathrm{E}\Big[\mathrm{tr}(S^+)\bar{X}^tS^+\bar{X}\Big]\qquad<\;0,
\end{aligned}$$

so $\frac{n-r-2}{n}S^+$ dominates S^+, as claimed. □

Proof of Proposition 7. We will apply 4, and we have here $\psi_k = \frac{n-r-3}{n}[1/l_k + t\mathrm{tr}^{-1}(S)]$ for $1 \le k \le r$, so

$$\begin{aligned}
\frac{n-r-2}{n}\frac{\psi_k}{l_k}&=\frac{(n-r-2)(n-r-3)}{n^2}\left[\frac{1}{l_k^2}+\frac{t}{l_k\mathrm{tr}(S)}\right],\\
\frac{2}{n}\sum_{k=1}^{r}\frac{\partial\psi_k}{\partial l_k}&=2\frac{n-r-3}{n^2}\left[-\frac{1}{l_k^2}-\frac{t}{\mathrm{tr}^2(S)}\right],\\
\frac{1}{n}\sum_{b\neq k}^{r}\frac{\psi_k-\psi_b}{l_k-l_b}&=\frac{n-r-3}{n^2}\sum_{b\neq k}^{r}\frac{l_k^{-1}-l_b^{-1}}{l_k-l_b}=\frac{n-r-3}{n^2}\left[\frac{1}{l_k^2}-\frac{\mathrm{tr}(S^+)}{l_k}\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\psi_k^*&=\frac{n-r-2}{n}\frac{\psi_k}{l_k}+\frac{2}{n}\frac{\partial\psi_k}{\partial l_k}+\frac{1}{n}\sum_{b\neq k}^{r}\frac{\psi_k-\psi_b}{l_k-l_b}\\
&=\frac{(n-r-3)^2}{n^2}\frac{1}{l_k^2}+\frac{(n-r-2)(n-r-3)}{n^2}t\,\frac{\mathrm{tr}^{-1}(S)}{l_k}\\
&\quad-2\frac{n-r-3}{n^2}t\,\mathrm{tr}^{-2}(S)-\frac{n-r-3}{n^2}\frac{\mathrm{tr}(S^+)}{l_k}
\end{aligned}$$

We can bound

$$\mathrm{E}\left[\left|\sum_{k=1}^{r}\psi_k\right|\right]\le\;\frac{n-r-3}{n}\,\mathrm{E}\big[\mathrm{tr}(S^+)\big]+\frac{(n-r-3)r}{n}|t|\,\mathrm{E}\big[\mathrm{tr}^{-1}(S)\big],$$

$$\begin{aligned}
\mathrm{E}\left[\sum_{k=1}^{r}\left|\psi_k^*\right|\right]\le\;&\frac{(n-r-3)^2}{n^2}\,\mathrm{E}\big[\mathrm{tr}(S^{+2})\big]+\frac{n-r-3}{n^2}\,\mathrm{E}\big[\mathrm{tr}^2(S^+)\big]\\
&+\frac{(n-r)(n-r-3)}{n^2}|t|\;\mathrm{E}\big[\mathrm{tr}^{-2}(S)\big],
\end{aligned}$$

so by $\mathrm{tr}^{-1} \leq \mathrm{tr}(S^+)/r^2$, inequality (3.11) and the fact that $n-r-4>0$ these two expressions are finite. Therefore, we can apply the results of Theorem 5 to obtain

$$
\begin{aligned}
\mathrm{E}\left[\left\|\hat{\eta}_t - \eta\right\|_2^2\right] &= 2\frac{n-r-3}{n^2}\,\mathrm{E}\big[\mathrm{tr}(S^+)\big] + 2\frac{(n-r-3)r}{n^2}t\,\mathrm{E}\big[\mathrm{tr}^{-1}(S)\big] \\
&\quad + \mathrm{E}\Bigg[\sum_{k=1}^{r}\Bigg(\frac{(n-r-3)^2}{n^2}\frac{1}{l_k^2} + 2\frac{(n-r-3)^2}{n^2}t\frac{\mathrm{tr}^{-1}(S)}{l_k} \\
&\qquad + \frac{(n-r-3)^2}{n^2}t^2\mathrm{tr}^{-2}(S) - 2\frac{(n-r-3)^2}{n^2}\frac{1}{l_k^2} \\
&\qquad - 2\frac{(n-r-2)(n-r-3)}{n^2}t\,\frac{\mathrm{tr}^{-1}(S)}{l_k} + 4\frac{n-r-3}{n^2}t\,\mathrm{tr}^{-2}(S) \\
&\qquad + 2\frac{n-r-3}{n^2}\frac{\mathrm{tr}(S^+)}{l_k}\Bigg)\left(O_1^t\bar{X}\bar{X}^tO_1\right)_{kk}\Bigg] \\
&\quad - \mathrm{E}\Big[(\bar{X}-\mu)^t\Sigma^{+2}(\bar{X}+\mu)\Big] \\
&= 2\frac{n-r-3}{n^2}\,\mathrm{E}\big[\mathrm{tr}(S^+)\big] + 2\frac{(n-r-3)r}{n^2}t\,\mathrm{E}\big[\mathrm{tr}^{-1}(S)\big] \\
&\quad + \mathrm{E}\Bigg[\sum_{k=1}^{r}\Bigg(-\frac{(n-r-3)^2}{n^2}\frac{1}{l_k^2} + 2\frac{n-r-3}{n^2}\frac{\mathrm{tr}(S^+)}{l_k} \\
&\qquad - 2\frac{n-r-3}{n^2}t\frac{\mathrm{tr}^{-1}(S)}{l_k} + 4\frac{n-r-3}{n^2}t\,\mathrm{tr}^{-2}(S) \\
&\qquad + \frac{(n-r-3)^2}{n^2}t^2\mathrm{tr}^{-2}(S)\Bigg)\left(O_1^t\bar{X}\bar{X}^tO_1\right)_{kk}\Bigg] \\
&\quad - \mathrm{E}\Big[(\bar{X}-\mu)^t\Sigma^{+2}(\bar{X}+\mu)\Big] \\
&= 2\frac{n-r-3}{n^2}\,\mathrm{E}\big[\mathrm{tr}(S^+)\big] - \frac{(n-r-3)^2}{n^2}\,\mathrm{E}\big[\bar{X}^tS^{+2}\bar{X}\big] \\
&\quad + 2\frac{n-r-3}{n^2}\,\mathrm{E}\big[\mathrm{tr}(S^+)\bar{X}^tS^+\bar{X}\big] + \Bigg(2\frac{(n-r-3)r}{n^2}\,\mathrm{E}\big[\mathrm{tr}^{-1}(S)\big] \\
&\quad - 2\frac{n-r-3}{n^2}\,\mathrm{E}\left[\frac{\bar{X}^tS^+\bar{X}}{\mathrm{tr}(S)}\right] + 4\frac{n-r-3}{n^2}\,\mathrm{E}\left[\frac{\bar{X}^t\bar{X}}{\mathrm{tr}^2(S)}\right]\Bigg)t \\
&\quad + \frac{(n-r-3)^2}{n^2}t^2\,\mathrm{E}\left[\frac{\bar{X}^t\bar{X}}{\mathrm{tr}^2(S)}\right] - \mathrm{E}\Big[(\bar{X}-\mu)^t\Sigma^{+2}(\bar{X}+\mu)\Big]
\end{aligned}
$$

for any $t \in \mathbb{R}$. Therefore, the difference in risk can be written

$$
\begin{aligned}
&\mathrm{E}\left[\left\|\hat{\eta}_t - \eta\right\|_2^2\right] - \mathrm{E}\left[\left\|\frac{n-r-3}{n}S^+\bar{X} - \eta\right\|_2^2\right] \\
&\quad = \Bigg(2\frac{(n-r-3)r}{n^2}\,\mathrm{E}\big[\mathrm{tr}^{-1}(S)\big] - 2\frac{n-r-3}{n^2}\,\mathrm{E}\left[\frac{\bar{X}^tS^+\bar{X}}{\mathrm{tr}(S)}\right]
\end{aligned}
$$

$$+ 4\frac{n-r-3}{n^2}\,\mathrm{E}\left[\frac{\bar{X}^t\bar{X}}{\mathrm{tr}^2(S)}\right]\Big)t + \frac{(n-r-3)^2}{n^2}t^2\,\mathrm{E}\left[\frac{\bar{X}^t\bar{X}}{\mathrm{tr}^2(S)}\right].$$

Note that $\mathrm{tr}(\bar{X}\bar{X}^t) = \mathrm{tr}(SS^+\bar{X}\bar{X}^t) \le \mathrm{tr}^{\frac{1}{2}}(S^2)\mathrm{tr}^{\frac{1}{2}}([S^+\bar{X}\bar{X}^t]^2) \le \mathrm{tr}(S)\mathrm{tr}(S^+\bar{X}\bar{X}^t)$, so we can bound

$$\begin{aligned}\le\ & 2\frac{(n-r-3)r}{n^2}t\,\mathrm{E}\big[\mathrm{tr}^{-1}(S)\big] + 2\frac{n-r-3}{n^2}t\,\mathrm{E}\left[\frac{\bar{X}^t\bar{X}}{\mathrm{tr}^2(S)}\right]\\ &+ \frac{(n-r-3)^2}{n^2}t^2\,\mathrm{E}\left[\frac{\bar{X}^t\bar{X}}{\mathrm{tr}^2(S)}\right].\end{aligned}$$

Next, write the reduced singular value decomposition of X as $X = \sqrt{n}V_1L^{1/2}O_1$ with V_1 $n\times r$ semi-orthogonal, $V_1^tV_1 = I_r$. Then

$$\begin{aligned}\bar{X}^t\bar{X} &= \mathrm{tr}\Big(X^t\frac{1_n1_n^t}{n^2}X\Big) = \mathrm{tr}\Big(LV_1^t\frac{1_n1_n^t}{n}V_1\Big)\\ &\le \mathrm{tr}(L)\sigma_{\max}\left(V_1^t\frac{1_n1_n^t}{n}V_1\right) \le \mathrm{tr}(S)\sigma_{\max}\left(\frac{1_n1_n^t}{n}\right) = \mathrm{tr}(S).\end{aligned}$$

Therefore, we can bound by

$$\le\ \frac{(n-r-3)}{n^2}\left[2(r+1)t + (n-r-3)t^2\right]\mathrm{E}\left[\frac{1}{\mathrm{tr}(S)}\right],$$

which proves (3.9). Since $n-r-3>0$, the quadratic coefficient has a minimum, at $t = -\frac{r+1}{n-r-3}$. In this case we have

$$\begin{aligned}&\mathrm{E}\left[\left\|\frac{n-r-3}{n}\left[S^+ - \frac{(r+1)\mathrm{tr}^{-1}(S)}{n-r-3}\right]\bar{X} - \eta\right\|_2^2\right] - \mathrm{E}\left[\left\|\frac{n-r-3}{n}S^+\bar{X} - \eta\right\|_2^2\right]\\ &\quad\le -\frac{(r+1)^2}{n^2}\,\mathrm{E}\left[\frac{1}{\mathrm{tr}(S)}\right] \qquad < \ 0.\end{aligned}$$

Thus $\hat{\eta}_{\mathrm{TK2}} = \frac{n-r-3}{n}\left[S^+ - \frac{r+1}{n-r-3}\mathrm{tr}^{-1}(S)\right]$ dominates $\hat{\eta}_{\mathrm{TK1}}$, as desired. □

CHAPTER 4

NOISE ESTIMATION IN THE SPIKED COVARIANCE MODEL

4.1 Introduction

The estimation of covariance matrices in a high dimensional framework has seen a surge of interest in the past years. The natural estimator, the sample covariance matrix, is well known to be inadequate in this context. The problem has been well studied under many sparsity scenarios: for example, zeros in the coordinates of the matrix [Bickel and Levina, 2008b, El Karoui, 2008b, Rothman et al., 2009, Cai and Liu, 2011] or its inverse [Meinshausen and Bühlmann, 2006, Friedman et al., 2008, Cai et al., 2011, Ravikumar et al., 2011, Rothman et al., 2008], bandedness [Bickel and Levina, 2008a, Bien et al., 2014] and many others. This paper will focus on the spiked model, first introducted by Johnstone [2001].

In the spiked model, the $p \times p$ covariance matrix Σ has distinct eigenvalues $\gamma_1 + \sigma^2 > ... > \gamma_\rho + \sigma^2$, and a smallest eigenvalue σ^2 of multiplicity $p - \rho$. It often provides good approximations in low and high dimensional settings, with small ρ being seen as a form of low rank sparsity in the data. It is also of substantial theoretical interest, being one of the few non-trivial settings in which random matrix theory has been extensively studied.

A related problem is principal components analysis. In PCA, one estimates eigenvectors associated with large eigenvalues of Σ, and perform dimension reduction using a truncated spectral decomposition. A traditional problem with the technique is that the number of eigenvectors to retain is not clear. However, if the true covariance matrix Σ is spiked, it is natural to associate its spiked rank ρ with the ideal number of eigenvectors to select, recasting the selection of the number of components as a rigorous statistical estimation problem.

Successful high-dimensional PCA usually requires good estimation of σ^2 (see e.g. Johnstone and Lu [2009]), a problem we will refer to as noise estimation. Although distinct from estimation of the covariance matrix itself, there is a context in which these two problems, estimation of Σ and σ^2, are analogous.

This context is as follows. Asymptotics are high-dimensional in the sense that p tends to infinity with the sample size n; for mathematical convenience we focus on the regime where the ratio p/n tends to a strictly positive constant as $n \to \infty$. The noise estimation problem is to estimate σ^2 under, say, absolute error loss $L(\hat{\sigma}^2, \sigma^2) = |\hat{\sigma}^2 - \sigma^2|$, while the covariance problem is to estimate the spiked Σ under the Frobenius loss $L_F(\hat{\Sigma}, \Sigma) = \|\hat{\Sigma} - \Sigma\|_F^2/p$ using a spiked estimator. This normalization is natural in this setting, since under normality the risk $\mathrm{E}[L_F(S, \Sigma)]$ of the sample covariance matrix S tends to a strictly positive constant.

Then, in essence, all that really matters in the covariance estimation problem is estimation of the noise level. Indeed, consider two spiked estimators $\hat{\Sigma}_i = \hat{\Gamma}_i + \hat{\sigma}^2 I$, $i = 1, 2$ with asymptotically finite spiked parts $\hat{\Gamma}_i$, which we take to mean that their ranks $\hat{\rho}_i = \mathrm{rk}(\hat{\Gamma}_i)$ and largest eigenvalues $\lambda_1(\hat{\Gamma}_i)$ are asymptotically finite. Then

$$\frac{\|\hat{\Sigma}_1 - \hat{\Sigma}_2\|_F^2}{p} \leq \frac{\hat{\rho}_1 + \hat{\rho}_2}{p} \Big[\lambda_1(\hat{\Gamma}_1) + \lambda_1(\hat{\Gamma}_2)\Big]^2 \underset{n\to\infty}{\longrightarrow} 0 \qquad \text{a.s.} \tag{4.1}$$

This means we can interpret the two problems as asymptotically analogous in practice. This reasoning is short of being a formal result of equivalence, but will serve as a guiding principle.

We propose a solution to these parallel problems as follows. We first restrict ourselves to orthogonally invariant estimators of the spiked form; this large class can be thought as performing spiked corrections of the eigenvalues of the sample covariance matrix. For this class, there exists an unbiased risk estimator (URE) in the closely related invariant loss $L_H(\hat{\Sigma}, \Sigma) = \mathrm{tr}\big[(\hat{\Sigma}\Sigma^{-1} - I)^2\big]/p$, which we will

refer to as the Haff loss. We propose to find an optimal choice of noise estimator by minimizing this URE using calculus of variations. This approach is close in spirit to the work of Stein [1975, 1986], where he considers a loss based on a normal log-likelihood, although it is not specifically high-dimensional. It is also close to the Bayesian approach of Haff [1991]. More generally, the idea to correct the eigenvalues of the sample covariance matrix is also found in previous work by Ledoit and Wolf [2004], El Karoui [2008a], Ledoit and Wolf [2012] and Donoho et al. [2014].

The URE of the covariance estimator depends on first and second derivatives of the noise estimator, so directly minimizing the risk would yield an estimator that depends on the truth. It however happily turns out that the "dominant" part of this URE does not depend on the derivatives. It is therefore possible to obtain, in closed form, an estimator optimal for the dominant part of the URE.

We prove that our proposed estimator is well-behaved; for example, it is strongly consistent for σ^2, even if the chosen estimators of γ_k and ρ are not. It is moreover essentially asymptotically normal of rate n, and we prove that this is the optimal minimax rate for the noise estimation problem. To illustrate concretely why this approach is interesting, we use it to construct a robust spiked covariance estimator. It seems to never perform worse than S in general, even in worst-case scenarios; while it performs remarkably well in spiked settings, and we show its eigenvalues are consistent.

We reiterate that in contrast with much work in high dimensional covariance estimation, we do not work with a sparsity assumption that many components of Σ or Σ^{-1} are zero. However, one can perfectly think of a spiked structure as a form of sparsity in itself, with ρ as sparsity parameter, which fits within the generally accepted principle that improved estimation in high dimensions is

difficult unless some form of sparsity holds with the truth. The fact that we can construct an estimator that can exploit that structure when present, yet be robust to the assumption is encouraging.

The chapter is divided as follows. The regularity conditions, construction of the unbiased risk estimator and construction of the noise estimator is in Section 4.2. Investigations of properties of the noise estimator is done in Section 4.3. The example construction and simulations are in Section 4.4. After some comments in Section 4.5, we cover the proofs of the claims in Section 4.6.

Notation The following notation will be used throughout. We write $H_p(\mathbb{R})$ for the simplex $\left\{x \in \mathbb{R}^p \,\middle|\, x_1 > ... > x_p > 0\right\}$. The real p-dimensional orthogonal group is denoted $O_p(\mathbb{R})$. The Frobenius norm of a matrix A is the sum of its squared eigenvalues, denoted $\|A\|_F = \mathrm{tr}(A^2)^{1/2}$, while the spectral norm is its largest singular value, denoted $\|A\|_2 = \sigma_{\max}(A)$. The notation $\mathrm{d_{TV}}(\mu_1, \mu_2)$ stands for the total variation distance between two probability measures μ_1, μ_2 on an underlying measurable space $(\Omega, \mathcal{B})$, which equals $\sup_{A \in \mathcal{B}} |\mu_1(A) - \mu_2(A)|$. The p-dimensional Wishart distribution with n degrees of freedom and covariance matrix Σ is written $W_p(n, \Sigma)$.

4.2 Construction

We work in the following setting. Assume the data is an i.i.d. sample $X_1, ..., X_n \sim \mathrm{N}_p(0, \Sigma_p)$, with $n \geq p$ and $\Sigma_p > 0$. For such a sample, one can stack the data into a matrix $X = (X_1', ..., X_n')$ and let $S = X'X/n = OLO'$, $L = \mathrm{diag}(l_1, ..., l_p)$ be the decreasing spectral decomposition of the sample covariance matrix, with $l_1 > ... > l_p > 0$ its ordered eigenvalues. The random matrix S, which is distributed as a scaled Wishart $n^{-1}\mathrm{W}_p(n, \Sigma_p)$, serves as a naive estimator of Σ upon which

we wish to improve. The normality and restriction to $n \geq p$ are necessary for the construction of the unbiased risk estimator that will follow; extensions will be discussed in Section 4.5.

As mentioned in the introduction, to adequately discuss high-dimensional behavior, we will also let this setting grow in complexity. We focus our attention on full-rank linear regimes, where a sequence of positive-definite covariance matrices of growing dimension $\Sigma_1, \Sigma_2, \Sigma_3, ...$ is fixed; and $p = p_n$, as a function of the sample size, grows in the sense that $p_n/n \to c$ for some $c \in (0, 1)$. It will then be assumed that for every (n, p_n), some i.i.d. sample $X_1, ..., X_n \sim \mathrm{N}_{p_n}(0, \Sigma_{p_n})$ will be available and a corresponding sample covariance matrix S constructed.

For such settings, the sequence $\{\Sigma_p\}$ is completely arbitrary beyond the requirement that each member be positive-definite. Of particular interest to us is the case where the covariance matrices form a *spiked sequence*, which we define as follows.

Definition 1. A sequence of covariance matrices $\{\Sigma_p\}$ is spiked if there exists a collection $\gamma_1 > ... > \gamma_\rho > 0$ of size $\rho \geq 0$ and a $\sigma^2 > 0$ such that for any p, $\Sigma_p = \mathrm{diag}(\gamma, 0, ..., 0) + \sigma^2 I_p$, where $\gamma = (\gamma_1, ..., \gamma_\rho)$.

When discussing asymptotics, we will sometimes need that the spiked eigenvalues $\gamma_1, ..., \gamma_\rho$ be sufficiently large with respect to the noise for efficient estimation to be possible. In practice, this will mean requiring that $\gamma_\rho/\sigma^2 > \sqrt{c}$, for c the asymptotic p_n/n ratio. The importance of this supercriticality condition for spiked eigenvalue estimation was first remarked by Baik et al. [2005] before being extended to the setting we are considering by Baik and Silverstein [2006], Paul [2007] and Nadler [2008]. They showed that for $1 \leq k \leq \rho$, the eigenvalues of the sample

covariance matrix satisfy

$$l_k \xrightarrow[n\to\infty]{\text{a.s.}} \begin{cases} \left(1 + c\sigma^2 \frac{\gamma_k + \sigma^2}{\gamma_k}\right)[\gamma_k + \sigma^2] & \text{if } \gamma_k > \sqrt{c}\sigma^2 \\ (1+\sqrt{c})^2\sigma^2 & \text{if } \gamma_k \leq \sqrt{c}\sigma^2 \end{cases}, \tag{4.2}$$

with the $l_{\rho+1}, ..., l_p$ asymptotically distributed like a scaled Marchenko-Pastur $\sigma^2\text{MP}(c)$ distribution. Therefore, the asymptotic spectrum of S do not contain any information about those γ_k below the critical threshold $\sqrt{c}\sigma^2$. But since their estimation is mostly tangential to our goals, supercriticality will not always be necessary, and we will make it clear when it will be.

Let us now turn our attention to the task at hand. The parallel problems we wish to solve are

(i) the estimation of σ^2 under the absolute error loss $L(\hat{\sigma}^2, \sigma^2) = |\hat{\sigma}^2 - \sigma^2|$;

(ii) the estimation of Σ under the Frobenius loss $L_F(\hat{\Sigma}, \Sigma) = \|\hat{\Sigma} - \Sigma\|_F^2/p$ using spiked estimators.

Under spikedness, these two problems are parallel to each other in the sense of (4.1). The approach we take begins with aspect (ii) - we seek a good covariance estimator $\hat{\Sigma}$ in spiked form $\hat{\Gamma} + \hat{\sigma}^2 I$, with $\hat{\rho} = \text{rk}(\hat{\Gamma})$ small with respect to p, which we interpret as $\hat{\rho}$ a.s. tending to a finite constant. By appealing again to (4.1), it is clear that for the Frobenius loss, the spiked part $\hat{\Gamma}$ is asymptotically dominated by noise estimation. We therefore might as well *choose* $\hat{\Gamma}$ based on convenience: for example, we can pick one with consistent eigenvalue estimators, or some other property. A specific choice will be considered in Section 4.4. Alternatively, the recent results of Donoho et al. [2014] could provide an attractive choice based on consideration of single eigenvalue corrections, and we comment on this further in Section 4.5. Once a choice of $\hat{\Gamma}$ is made, we can then look for an optimal $\hat{\sigma}^2$, which would simultaneously solve aspects ((i)) and ((ii)) of the problem, while being asymptotically independent of our choice of $\hat{\Gamma}$.

Being quite free in selecting the spiked part, let us focus on mathematically convenient possibilities. A first restriction is to take $\hat{\Gamma}$ orthogonally invariant – that is, of the form $\hat{\Gamma} = O\text{diag}(\hat{\gamma}, 0)\, O'$ for some estimators $\hat{\gamma}_1 > ... > \hat{\gamma}_{\hat{\rho}} > 0$ and $O \in O_p(\mathbb{R})$ the matrix of ordered eigenvectors of S. With this choice, our estimators $\hat{\Sigma}$ can be thought as performing spiked corrections on the sample covariance matrix S.

A second restriction will be necessary. At this point in our discussion the spiked, rank and noise estimators $\hat{\gamma}$, $\hat{\rho}$, $\hat{\sigma}^2$ have been essentially arbitrary. This is too general for the construction of the URE that will follow, so we must to restrict ourselves to sufficiently regular estimators. The regularity conditions come in two flavors, weak and strong, and are statements of integrability; these conditions simply guarantee that expected values appearing in the construction of the URE are convergent. Combining the invariance and regularity restrictions, we define the following.

Definition 2. A spiked eigenvalue estimator $\hat{\Gamma}$ satisfies the weak regularity conditions if it is of the form $\hat{\Gamma} = O\text{diag}(\hat{\gamma})\, O'$ for $S = OLO'$ and satisfies the following. Let $\hat{\rho} = \text{rk}(\hat{\Gamma})$. For each $1 \leq k \leq p$, $\hat{\gamma}_k$ are a.s. $C^2(H_p(\mathbb{R}); \mathbb{R})$ functions of $l_1, ..., l_p$ with boundary cases $\mathbb{1}[\hat{\rho} < k]\, \hat{\gamma}_k = 0$ and $\mathbb{1}[\hat{\rho} = p]\, \hat{\gamma}_k = \mathbb{1}[\hat{\rho} = p]\, l_k$ for which both expectations

$$\text{E}\left[\left|\frac{\hat{\gamma}_k}{l_k}\right|^{9(1+\epsilon)}\right] \text{ and } \text{E}\left[\left|\frac{\partial \hat{\gamma}_k}{\partial l_k}\right|^{4.5}\right]$$

are finite for some $\epsilon > 0$. Similarly, a noise estimator $\hat{\sigma}^2$ satisfies the weak regularity conditions for a weak spiked eigenvalue estimator $\hat{\Gamma}$ if it is a $C^2(H_p(\mathbb{R}); \mathbb{R})$ function based on $l_1, ..., l_p$ such that for each $1 \leq k \leq p$,

$$\text{E}\left[\left|\frac{\hat{\sigma}^2}{l_k}\right|^{9(1+\epsilon)}\right], \text{E}\left[\left|\frac{\partial \hat{\sigma}^2}{\partial l_k}\right|^{4.5}\right] \text{ and } \text{E}\left[\left|\hat{\gamma}_k + \hat{\sigma}^2\right| \left|\frac{\partial^2 \hat{\gamma}_k}{\partial l_k^2} + \frac{\partial^2 \hat{\sigma}^2}{\partial l_k^2}\right|\right]$$

are all finite for some $\epsilon > 0$.

The previous conditions assert integrability of quantities associated with the estimators for a given p, and are dimension dependent. In contrast, the following conditions assert similar integrability as p grows. To make the dependence explicit, we superscript the dimension.

Definition 3. A spiked eigenvalue estimator $\hat{\Gamma}^p$ satisfies the strong regularity conditions if it satisfies the weak regularity conditions for each $p > 0$, and moreover

$$\mathrm{E}\left[\sup_{p>0}\max_{1\le k\le p}\left|\frac{\hat{\gamma}_k^p}{l_k^p}\right|^{9(1+\epsilon)}\right],\ \mathrm{E}\left[\sup_{p>0}\max_{1\le k\le p}\left|\frac{\partial\hat{\gamma}_k^p}{\partial l_k^p}\right|^{4.5}\right],\ \mathrm{E}\left[\sup_{p>0}\max_{1\le k\le p}\left|\hat{\gamma}_k^p\frac{\partial^2\hat{\gamma}_k^p}{\partial l_k^{p2}}\right|\right],$$
$$\mathrm{E}\left[\sup_{p>0}\max_{1\le k\ne b\le\hat{\rho}}\left|\frac{\hat{\gamma}_k^p-\hat{\gamma}_b^p}{l_k^p-l_b^p}\right|^2\right],\ \mathrm{E}\left[\sup_{p>0}\max_{\substack{1\le k\ne b\\ \ne e\le\hat{\rho}}}\left|\frac{l_k^p}{l_k^p-l_b^p}\right|^2\left|\frac{\hat{\gamma}_k^p-\hat{\gamma}_e^p}{l_k^p-l_e^p}-\frac{\hat{\gamma}_r^p-\hat{\gamma}_e^p}{l_b^p-l_e^p}\right|^2\right]$$
$$\text{and }\ \mathrm{E}\left[\sup_{p>0}\max_{1\le k\ne b\le\hat{\rho}<e\le p}\left|\frac{l_k^p}{l_k^p-l_b^p}\right|^2\left|\frac{\hat{\gamma}_k^p}{l_k^p-l_e^p}-\frac{\hat{\gamma}_b^p}{l_b^p-l_e^p}\right|^2\right]$$

are all finite for some $\epsilon > 0$. Similarly, a noise estimator $\hat{\sigma}^2$ satisfies the strong regularity conditions for a strong spiked eigenvalue estimator $\hat{\Gamma}$ if it satisfies the weak, and the following holds:

$$\mathrm{E}\left[\sup_{p>0}\max_{0\le k\le p}\left|\frac{\hat{\sigma}^{2p}}{l_k^p}\right|^{9(1+\epsilon)}\right],\ \mathrm{E}\left[\sup_{p>0}\max_{0\le k\le p}\left|\frac{\partial\hat{\sigma}^{2p}}{\partial l_k^p}\right|^{4.5}\right]$$
$$\text{and }\ \mathrm{E}\left[\sup_{p>0}\max_{1\le k\le p}\left|\hat{\gamma}_k^p+\hat{\sigma}^{2p}\right|\left|\frac{\partial^2\hat{\gamma}_k^p}{\partial l_k^{p2}}+\frac{\partial^2\hat{\sigma}^{2p}}{\partial l_k^{p2}}\right|\right]$$

are all finite for some $\epsilon > 0$.

Careful inspection of the proofs reveal that regularity conditions in this spirit are inevitable; however, we emphasize that by no means we believe those precise conditions to be necessary, merely sufficient. In any case, with these conditions in hand we can formally define the classes of estimators in which we look for an optimal $\hat{\sigma}^2$.

Definition 4. For $\hat{\Gamma}$ a weak (strong) spiked eigenvalue estimator, the associated

weak (strong) class of spiked corrections to the sample covariance matrix is

$$V_p(\hat{\Gamma}) = \left\{ \hat{\Gamma} + \hat{\sigma}^2 I_p \,\middle|\, \hat{\sigma}^2 \text{ is } \hat{\Gamma}\text{-weak} \right\} \text{ and } \bar{V}_p(\hat{\Gamma}) = \left\{ \hat{\Gamma} + \hat{\sigma}^2 I_p \,\middle|\, \hat{\sigma}^2 \text{ is } \hat{\Gamma}\text{-strong} \right\}.$$

We would like to find an optimal estimator over these two classes. Recall we are evaluating performance in Frobenius loss $L_F(\hat{\Sigma}, \Sigma) = \|\hat{\Sigma} - \Sigma\|_F^2/p$. Although natural and common within the literature, we find it more convenient to move to the closely related "invariant" loss

$$L_H(\hat{\Sigma}, \Sigma) = \frac{\operatorname{tr}\big[(\hat{\Sigma}\Sigma^{-1} - I)^2\big]}{p}.$$

This loss was, up to the high-dimensional p^{-1} normalization, mentioned by James and Stein [1961] early but first thoroughly investigated by Haff [1977], and we will refer to it as Haff's loss. A modification of the argument behind (4.1) shows that estimation of a spiked covariance matrix under this loss can also be thought as a noise estimation problem, just like for the Frobenius case. In this sense the problem stays similar.

A great advantage of the Haff loss is that it is one of the few for which an unbiased estimator of the risk is known, at least in the orthogonally invariant case. There is a rich body of literature behind that construction [Haff, 1977, 1979a, 1980], in different shapes and under different conditions. A remarkable feature is that if we collect and split the terms of the URE between the terms of leading and smaller order, the dominant part does not depend on the derivatives of the eigenvalue estimators. More precisely, we have this construction.

Theorem 6. *Let $n \geq p+1$. Then for any weak spiked estimator $\hat{\Sigma} \in V_p(\hat{\Gamma})$ whose spiked rank $\hat{\rho}$ is independent of S, we find its Haff risk to satisfy $E\big[L_H(\hat{\Sigma}, \Sigma)\big] = E\big[F + G\big]$ with $E\big[|F + G|\big] < \infty$, where*

$$F = F\Big(l, \hat{\rho}, \hat{\gamma}, \hat{\sigma}^2\Big) \quad \text{and} \quad G = G\Big(l, \hat{\rho}, \hat{\gamma}, \hat{\sigma}^2, \frac{\partial \hat{\gamma}}{\partial l}, \frac{\partial \hat{\sigma}^2}{\partial l}, \frac{\partial^2 \hat{\gamma}}{\partial l^2}, \frac{\partial^2 \hat{\sigma}^2}{\partial l^2}\Big)$$

are functionals that do not depend on Σ*. In addition, if the estimator is strong in the sense that* $\hat{\Sigma} \in \bar{V}_p(\hat{\Gamma})$*, then asymptotically* F *is the dominant term and* G *the dominated term, respectively, in the sense that*

$$\lim_{n\to\infty} \mathrm{E}\Big[|F|\Big] < \infty \quad \textit{and} \quad \lim_{n\to\infty} p\,\mathrm{E}\Big[|G|\Big] < \infty.$$

Explicit expressions for F *and* G *are given in (4.12)–(4.13).*

In practice, we found the decomposition most useful for a deterministic spiked rank $\hat{\rho} = r$, in which case we might consider estimates of the form $\hat{\Gamma}_r + \sigma_r^2 I$; this is the approach we take in Section 4.4 when constructing a specific estimator. But it is reasonable to think of a context in which some estimate of the true rank ρ based on prior independent data is available, in which case the construction applies equally.

Now fix a weak spiked eigenvalue estimator $\hat{\Gamma}$, and consider the task of finding a $\hat{\sigma}^2$ that minimizes the Haff risk: minimizing the construction from Theorem 6 makes the task plausible. Since G depends on the derivatives of $\hat{\sigma}^2$, formally proceeding with calculus of variations would yield a minimizer that depends on the unknown density of the eigenvalues which is, of course, unavailable. However since the dominant part only depends on $\hat{\sigma}^2$ itself, one can obtain a minimizer of $\mathrm{E}[F]$ whose expression is independent of Σ.

Proposition 8. *Let* $n \geq p$*. If* $\hat{\Sigma} \mapsto \mathrm{E}[F]$ *has a minimum over* $V_p(\hat{\Gamma})$*, it is given by* $\tilde{\Sigma} = \hat{\Gamma} + \tilde{\sigma}^2 I$*, where* $\tilde{\sigma}^2 = A/B$ *where*

$$\begin{aligned} A = {} & \frac{n-p-1}{np} \sum_{c=1}^{p} \frac{1}{l_c} - \frac{(n-p-1)(n-p-2)}{n^2 p} \sum_{k=1}^{\hat{\rho}} \frac{\hat{\gamma}_k}{l_k^2} \\ & + \frac{n-p-1}{n^2 p} \sum_{k=1}^{\hat{\rho}} \sum_{c=1}^{p} \frac{\hat{\gamma}_k}{l_k} \frac{1}{l_c} - 2\frac{n-p-1}{n^2 p} \sum_{k=1}^{\hat{\rho}} \sum_{c=\hat{\rho}+1}^{p} \frac{1}{l_c} \frac{\hat{\gamma}_k}{l_k - l_c} \\ & + \frac{3}{n^2 p} \sum_{k\neq b}^{\hat{\rho}} \sum_{c=\hat{\rho}+1}^{p} \frac{1}{l_k - l_c} \frac{\hat{\gamma}_b}{l_b - l_c} - \frac{3}{n^2 p} \sum_{k=1}^{\hat{\rho}} \sum_{c\neq d=\hat{\rho}+1}^{p} \frac{1}{l_k - l_c} \frac{\hat{\gamma}_k}{l_k - l_d} \end{aligned}$$

$$- \frac{3}{n^2 p} \sum_{k \neq b}^{\hat{\rho}} \sum_{c=\hat{\rho}+1}^{p} \frac{\hat{\gamma}_k - \hat{\gamma}_b}{l_k - l_b} \frac{1}{l_k - l_c},$$

$$B = \frac{(n-p-1)(n-p-2)}{n^2 p} \sum_{c=1}^{p} \frac{1}{l_c^2} - \frac{n-p-1}{n^2 p} \sum_{c=1}^{p} \frac{1}{l_c} \sum_{c=1}^{p} \frac{1}{l_c}.$$

In addition, if $n \geq 2p+2$, a minimum must exist (and therefore at $\tilde{\Sigma}$).

The matter of whether a minimizer should exist at all in a given context is delicate. A proof of existence for some large class of covariance matrices would be quite interesting. In the spiked case, the remarks following Lemma 4 hint at a plausible approach.

4.3 Properties

The previous chapter was concerned with the construction of a good estimator $\tilde{\sigma}^2$ that satisfies some optimality property, namely minimizing the dominant part of the Haff URE over $V_p(\hat{\Gamma})$. Let us now turn our attention to its performance in estimating σ^2 under spikedness. We will make repeated use of the following lemma, which extends the results of Nadler [2008].

Lemma 4. *Suppose the underlying sequence of covariance matrices $\{\Sigma_p\}$ is spiked and $p_n/n \to c \in (0,1)$.*

(i) If $\gamma_\rho/\sigma^2 > \sqrt{c}$, then for any $1 \leq k \leq \rho$,

$$\frac{1}{p-\rho} \sum_{d=\rho+1}^{p} \frac{l_d}{l_k - l_d} \xrightarrow[n\to\infty]{a.s.} \frac{\sigma^2}{\gamma_k};$$

(ii) For any $m > 1$,

$$\frac{1}{p-\rho} \sum_{d=\rho+1}^{p} \frac{1}{l_d^m} \xrightarrow[n\to\infty]{a.s.} \frac{1}{(1-c)^{2m-1}} \frac{1}{\sigma^{2m}}.$$

The supercriticality assumption $\gamma_\rho/\sigma^2 > \sqrt{c}$ in (i) is necessary for the expression to converge. Two remarks are in order. First, as a consequence of this result, it is easy to show that

$$\frac{1}{p-\rho}\sum_{d=\rho+1}^{p}\frac{1}{l_k-l_d} \xrightarrow[n\to\infty]{\text{a.s.}} \frac{1}{\gamma_k+c\sigma^2},$$
$$\frac{1}{p-\rho}\sum_{d=\rho+1}^{p}\frac{1}{l_d(l_k-l_d)} \xrightarrow[n\to\infty]{\text{a.s.}} \frac{1}{1-c}\frac{\gamma_k}{\sigma^2(\gamma_k+c\sigma^2)^2},$$

a result we will use later in Section 4.4. Second, in connection with the proof of Proposition 8, we see that when the underlying sequence of covariance matrices $\{\Sigma_p\}$ is spiked

$$\frac{1}{n^2}\left[(n-p-2)\sum_{d=1}^{p}\frac{1}{l_d^2}-\left(\sum_{d=1}^{p}\frac{1}{l_d}\right)^2\right] \xrightarrow[n\to\infty]{\text{a.s.}} \frac{c}{(1-c)\sigma^4} > 0,$$

by Lemma 4. From (4.15), one would therefore expect the estimator also to be a minimizer under spikedness. Although we haven't been successful in formalizing this intuition, this could be a plausible approach towards proving existence of minimizers for spiked covariance matrices.

Let us now turn our attention to the behavior of $\tilde{\sigma}^2$. The following theorem summarizes important aspects of its asymptotic behavior.

Theorem 7. *Suppose the underlying sequence of covariance matrices $\{\Sigma_p\}$ is spiked and $p_n/n \to c \in (0,1)$ with $\gamma_\rho/\sigma^2 > \sqrt{c}$. For a given weak $\hat{\Gamma}$, let $\tilde{\sigma}^2$ be the associated minimizer of Proposition 8. Then*

(i) If $\hat{\rho}$ a.s. converges to a finite constant, then $\tilde{\sigma}^2 \xrightarrow[n\to\infty]{a.s.} \sigma^2$.

(ii) If $\hat{\rho}$ is strongly consistent and for all $1 \le k \le \rho$, $\hat{\gamma}_k$ a.s. converges to some finite constant, then we have bounds $X_n^- \le n(\tilde{\sigma}^2-\sigma^2) \le X_n^+$ with

$$X_n^- \xrightarrow[n\to\infty]{\mathcal{D}} N\left(\mu^-, \frac{2c(1+c)^2}{(1-c)^4}\sigma^4\right), \quad X_n^+ \xrightarrow[n\to\infty]{\mathcal{D}} N\left(\mu^+, \frac{2c(1+c)^2}{(1-c)^4}\sigma^4\right),$$

where μ^- and μ^+ have explicit expressions given in (4.27)–(4.28).

An immediate consequence of this result is that $\tilde{\sigma}^2$ estimates σ^2 with rate n in our absolute error loss $L(\hat{\sigma}^2, \sigma^2) = |\hat{\sigma}^2 - \sigma^2|$. We should mention that, given good estimators of ρ and γ_k, one could perhaps build approximate high-dimensional confidence intervals for σ^2 with part (ii) of Theorem 7. We will not investigate this further, but rather turn our attention to minimax rates for the noise estimation problem. For any spiked sequence $\{\Sigma_p\}$, define a δ-ball of order r as

$$\mathrm{B}_r(\Sigma, \delta) = \left\{ \{\Sigma_p'\} \text{ spiked } \;\middle|\; \left|\lambda_i(\Sigma_p) - \lambda_i(\Sigma_p')\right| < \delta \frac{\lambda_i(\Sigma_p)}{n^r} \quad \forall p > 0 \right\}.$$

We start with a lemma. Recall that $\mathrm{d_{TV}}$ stands for the total variation distance between two probability measures.

Lemma 5. *Let $\{\Sigma_p\}$ be a spiked sequence of covariance matrices and $M > 0$. Then, as $p_n/n \to c \in (0,1)$,*

$$\lim_{n\to\infty} \sup_{\Sigma' \in B_r(\Sigma, 2M)} d_{TV}\Big(N(0,\Sigma_p)^n, N(0,\Sigma_p')^n \Big) \leq \sqrt{1 - \exp\left(-\frac{cM^2}{2}\right)}$$

when $r = 1$, while this limit is zero when $r > 1$.

Using this result, we now proceed to show a lower bound on the local minimax rate of convergence for estimating σ^2, using the classic two-point test argument of Le Cam [1973].

Theorem 8. *Say the underlying sequence of covariance matrices $\{\Sigma_p\}$ is spiked and $p_n/n \to c \in (0,1)$. Let $\epsilon > 0$ and define*

$$M_\epsilon = \sqrt{-\frac{2}{c}\log\Big(1 - (1-4\epsilon)^2\Big)}.$$

Then no estimator can estimate σ^2 with speed $\sigma^2 M_\epsilon/n$ over the shrinking neighborhoods $B(\Sigma_p, 2M_\epsilon)$, in the sense that

$$\liminf_{n\to\infty} \inf_{\hat{\sigma}^2} \sup_{\Sigma' \in B_1(\Sigma, 2M_\epsilon)} P_{\Sigma_p'}\left[|\hat{\sigma}^2 - \sigma^{2\prime}| > \sigma^2 \frac{M_\epsilon}{n} \right] \geq \epsilon.$$

Thus, the minimax rate of estimation of σ^2 over n-shrinking neighborhoods cannot be faster than $O_P(1/n)$ (so in particular over, say, fixed neighborhoods.) Using Theorem 7, we can show our noise estimator $\tilde{\sigma}^2$ essentially achieves this rate, in the sense that it is $o_P(1/n^r)$ over n^r-shrinking neighborhoods for any $r > 1$.

Proposition 9. *Let the underlying sequence of covariance matrices $\{\Sigma_p\}$ be spiked and $p_n/n \to c \in (0,1)$ with $\gamma_\rho/\sigma^2 > \sqrt{c}$. For a given weak $\hat{\Gamma}$, let $\tilde{\sigma}^2$ be the associated extremizer of proposition 8. If $\hat{\rho}$ is strongly consistent and for all $1 \leq k \leq \rho$, $\hat{\gamma}_k$ a.s. converges to some finite constant, then for any $r > 1$ and $M > 0$, $\tilde{\sigma}^2$ estimates σ^2 with rate at least $\sigma^2 M/n^r$ over the shrinking neighborhoods $B_r(\Sigma, 2M)$, in the sense that*

$$\lim_{n\to\infty} \sup_{\Sigma' \in B_r(\Sigma, 2M)} P_{\Sigma'_p}\left[|\tilde{\sigma}^2 - \sigma^{2\prime}| > \sigma^2 \frac{M}{n^r}\right] = 0.$$

Thus we can conclude that, despite choosing our noise estimator to minimize a covariance problem, good behavior has been transferred to the noise estimation problem, which is not surprising in light of (4.1). In particular, we see that by Theorem 7 (i), strong consistency of the noise estimator follows even when the eigenvalues of $\hat{\Gamma}$ itself are not consistent – a robustness which is certainly welcome.

4.4 Application

Having built and analyzed our noise estimator, we now proceed to illustrate our construction by building a specific covariance estimator. We hope this concrete example will help clarify the approach taken and its behavior in the covariance problem.

4.4.1 Example

We build a spiked covariance estimator as follows. The first step is to specify an asymptotically negligible spiked component $\hat{\Gamma}$. For r some fixed rank strictly smaller than p, we take $\tilde{\Gamma}_r = O\text{diag}(\tilde{\gamma})\, O'$ with

$$\tilde{\gamma}_k = \sum_{c=r+1}^{p} l_c \left(\sum_{c=r+1}^{p} \frac{l_c}{l_k - l_c} \right)^{-1}$$

for $1 \leq k \leq r$, and 0 otherwise. These estimators are strongly consistent when $r = \rho$, as we will soon show using Lemma 4; this is the main motivation for our choice. Note that this choice does not quite fit within the framework considered by Donoho et al. [2014], since it is not a function of l_k only. With this choice of spiked part, let $\tilde{\sigma}_r^2$ be the minimizer of Proposition 8 associated with our spiked component. We then have a family $r \to \tilde{\Sigma}_r = \tilde{\Gamma}_r + \tilde{\sigma}_r^2 I$ for all $0 \leq r < p$, which we naturally extend to the $r = p$ case through $\tilde{\Sigma}_p = \tilde{\Gamma}_p = S$.

Next, we select the rank r based on the data. Motivated again by the results of Lemma 4, we define the rank estimator

$$\tilde{\rho} = \underset{0 \leq r \leq p}{\arg\min} \left\{ \frac{\mathbb{1}[r < p]}{l_{r+1}} \frac{(1 + \sqrt{p/n})^2}{p - r} \sum_{c=r+1}^{p} l_c \geq 1, \quad \left| F_r + G_r \right| \leq \frac{p+1}{n} \right\},$$

with F_r, G_r the F, G of Theorem 6 applied to $\tilde{\Gamma}_r$ and $\tilde{\sigma}_r^2$. This choice aims to select the smallest rank that lies both above the critical threshold and yields improvement in Haff risk over S. Since $r = p$ satisfies both criteria, the set is never empty and in the worst case we simply do not correct the eigenvalues of S. This will happen when there is strong departure from spikedness, which means that the construction is in some sense robust to this situation: it exploits it when present and reverts to S when not.

Finally, we simply set $\tilde{\Sigma} = \tilde{\Sigma}_{\tilde{\rho}}$ as our estimator. In practice, the computation is straightforward, since everything is in closed-form, with polynomial complexity.

An implementation in R is available at `http://stat.cornell.edu/~chetelat`. At the same time, as previously hinted, the estimator has strongly consistent eigenvalues under spikedness. The proof is a simple application of results from Section 4.3.

Proposition 10. *If the underlying sequence of covariance matrices* $\{\Sigma_p\}$ *is spiked and* $p_n/n \to c \in (0,1)$ *with* $\gamma_\rho/\sigma^2 > \sqrt{c}$*, then* $\tilde{\rho}$*,* $\tilde{\sigma}^2$ *and* $\tilde{\gamma}_k$ *for* $1 \leq k \leq \rho$ *are all strongly consistent.*

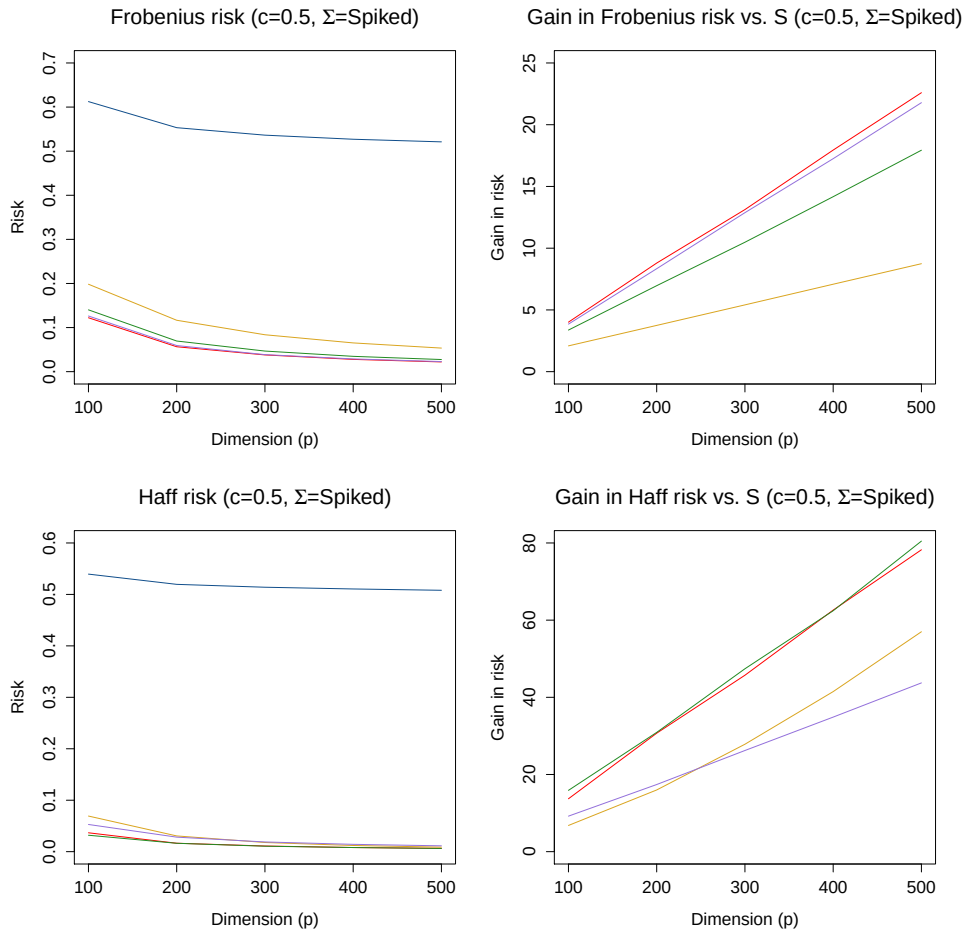
This stands, of course, in contrast with the eigenvalues of the sample covariance matrix S, which converge to the wrong values (4.2) in spiked settings.

4.4.2 Numerical comparisons

We display the performance of the constructed estimator through simulations. The setting is as follows. We fix the dimension to sample ratio c at 0.5 and vary n, p. For each n, p, we simulate data from a normal $\mathrm{N}(0, \Sigma)$ and approximate its Haff and Frobenius risk using a law of large numbers approximation with 100 iterations. Four true covariance matrices Σ are considered. The first is a spiked setting $\Sigma = \mathrm{diag}(5, 4, 3, 2, 1, ..., 1)$, while the other three correspond to autoregressive settings $\Sigma_{ij} = \kappa^{|i-j|}$ for $\kappa = 0.05$, 0.5 and 0.95. The case $\kappa = 0.95$ is particularly difficult for the constructed estimator as it is very far from spikedness.

The risks are computed for S, our estimator $\tilde{\Sigma}$ and three benchmark competitors. The first is Stein's isotonized covariance estimator, with well regarded overall performance. We follow the implementation of Lin and Perlman [1985]. The second is the popular linear shrinkage covariance estimator of Ledoit and Wolf [2004], specifically designed for high-dimensional settings. The third is a naive spiked estimator, given by the same estimators of spikes as $\tilde{\Sigma}$, the noise estimator

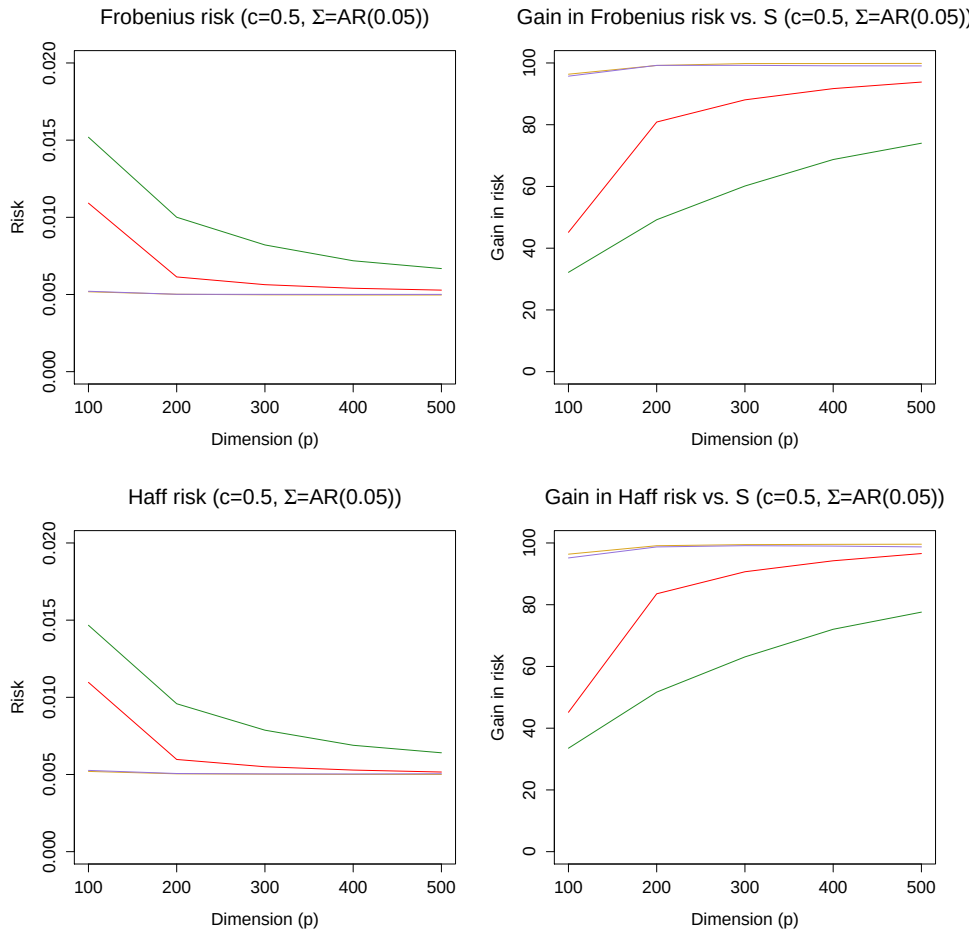
Figure 4.1: Spiked covariance setting.



$\hat{\sigma}^2 = \frac{1}{n-r} \sum_{c=\hat{\rho}+1}^{p} l_c$ and $\hat{\rho}$ chosen by cross-validation on the Frobenius loss. We plot the risks and the gain in risk with respect to S, defined as $\text{Risk}(S)/\text{Risk}(\hat{\Sigma}) - 1$. The computations were performed using the R package, and the results are given as Figures 4.1-4.4. Blue corresponds to S, red to $\tilde{\Sigma}$, green to Stein's isotonized estimator, yellow to the Ledoit-Wolf estimator and purple to the naive spiked estimator.

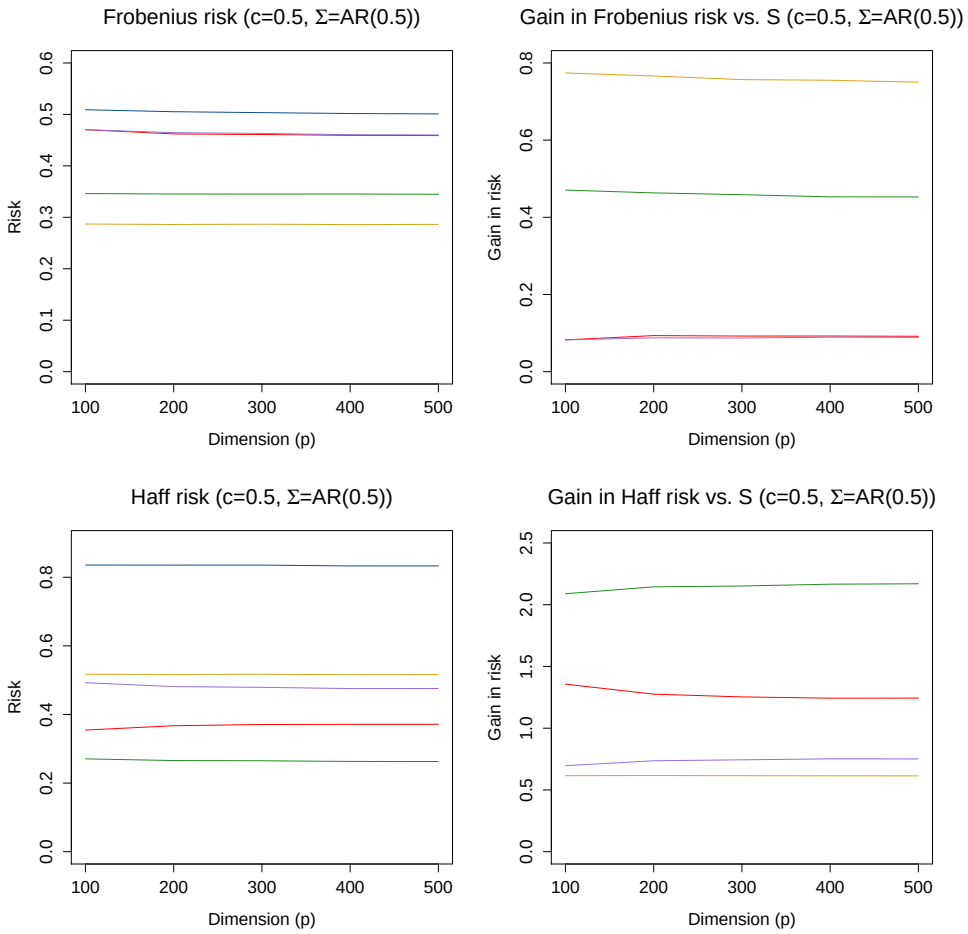
We can see from the results of the simulations that the expected good performance in Haff loss of our estimator seems to translate well into the more standard Frobenius loss. The estimator performs particularly well in supercritical spiked set-

Figure 4.2: AR(0.05) setting. The sample covariance matrix is omitted.



tings, with a 23-times improvement over S in Frobenius risk for $p = 500$, $n = 1000$ in our setting. In particular, in this setting it outperforms the naive spiked estimator. In addition, the estimator is quite robust to deviations from spikedness, as even in worst-case scenarios such as an AR(0.95) setting, we do not do worse than S in Haff or Frobenius risk. There is therefore little to lose by using it rather than the sample covariance matrix, and as far as such a thing can exist, it could be advocated as some kind of generic high-dimensional covariance estimator.

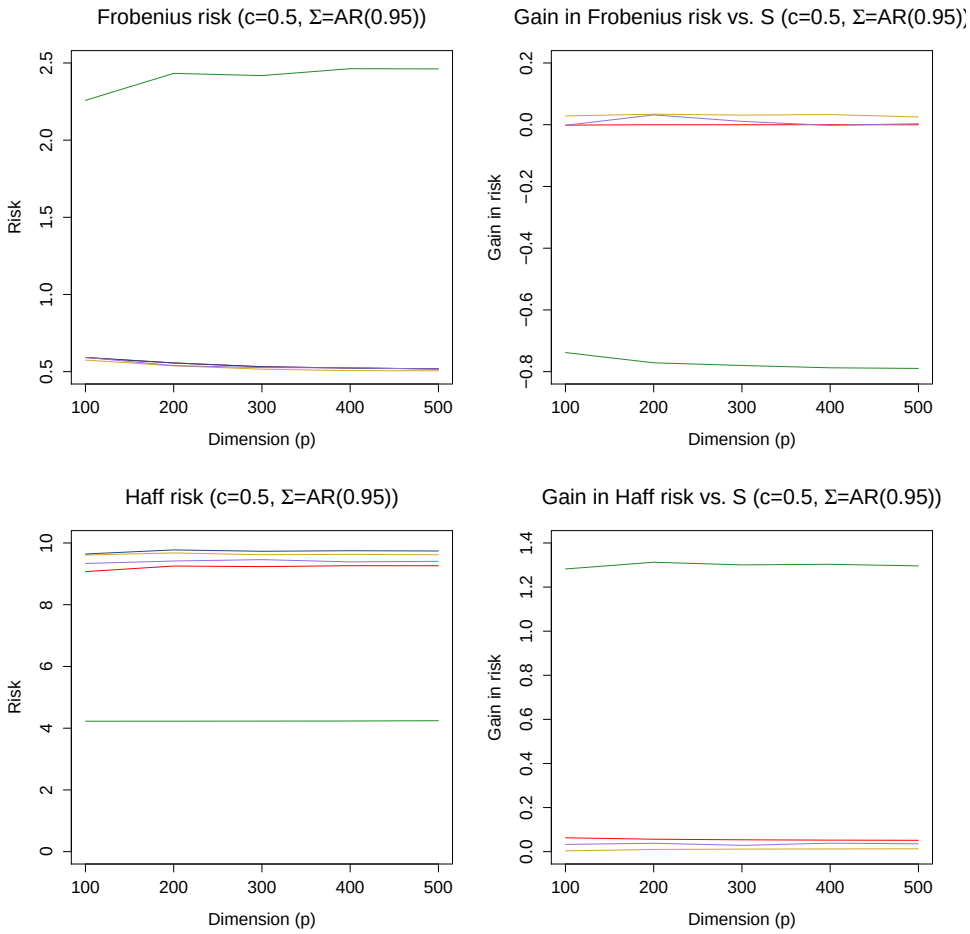
Figure 4.3: AR(0.50) setting.



4.5 Comments

In this work we considered two parallel high-dimensional problems, the estimation of noise in principal components analysis under absolute error loss and the estimation of a spiked covariance matrix under Frobenius loss. We proposed a variational solution, by restricting ourselves to regular estimators and minimizing an unbiased covariance risk estimator in the invariant analogue of the loss. The resulting noise estimator was shown to be strongly consistent and almost asymptotically normal and minimax for the noise problem, and we used the construction to build a robust spiked covariance estimator with good simulation performance. Beyond this,

Figure 4.4: AR(0.95) setting.



however, there are several aspects of our solution that warrant further discussion.

First, we assumed throughout this work that the underlying data was normal, and the construction and proofs depend quite heavily on it. This could be a point of discord between practice and the theory outlined here. However, we feel that, unlike many statistical problems where normality is convenient but unrealistic, it is quite natural here. Indeed, the construction and its properties only depends on the data through the eigenstructure of S, unlike other estimators such as the one of Ledoit and Wolf [2004]. The sample covariance matrix being an empirical average, one can expect it to behave asymptotically like a Wishart, and in that sense the assumption does not appear particularly restrictive.

Another assumption running through the work is that although we are in high-dimensions, we keep $p \leq n$. The extension to a $p > n$ setting is attractive, as in addition to the properties described above, a corresponding robust spiked covariance estimator would be automatically invertible, in contrast with S. The single obstacle appears to be the absence of an appropriate unbiased risk estimator for the Haff loss when $p > n$. This is therefore more an obstruction by knowledge than mathematics, as if such a construction would be found, the method outlined in this work could easily be applied.

As we considered minimization of a covariance loss, it might be surprising that we did not present any results on the behavior of the estimator in the covariance problem. We strongly believe that the Haff risk must tend to zero under spikedness since, as some algebra shows, the unbiased risk estimator of our estimate tends a.s. to zero. This is quite interesting since the Haff risk of S equals, in contrast, $(p+1)/n \to c > 0$. Unfortunately, we haven't been able to prove this statement. Although the literature on the probabilistic behavior of Wishart eigenvalues is extensive, it is more scant on their L^1 behavior, and this limits what can be proven as of now.

In Section 4.2, we considered an invariant analogue of the Frobenius loss $\|\hat{\Sigma} - \Sigma\|_F^2/p$, the Haff loss $\|\hat{\Sigma}\Sigma^{-1} - I\|_F^2/p$ which allowed for the existence of an unbiased risk estimator. Since our estimator is particularly adapted to this invariant covariance loss, it might also be of interest to study an invariant noise loss such as $|\hat{\sigma}^2/\sigma^2 - 1|$.

We did not tackle the problem of selecting the spiked eigenvalue estimators $\hat{\gamma}_k$ in an optimal way, beyond the suggestion in Section 4.4. The recent work of Donoho et al. [2014] could offer a solution. The authors consider the spiked covariance estimation problem where the noise is known and fixed at $\sigma^2 = 1$, and

look for optimal shrinking of the spiked eigenvalues l_k, $1 \leq k \leq \rho$. In the Frobenius and Haff losses, their optimal estimators coincide and equal

$$\hat{\gamma}_k = \left[l_k - 1 + \frac{c l_k}{l_k - 1} \right] \frac{1 - c/(l_k - 1)^2}{1 + c/(l_k - 1)}$$

for $l_k > (1 + \sqrt{c})^2$. An appealing feature of this estimate is that it accounts for the deterministic angles between the top sample and population eigenvectors. It would be interesting to study the behavior of the noise estimator $\tilde{\sigma}^2$ from Theorem 8 applied to these spiked estimators, with perhaps adjustments for not knowing σ^2.

Finally, we should remark that our construction automatically provides well-conditioned covariance estimators, which is quite important for applications. Therefore, when the parameter of interest is the precision rather than the covariance matrix, using $\tilde{\Sigma}^{-1}$ as estimator appears reasonable, although we currently do not have any formal results on its behavior for this problem.

4.6 Technical results and proofs

4.6.1 Proofs for Section 4.2

The following Stein-Haff identity is used to compute an unbiased estimator of risk for orthogonally invariant estimators in proposition 6. The general identity dates back to Haff [1979a] and its specialization to orthogonally invariant estimators for $n \geq p$ first implicitly used by Sheena [1995]. Unfortunately, the approach taken by the author requires regularity conditions that are difficult to verify in practice (conditions 1–3 in his Section 1 and 2). The following lemma follows the approach of Konno [2009] and Kubokawa and Srivastava [2008] to obtain the same $n \geq p$ result, but under weaker, simpler conditions.

Lemma 6. *Let $W \sim W_p(n, \Sigma)$ with $n \geq p$, and let $W/n = OLO'$ be the spectral decomposition of the associated sample covariance matrix. Let $\psi_1(L), ..., \psi_p(L)$ be differentiable functions of the eigenvalues of W/n satisfying:*

$$\mathrm{E}\left[\left|\sum_{k=1}^{p} \frac{n-p-1}{n}\frac{\psi_k}{l_k} + \frac{2}{n}\sum_{k=1}^{p}\frac{\partial \psi_k}{\partial l_k} + \frac{1}{n}\sum_{k\neq b}^{p}\frac{\psi_k - \psi_b}{l_k - l_b}\right|\right] < \infty. \tag{4.3}$$

Define $\Psi = diag(\psi_1, ..., \psi_p)$. Then

$$\mathrm{E}\big[\mathrm{tr}\big(\Sigma^{-1}O\Psi O'\big)\big] = \mathrm{E}\left[\sum_{k=1}^{p} \frac{n-p-1}{n}\frac{\psi_k}{l_k} + \frac{2}{n}\sum_{k=1}^{p}\frac{\partial \psi_k}{\partial l_k} + \frac{1}{n}\sum_{k\neq b}^{p}\frac{\psi_k - \psi_b}{l_k - l_b}\right].$$

Proof. We use Lemma 3 in Chételat and Wells [2012]. Decompose $W = X'X$ for some $X \sim N_{n\times p}(0, I_n \otimes \Sigma)$. In the spirit of Lemma 4.1 in Konno [2009], we find

$$\begin{aligned}
(dX')X + X'dX &= d(X'X) = n(dO)LO' + nOdLO' + nOLdO' \\
\Rightarrow \qquad O'\left[(dX')X + X'dX\right]O &= nO'dOL + nL(dO')O + ndL.
\end{aligned}$$

Since $O'dO + (dO')O = 0$, we get

$$O'\left[(dX')X + X'dX\right]O = nO'dOL - nLO'dO + ndL,$$

so that for $k \neq l$,

$$\begin{aligned}
O'dO_{kl} &= \frac{1}{n}\frac{1}{l_l - l_k}O'\left[(dX')X + X'dX\right]O_{kl}, \\
dL_{kk} &= \frac{1}{n}O'\left[(dX')X + X'dX\right]O_{kk}
\end{aligned}$$

and $O'dO_{kk} = 0$. Then

$$\begin{aligned}
\frac{\partial l_k}{\partial X_{ij}} &= \frac{1}{n}\sum_{\alpha,\beta,\gamma} O'_{k\alpha}\frac{X'_{\alpha\beta}}{dX_{ij}}X_{\beta\gamma}O_{\gamma k} + \frac{1}{n}\sum_{\alpha,\beta,\gamma} O'_{k\alpha}X'_{\alpha\beta}\frac{X_{\beta\gamma}}{dX_{ij}}O_{\gamma k} \\
&= \frac{2}{n}\sum_{\gamma} O'_{kj}X_{i\gamma}O_{\gamma k}
\end{aligned} \tag{4.4}$$

and

$$\begin{aligned}
\frac{\partial O_{kl}}{\partial X_{ij}} &= \frac{1}{n}\sum_{\alpha\neq l,\beta,\gamma,\epsilon} O_{k\alpha}\frac{1}{l_l - l_\alpha}O'_{\alpha\beta}\left[\frac{\partial X'_{\beta\gamma}}{\partial X_{ij}}X_{\gamma\epsilon} + X'_{\beta\gamma}\frac{\partial X'_{\gamma\epsilon}}{\partial X_{ij}}\right]O_{\epsilon l} \\
&= \frac{1}{n}\sum_{\alpha\neq l,\beta} O_{k\alpha}\frac{O'_{\alpha j}O_{\beta l} + O'_{\alpha\beta}O_{jl}}{l_l - l_\alpha}X_{i\beta}.
\end{aligned} \tag{4.5}$$

Now define $\tilde{X} = X\Sigma^{-1/2}$ and $H = \frac{1}{n}\Sigma^{1/2}OL^{-1}\Psi O'\Sigma^{-1/2}$ – we need to compute $\mathrm{div}_{\mathrm{vec}(\tilde{X})}\ \mathrm{vec}\left(\tilde{X}H\right)$. We find

$$\begin{aligned}
\mathrm{div}_{\mathrm{vec}(\tilde{X})}\ \mathrm{vec}\left(\tilde{X}H\right) &= \sum_{\alpha,i,j}\frac{\partial}{\partial\tilde{X}_{\alpha i}}\left\{\tilde{X}_{\alpha j}H_{ji}\right\} = n\sum_{i}H_{ii} + \sum_{\alpha,j}\tilde{X}_{\alpha j}\frac{\partial H_{ji}}{\partial\tilde{X}_{\alpha i}} \\
&= \sum_{\gamma}\frac{\psi_\gamma}{l_\gamma} + \frac{1}{n}\sum_{\alpha,\beta,i,j,k,l}\tilde{X}_{\alpha j}\Sigma^{1/2}_{\beta i}\Sigma^{1/2}_{jk}\frac{\partial}{\partial X_{\alpha\beta}}\left\{OL^{-1}\Psi O'_{kl}\right\}\Sigma^{-1/2}_{li} \\
&= \sum_{\gamma}\frac{\psi_\gamma}{l_\gamma} + \frac{1}{n}\sum_{\alpha,k,l}X_{\alpha k}\frac{\partial}{\partial X_{\alpha l}}\left\{OL^{-1}\Psi O'_{kl}\right\} \qquad (4.6) \\
&= \sum_{\gamma}\frac{\psi_\gamma}{l_\gamma} + \frac{1}{n}\sum_{\alpha,k,l,\beta}X_{\alpha k}\frac{\partial O_{k\beta}}{\partial X_{\alpha l}}\left[L^{-1}\Psi\right]_{\beta\beta}O'_{\beta l} \\
&\quad + \frac{1}{n}\sum_{\alpha,k,l,\beta}X_{\alpha k}O_{k\beta}\frac{\partial\left[L^{-1}\Psi\right]_{\beta\beta}}{\partial X_{\alpha l}}O'_{\beta l} + \frac{1}{n}\sum_{\alpha,k,l,\beta}X_{\alpha k}O_{k\beta}\left[L^{-1}\Psi\right]_{\beta\beta}\frac{\partial O'_{\beta l}}{\partial X_{\alpha l}}.
\end{aligned}$$

Using (4.4) and (4.5), we obtain

$$\begin{aligned}
&= \sum_{\gamma}\frac{\psi_\gamma}{l_\gamma} + \frac{1}{n^2}\sum_{\alpha,k,l,\beta,\gamma\neq\beta,\epsilon}X_{\alpha k}O_{k\gamma}\frac{O'_{\gamma l}O_{\epsilon\beta} + O'_{\gamma\epsilon}O_{l\beta}}{l_\beta - l_\gamma}X_{\alpha\epsilon}\left[L^{-1}\Psi\right]_{\beta\beta}O'_{\beta l} \\
&\quad + \frac{2}{n^2}\sum_{\alpha,k,l,\beta,\gamma,\epsilon}X_{\alpha k}O_{k\beta}O'_{\gamma l}X_{\alpha\epsilon}O_{\epsilon\gamma}\frac{\partial\left[\psi_\beta/l_\beta\right]}{\partial l_\gamma}O'_{\beta l} \\
&\quad + \frac{1}{n^2}\sum_{\alpha,k,l,\beta,\gamma\neq\beta,\epsilon}X_{\alpha k}O_{k\beta}\left[L^{-1}\Psi\right]_{\beta\beta}O_{l\gamma}\frac{O'_{\gamma l}O_{\epsilon\beta} + O'_{\gamma\epsilon}O_{l\beta}}{l_\beta - l_\gamma}X_{\alpha\epsilon} \\
&= \sum_{\gamma}\frac{\psi_\gamma}{l_\gamma} + \frac{1}{n}\sum_{\gamma\neq\beta}\frac{l_\gamma\psi_\beta}{(l_\beta - l_\gamma)l_\beta} + \frac{2}{n}\sum_{\gamma}l_\gamma\frac{\partial\left[\psi_\gamma/l_\gamma\right]}{\partial l_\gamma} + \frac{1}{n}\sum_{\gamma\neq\beta}\frac{\psi_\beta}{l_\beta - l_\gamma} \\
&= \frac{n-p-1}{n}\sum_{\gamma}\frac{\psi_\gamma}{l_\gamma} + \frac{2}{n}\sum_{\gamma}\frac{\partial\psi_\gamma}{\partial l_\gamma} + \frac{1}{n}\sum_{\gamma\neq\beta}\frac{\psi_\beta - \psi_\gamma}{l_\beta - l_\gamma}.
\end{aligned}$$

By (4.3), we conclude

$$\begin{aligned}
&\mathrm{E}\left[\left|\mathrm{div}_{\mathrm{vec}(\tilde{X})}\ \mathrm{vec}\left(\tilde{X}H\right)\right|\right] \\
&\quad = \mathrm{E}\left[\left|\sum_{k=1}^{p}\frac{n-p-1}{n}\frac{\psi_k}{l_k} + \frac{2}{n}\sum_{k=1}^{p}\frac{\partial\psi_k}{\partial l_k} + \frac{1}{n}\sum_{k\neq b}^{p}\frac{\psi_k - \psi_b}{l_k - l_b}\right|\right] < \infty.
\end{aligned}$$

Therefore, we can apply Lemma 3 in Chételat and Wells [2012] to $G = \frac{1}{n}OL^{-1}\Psi O'$,

which holds for any (p, n). We obtain that

$$\begin{aligned}\mathrm{E}\big[\mathrm{tr}\big(\Sigma^{-1}O\Psi O'\big)\big] &= \mathrm{E}\big[\mathrm{tr}\big(L^{-1}\Psi\big) + \mathrm{tr}(X'\nabla_X G')\big] \\ &= \mathrm{E}\left[\sum_\gamma \frac{\psi_\gamma}{l_\gamma} + \frac{1}{n}\sum_{k,l,\alpha} X'_{k\alpha}\frac{\partial}{\partial X_{\alpha l}} OL^{-1}\Psi O'_{kl}\right].\end{aligned}$$

But the expression inside the expected value is precisely eq. (4.6), so

$$= \mathrm{E}\left[\sum_{k=1}^p \frac{n-p-1}{n}\frac{\psi_k}{l_k} + \frac{2}{n}\sum_{k=1}^p \frac{\partial \psi_k}{\partial l_k} + \frac{1}{n}\sum_{k\neq b}^p \frac{\psi_k - \psi_b}{l_k - l_b}\right]$$

as desired. □

Lemma 7. *Let $W \sim W_p(n, \Sigma)$ with $n \geq p$, and let $W/n = OLO'$ be the spectral decomposition of the associated sample covariance matrix. Let $\psi_1(L), ..., \psi_p(L)$ be twice-differentiable functions of the eigenvalues of W/n, and define the associated quantities*

$$\psi_k^* = \frac{n-p-1}{n}\frac{\psi_k^2}{l_k} + \frac{4}{n}\psi_k\frac{\partial \psi_k}{\partial l_k} + \frac{2}{n}\psi_k\sum_{b\neq k}^p \frac{\psi_k - \psi_b}{l_k - l_b} \qquad \text{for } k = 1, ..., p$$

with $\Psi^* = diag(\psi_1, ..., \psi_p)$. Assume

$$\mathrm{E}\left[\sum_{k=1}^p \left|\frac{\psi_k^*}{l_k}\right|^{1+\epsilon}\right] < \infty \qquad \text{and}$$

$$\mathrm{E}\left[\left|\sum_{k=1}^p \frac{n-p-1}{n}\frac{\psi_k^*}{l_k} + \frac{2}{n}\sum_{k=1}^p \frac{\partial \psi_k^*}{\partial l_k} + \frac{1}{n}\sum_{k\neq b}^p \frac{\psi_k^* - \psi_b^*}{l_k - l_b}\right|\right] < \infty.$$

for some $\epsilon > 0$. Then

$$\mathrm{E}\left[\mathrm{tr}\Big(\big[\Sigma^{-1}O\Psi O'\big]^2\Big)\right] = \mathrm{E}\left[\sum_{k=1}^p \frac{n-p-1}{n}\frac{\psi_k^*}{l_k} + \frac{2}{n}\sum_{k=1}^p \frac{\partial \psi_k^*}{\partial l_k} + \frac{1}{n}\sum_{k\neq b}^p \frac{\psi_k^* - \psi_b^*}{l_k - l_b}\right].$$

Proof. We use Lemma 3 in Chételat and Wells [2012] again. Decompose $W = X'X$ for some $X \sim N_{n\times p}(0, I_n \otimes \Sigma)$, and define $\tilde{X} = X\Sigma^{-1/2}$ and

$H = \frac{1}{n}\Sigma^{1/2}OL^{-1}\Psi O'\Sigma^{-1}O\Psi O'\Sigma^{-1/2}$. Then

$$\begin{aligned}
\mathrm{div}_{\mathrm{vec}(\tilde{X})}\ \mathrm{vec}\left(\tilde{X}H\right) &= \sum_{\alpha,i,j}\frac{\partial}{\partial\tilde{X}_{\alpha i}}\left\{\tilde{X}_{\alpha j}H_{ji}\right\} = n\sum_i H_{ii} + \sum_{\alpha,j}\tilde{X}_{\alpha j}\frac{\partial H_{ji}}{\partial\tilde{X}_{\alpha i}} \\
&= \sum_{i,j,\gamma}\Sigma^{-1}_{ij}O_{j\gamma}\frac{\psi^2_\gamma}{l_\gamma}O'_{\gamma i} \qquad (4.7) \\
&\quad + \frac{1}{n}\sum_{\alpha,\beta,i,j,k,l}\tilde{X}_{\alpha j}\Sigma^{1/2}_{\beta i}\Sigma^{1/2}_{jk}\frac{\partial}{\partial X_{\alpha\beta}}\left\{OL^{-1}\Psi O'\Sigma^{-1}O\Psi O'\right\}_{kl}\Sigma^{-1/2}_{li} \\
&= \sum_{i,j,\gamma}\Sigma^{-1}_{ij}O_{j\gamma}\frac{\psi^2_\gamma}{l_\gamma}O'_{\gamma i} + \frac{1}{n}\sum_{\alpha,k,l}X_{\alpha k}\frac{\partial}{\partial X_{\alpha l}}OL^{-1}\Psi O'_{ki}\Sigma^{-1}_{ij}O\Psi O'_{jl} \qquad (4.8) \\
&= \sum_{i,j,\gamma}\Sigma^{-1}_{ij}O_{j\gamma}\frac{\psi^2_\gamma}{l_\gamma}O'_{\gamma i} + \frac{1}{n}\sum_{i,j,\alpha,k,l,\beta}\Sigma^{-1}_{ij}O\Psi O'_{jl}X_{\alpha k}\frac{\partial O_{k\beta}}{\partial X_{\alpha l}}\left[L^{-1}\Psi\right]_{\beta\beta}O'_{\beta i} \\
&\quad + \frac{1}{n}\sum_{i,j,\alpha,k,l,\beta}\Sigma^{-1}_{ij}O\Psi O'_{jl}X_{\alpha k}O_{k\beta}\frac{\partial\left[L^{-1}\Psi\right]_{\beta\beta}}{\partial X_{\alpha l}}O'_{\beta i} \\
&\quad + \frac{1}{n}\sum_{i,j,\alpha,k,l,\beta}\Sigma^{-1}_{ij}O\Psi O'_{jl}X_{\alpha k}O_{k\beta}\left[L^{-1}\Psi\right]_{\beta\beta}\frac{\partial O_{i\beta}}{\partial X_{\alpha l}} \\
&\quad + \frac{1}{n}\sum_{i,j,\alpha,k,l}X_{\alpha k}OL^{-1}\Psi O'_{ki}\Sigma^{-1}_{ij}\frac{\partial O_{j\beta}}{\partial X_{\alpha l}}\Psi_{\beta\beta}O'_{\beta l} \\
&\quad + \frac{1}{n}\sum_{i,j,\alpha,k,l}X_{\alpha k}OL^{-1}\Psi O'_{ki}\Sigma^{-1}_{ij}O_{j\beta}\frac{\partial\Psi_{\beta\beta}}{\partial X_{\alpha l}}O'_{\beta l} \\
&\quad + \frac{1}{n}\sum_{i,j,\alpha,k,l}X_{\alpha k}OL^{-1}\Psi O'_{ki}\Sigma^{-1}_{ij}O_{j\beta}\Psi_{\beta\beta}\frac{\partial O_{l\beta}}{\partial X_{\alpha l}} \\
&= \sum_{i,j,\gamma}\Sigma^{-1}_{ij}O_{j\gamma}\frac{\psi^2_\gamma}{l_\gamma}O'_{\gamma i} \\
&\quad + \frac{1}{n^2}\sum_{i,j,k,l,\alpha,\beta,\gamma\neq\beta,\epsilon}\Sigma^{-1}_{ij}O\Psi O'_{jl}X_{\alpha k}O_{k\gamma}\frac{O'_{\gamma l}O_{\epsilon\beta}+O'_{\gamma\epsilon}O_{l\beta}}{l_\beta-l_\gamma}X_{\alpha\epsilon}\left[L^{-1}\Psi\right]_{\beta\beta}O'_{\beta i} \\
&\quad + \frac{2}{n^2}\sum_{i,j,\alpha,\beta,\gamma,\epsilon}\Sigma^{-1}_{ij}O\Psi O'_{jl}X_{\alpha k}O_{k\beta}O'_{\gamma l}X_{\alpha\epsilon}O_{\epsilon\gamma}\frac{\partial\left[\psi_\beta/l_\beta\right]}{\partial l_\gamma}O'_{\beta i} \\
&\quad + \frac{1}{n^2}\sum_{i,j,k,l,\alpha,\beta,\gamma\neq\beta,\epsilon}\Sigma^{-1}_{ij}O\Psi O'_{jl}X_{\alpha k}O_{k\beta}\left[L^{-1}\Psi\right]_{\beta\beta}O_{i\gamma}\frac{O'_{\gamma l}O_{\epsilon\beta}+O'_{\gamma\epsilon}O_{l\beta}}{l_\beta-l_\gamma}X_{\alpha\epsilon} \\
&\quad + \frac{1}{n^2}\sum_{i,j,\alpha,\beta,\gamma\neq\beta,\epsilon}X_{\alpha k}OL^{-1}\Psi O'_{ki}\Sigma^{-1}_{ij}O_{j\gamma}\frac{O'_{\gamma l}O_{\epsilon\beta}+O'_{\gamma\epsilon}O_{l\beta}}{l_\beta-l_\gamma}X_{\alpha\epsilon}\Psi_{\beta\beta}O'_{\beta l} \\
&\quad + \frac{2}{n^2}\sum_{i,j,\alpha,\beta,\gamma,\epsilon}X_{\alpha k}OL^{-1}\Psi O'_{ki}\Sigma^{-1}_{ij}O_{j\beta}O'_{\gamma l}X_{\alpha\epsilon}O_{\epsilon\gamma}\frac{\partial\psi_\beta}{\partial l_\gamma}O'_{\beta l}
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{n^2}\sum_{i,j,k,l,\alpha,\beta,\gamma\neq\beta,\epsilon} X_{\alpha k}OL^{-1}\Psi O'_{ki}\Sigma^{-1}_{ij}O_{j\beta}\Psi_{\beta\beta}O_{l\gamma}\frac{O'_{\gamma l}O_{\epsilon\beta}+O'_{\gamma\epsilon}O_{l\beta}}{l_\beta-l_\gamma}X_{\alpha\epsilon}\\
&=\sum_{i,j,\gamma}\Sigma^{-1}_{ij}O_{j\gamma}\frac{\psi^2_\gamma}{l_\gamma}O'_{\gamma i}\\
&+\frac{1}{n}\sum_{i,j,l,\beta,\gamma\neq\beta}\Sigma^{-1}_{ij}O_{j\beta}\frac{l_\gamma\psi^2_\beta}{(l_\beta-l_\gamma)l_\beta}O'_{\beta i}+\frac{2}{n}\sum_{i,j,\gamma}\Sigma^{-1}_{ij}O_{j\gamma}\psi_\gamma l_\gamma\frac{\partial\left[\psi_\gamma/l_\gamma\right]}{\partial l_\gamma}O'_{\gamma i}\\
&+\frac{1}{n}\sum_{i,j,\beta,\gamma\neq\beta}\Sigma^{-1}_{ij}O_{j\gamma}\frac{\psi_\beta\psi_\gamma}{l_\beta-l_\gamma}O'_{\gamma i}+\frac{1}{n}\sum_{i,j,\beta,\gamma\neq\beta}\Sigma^{-1}_{ij}O_{j\gamma}\frac{\psi_\gamma\psi_\beta}{l_\beta-l_\gamma}O'_{\gamma i}\\
&+\frac{2}{n}\sum_{i,j,\gamma}\Sigma^{-1}_{ij}O_{j\gamma}\psi_\gamma\frac{\partial\psi_\gamma}{\partial l_\gamma}O'_{\gamma i}+\frac{1}{n}\sum_{i,j,\beta,\gamma\neq\beta,\epsilon}\Sigma^{-1}_{ij}O_{j\beta}\frac{\psi^2_\beta}{l_\beta-l_\gamma}O'_{\beta i}\\
&=\frac{n-p-1}{n}\sum_{i,j,\gamma}\Sigma^{-1}_{ij}O_{j\gamma}\frac{\psi^2_\gamma}{l_\gamma}O'_{\gamma i}+\frac{4}{n}\sum_{i,j,\gamma}\Sigma^{-1}_{ij}O_{j\gamma}\psi_\gamma\frac{\partial\psi_\gamma}{\partial l_\gamma}O'_{\gamma i}\\
&+\frac{2}{n}\sum_{i,j,l,\beta,\gamma\neq\beta}\Sigma^{-1}_{ij}O_{j\gamma}\psi_\gamma\frac{(\psi_\gamma-\psi_\beta)}{(l_\gamma-l_\beta)}O'_{\gamma i} \qquad (4.9)
\end{aligned}
$$

Thus

$$
\begin{aligned}
\mathrm{E}\Big[\Big|\mathrm{div}_{\mathrm{vec}(\tilde{X})}\ \mathrm{vec}\left(\tilde{X}H\right)\Big|\Big] &= \frac{1}{n}\,\mathrm{E}\left[\left|\sum_{i,j=1}^{p}\Sigma^{-1}_{ij}O\Psi^*O'_{ji}\right|\right]\\
&\le \frac{1}{n}\,\mathrm{E}\left[\sum_{k=1}^{p}\left|\left[L^{1/2}O'\Sigma^{-1}OL^{1/2}\right]_{kk}\right|\left|\frac{\psi^*_k}{l_k}\right|\right]\\
&\le \frac{1}{n}\sum_{k=1}^{p}\mathrm{E}\Big[\left[L^{1/2}O'\Sigma^{-1}OL^{1/2}\right]^{1+\frac{1}{\epsilon}}_{kk}\Big]^{\frac{\epsilon}{1+\epsilon}}\,\mathrm{E}\left[\left|\frac{\psi^*_k}{l_k}\right|^{1+\epsilon}\right]^{\frac{1}{1+\epsilon}}\\
&\le \frac{1}{n}\left(\mathrm{E}\left[\sum_{k=1}^{p}\left[L^{1/2}O'\Sigma^{-1}OL^{1/2}\right]^{1+\frac{1}{\epsilon}}_{kk}\right]\right)^{\frac{\epsilon}{1+\epsilon}}\left(\mathrm{E}\left[\sum_{k=1}^{p}\left|\frac{\psi^*_k}{l_k}\right|^{1+\epsilon}\right]\right)^{\frac{1}{1+\epsilon}}
\end{aligned}
$$

Since

$$
\begin{aligned}
\sum_{k=1}^{p}\left[L^{1/2}O'\Sigma^{-1}OL^{1/2}\right]^{1+\frac{1}{\epsilon}}_{kk} &\le \left(\sum_{k=1}^{p}\left|L^{1/2}O'\Sigma^{-1}OL^{1/2}\right|_{kk}\right)^{1+\frac{1}{\epsilon}}\\
&= \mathrm{tr}\big(L^{1/2}O'\Sigma^{-1}OL^{1/2}\big)^{1+\frac{1}{\epsilon}} = \mathrm{tr}\big(\Sigma^{-1}S\big)^{1+\frac{1}{\epsilon}} \sim (\chi^2_{pn})^{1+\frac{1}{\epsilon}}
\end{aligned}
$$

we get

$$
\mathrm{E}\Big[\Big|\mathrm{div}_{\mathrm{vec}(\tilde{X})}\ \mathrm{vec}\left(\tilde{X}H\right)\Big|\Big]
$$

$$\leq \frac{2\Gamma\left(1+\frac{1}{\epsilon}+\frac{np}{2}\right)^{\frac{\epsilon}{1+\epsilon}}}{n\Gamma\left(\frac{np}{2}\right)^{\frac{\epsilon}{1+\epsilon}}}\left(\mathrm{E}\left[\sum_{k=1}^{p}\left|\frac{\psi_k^*}{l_k}\right|^{1+\epsilon}\right]\right)^{\frac{1}{1+\epsilon}}<\infty$$

by assumption of the lemma. Therefore by Lemma 3 in Chételat and Wells [2012],

$$\begin{aligned}\mathrm{E}\Big[\mathrm{tr}\Big(\big[\Sigma^{-1}O\Psi O'\big]^2\Big)\Big] &= \mathrm{E}\big[\mathrm{tr}\big(L^{-1}\Psi\big)+\mathrm{tr}(X'\nabla_X G')\big]\\ &= \mathrm{E}\Bigg[\sum_{i,j,k=1}^{p}\Sigma_{ij}^{-1}O_{jk}\frac{\psi_k^2}{l_k}O'_{ki}\\ &\qquad +\frac{1}{n}\sum_{\alpha=1}^{n}\sum_{k,l=1}^{p}X_{\alpha k}\frac{\partial}{\partial X_{\alpha l}}OL^{-1}\Psi O'_{ki}\Sigma_{ij}^{-1}O\Psi O'_{jl}\Bigg]\\ &= \mathrm{E}\Bigg[\sum_{i,j,k=1}^{p}\Sigma_{ij}^{-1}O_{jk}\psi_k^*O'_{ki}\Bigg] \qquad \big(\text{ by (4.8) }\big).\end{aligned}$$

Finally, by Lemma 6, we conclude

$$= \mathrm{E}\Bigg[\sum_{k=1}^{p}\frac{n-p-1}{n}\frac{\psi_k^*}{l_k}+\frac{2}{n}\sum_{k=1}^{p}\frac{\partial\psi_k^*}{\partial l_k}+\frac{1}{n}\sum_{k\neq b}^{p}\frac{\psi_k^*-\psi_b^*}{l_k-l_b}\Bigg]$$

as desired. □

Lemma 8. *Let $l_1 > ... > l_p > 0$ be the eigenvalues of a $W_p(n,\Sigma)$-distributed matrix ,for some $\Sigma > 0$. If $n \geq p+1$, then*

(i) for any $1 \leq k \leq p$ and $0 \leq m < \frac{n-p-1}{2}$, $\mathrm{E}\Big[\frac{1}{|l_k|^m}\Big] < \infty$;

(ii) for any $1 \leq k \neq b \leq p$ and $1 \leq m < 2$, $\mathrm{E}\Big[\frac{1}{|l_k-l_b|^m}\Big] < \infty$;

(iii) for any $1 \leq k \neq b \neq e \leq p$ and $1 \leq m < 2$, $\mathrm{E}\Big[\frac{1}{|l_k-l_b|^m|l_k-l_e|^m}\Big] < \infty$.

Proof. First, notice that in (ii), we can take $k < b$ without loss of generality. Then

$$\mathrm{E}\left[\frac{1}{|l_k-l_b|^m}\right] \leq \mathrm{E}\left[\frac{1}{|l_k-l_{k+1}|^m}\right],$$

and it would be enough to show the r.h.s. finite for all $1 \leq k < p$ to show (ii).

Similarly, in (iii), we can take $b < e$ without loss of generality, and there are then three possibilities. Either $k < b < e$, in which case

$$\mathrm{E}\left[\frac{1}{|l_k - l_b|^m |l_k - l_e|^m}\right] \leq \mathrm{E}\left[\frac{1}{|l_k - l_{k+1}|^m |l_{e-1} - l_e|^m}\right],$$

or $b < k < e$ in which case

$$\mathrm{E}\left[\frac{1}{|l_k - l_b|^m |l_k - l_e|^m}\right] \leq \mathrm{E}\left[\frac{1}{|l_b - l_{b+1}|^m |l_{e-1} - l_e|^m}\right]$$

or $b < e < k$ in which case

$$\mathrm{E}\left[\frac{1}{|l_k - l_b|^m |l_k - l_e|^m}\right] \leq \mathrm{E}\left[\frac{1}{|l_b - l_{b+1}|^m |l_e - l_{e+1}|^m}\right]$$

Thus in any case it is enough to show that

$$\mathrm{E}\left[\frac{1}{|l_k - l_{k+1}|^m |l_b - l_{b+1}|^m}\right] < \infty$$

for all $1 \leq k < b < p$ to show (iii).

By Muirhead [1982], Theorem 3.2.18, the joint density of $l_1 > ... > l_p$ is given by

$$f_{l_1,...,l_p}(l_1, ..., l_p) = \frac{\pi^{p^2/2} 2^{-pn} |\Sigma|^{-n/2}}{\Gamma_p(p/2)\Gamma_p(n/2)} \prod_{i=1}^{p} l_i^{\frac{n-p-1}{2}} \prod_{1 \leq i < j \leq p} (l_i - l_j) \tag{4.10}$$

$$\int_{O(p)} \mathrm{etr}\left(-\frac{1}{2}\Sigma^{-1} H L H'\right) dH \; \mathbb{1}[l_1 > ... > l_p > 0] \tag{4.11}$$

for $L = \mathrm{diag}(l_1, ..., l_p)$. Define $I_2 = \{(i,j) \,|\, i < j \wedge (i,j) \neq (k, k+1)\}$ and $I_3 = \{(i,j) \,|\, i < j \wedge (i,j) \neq (k, k+1), (b, b+1)\}$. The expressions

$$P_1(l_1, ..., l_p) = \prod_{i \neq k}^{p} l_i^{n-p-1} \prod_{i<j}^{p} (l_i - l_j)^2$$

$$P_2(l_1, ..., l_p) = \prod_{i=1}^{p} l_i^{n-p-1} \prod_{(i,j) \in I_2} (l_i - l_j)^2$$

$$P_3(l_1, ..., l_p) = \prod_{i=1}^{p} l_i^{n-p-1} \prod_{(i,j) \in I_3} (l_i - l_j)^2$$

and $K = \frac{\pi^{p^2/2} 2^{-pn} |\Sigma|^{-n/2}}{\Gamma_p(p/2)\Gamma_p(n/2)}$ can then be defined to write

$$
\begin{aligned}
f_{l_1,...,l_p}(l_1,...,l_p) &= K l_k^{\frac{n-p-1}{2}} P_1^{1/2}(l_1,...,l_p) \\
&\qquad \int_{O(p)} \operatorname{etr}\left(-\frac{1}{2}\Sigma^{-1}HLH'\right) dH \; \mathbb{1}[l_1 > ... > l_p > 0] \\
&= K|l_k - l_{k+1}| P_2^{1/2}(l_1,...,l_p) \\
&\qquad \int_{O(p)} \operatorname{etr}\left(-\frac{1}{2}\Sigma^{-1}HLH'\right) dH \; \mathbb{1}[l_1 > ... > l_p > 0] \\
&= K|l_k - l_{k+1}||l_b - l_{b+1}| P_3^{1/2}(l_1,...,l_p) \\
&\qquad \int_{O(p)} \operatorname{etr}\left(-\frac{1}{2}\Sigma^{-1}HLH'\right) dH \; \mathbb{1}[l_1 > ... > l_p > 0] \,.
\end{aligned}
$$

The important point is that since $n - p - 1 \geq 0$, P_1, P_2 and P_3 are polynomials in $l_1, ..., l_p$. Define $x = l_{k+1} - l_k$, $y = l_{k+1} - l_b$ and $z = l_b - l_{b+1}$, so that $l_k = l_{b+1} + z + y + x$, $l_{k+1} = l_{b+1} + z + y$ and $l_b = l_{b+1} + z$. (It might happen that $k+1{=}b$, something which should be kept in mind.) Then

$$
\begin{aligned}
&P_2(l_1, ..., l_{k+1} + x, ..., l_{k+1}, ..., l_p), \\
&P_3(l_1, ..., l_{b+1} + z + x, ..., l_{b+1} + z, ..., l_b) && \text{if } k+1 = b, \\
&P_3(l_1, ..., l_{b+1} + z + y + x, ..., l_{b+1} + z + y, ..., l_{b+1} + z, ..., l_b) && \text{if } k+1 \neq b,
\end{aligned}
$$

must still be polynomials, in $\{l_i\}\backslash\{l_k\} \cup \{x\}$, $\{l_i\}\backslash\{l_k, l_b\} \cup \{x, z\}$ and $\{l_i\}\backslash\{l_k, l_{k+1}, l_b\} \cup \{x, y, z\}$ respectively. Therefore, for some finite degrees $D_1, ..., D_4$ one can write

$$
\begin{aligned}
P_1(l_1,...,l_p) &= \sum_{\substack{d_1+...+d_p \\ \leq D_1}} A^1_{d_1,...,d_p} l_1^{d_1} \cdots l_p^{d_p}, \\
P_2(l_1,...,l_{k+1}+x,...,l_{k+1},...,l_p) &= \sum_{\substack{d_1+...+d_p \\ \leq D_2}} A^2_{d_1,...,d_p} l_1^{d_1} \cdots x^{d_k} \cdots l_p^{d_p}, \\
&P_3(l_1,...,l_{b+1}+z+x,...,l_{b+1}+z,...,l_b) \\
&\quad = \sum_{\substack{d_1+...+d_p \\ \leq D_3}} A^3_{d_1,...,d_p} l_1^{d_1} \cdots x^{d_k} \cdots z^{d_b} \cdots l_p^{d_p}, \qquad \text{if } k+1 = b,
\end{aligned}
$$

$$P_3(l_1, ..., l_{b+1} + z + y + x, ..., l_{b+1} + z + y, ..., l_{b+1} + z, ..., l_b)$$
$$= \sum_{\substack{d_1+...+d_p \\ \leq D_4}} A^4_{d_1,...,d_p} l_1^{d_1} \cdots x^{d_k} \cdots y^{d_{k+1}} \cdots z^{d_b} \cdots l_p^{d_p}, \quad \text{if } k+1 \neq b.$$

for coefficients $A^1_{d_1,...,d_p}, ..., A^4_{d_1,...,d_p} \in \mathbb{R}$.

If we denote the greatest eigenvalue of Σ by $\lambda_{\max}$, then $\Sigma^{-1} \geq \lambda_{\max}^{-1} I$ so that

$$\int_{O(p)} \operatorname{etr}\left(-\frac{1}{2}\Sigma^{-1} HLH'\right) dH \leq \exp\left(-\frac{1}{2\lambda_{\max}} \sum_{i=1}^{p} l_i\right)$$

for any $l_1, ..., l_p \geq 0$.

Now, for (i), we can use (4.11) and $\mathbb{1}[l_1 > ... > l_p > 0] \leq \prod_{i=1}^{p} \mathbb{1}[l_i > 0]$ to find

$$\mathrm{E}\left[\frac{1}{|l_k|^m}\right] \leq K \int_{\mathbb{R}^p} \frac{1}{l_k^{m-\frac{n-p-1}{2}}} P_1^{1/2}(l_1, ..., l_p) \exp\left(-\frac{1}{2\lambda_{\max}} \sum_{i=1}^{p} l_i\right)$$
$$\mathbb{1}[l_1 > ... > l_p > 0]\, dl_1 \cdots dl_p$$
$$\leq K \sum_{\substack{d_1+...+d_p \\ \leq D_1}} \sqrt{\left|A^1_{d_1,...,d_p}\right|} \int_0^\infty l_1^{d_1/2} e^{-l_1/2\lambda_{\max}}\, dl_1 \cdots$$
$$\int_0^\infty l_k^{d_k/2-m+\frac{n-p-1}{2}} e^{-l_k/2\lambda_{\max}}\, dl_k \cdots \int_0^\infty l_p^{d_p/2} e^{-l_p/2\lambda_{\max}}\, dl_p.$$

Notice that for $i \neq k$, $\int_0^\infty l_i^{d_i/2} e^{-l_i/2\lambda_{\max}}\, dl_i < \infty$ for any $d_i \geq 0$, and $\int_0^\infty l_k^{d_k/2-m+\frac{n-p-1}{2}} e^{-l_k/2\lambda_{\max}}\, dl_k < \infty$ for all $d_k \geq 0$ iff $0 \leq m < \frac{n-p-1}{2}$. Thus $\mathrm{E}[1/|l_k|^m] < \infty$, as desired.

For (ii), we proceed similarly, but though a change of variables $(l_k, l_{k+1}) \to (l_{k+1} + x, l_k)$. Then, using

$$\mathbb{1}\Big[l_1 > ... > l_{k+1} + x > ... > l_{k+1} > ... > l_p > 0\Big] \leq \mathbb{1}[x > 0] \prod_{i \neq k}^{p} \mathbb{1}[l_i > 0],$$

we obtain

$$\mathrm{E}\left[\frac{1}{|l_k - l_{k+1}|^m}\right] \leq K \int_{\mathbb{R}^p} \frac{1}{|l_k - l_{k+1}|^{m-1}} P_2^{1/2}(l_1, ..., l_p)$$
$$\exp\left(-\frac{1}{2\lambda_{\max}} \sum_{i=1}^{p} l_i\right)\ \mathbb{1}[l_1 > ... > l_p > 0]\, dl_1 \cdots dl_p$$

$$\leq K \sum_{\substack{d_1+\ldots+d_p \\ \leq D_2}} \sqrt{\left|A^2_{d_1,\ldots,d_p}\right|} \int_0^\infty l_1^{d_1/2} e^{-l_1/2\lambda_{\max}}\, dl_1 \cdots$$
$$\int_0^\infty x^{d_k/2-m+1} e^{-x/2\lambda_{\max}}\, dx \cdots \int_0^\infty l_{k+1}^{d_{k+1}/2} e^{-l_{k+1}/\lambda_{\max}}\, dl_{k+1} \cdots$$
$$\int_0^\infty l_p^{d_p/2} e^{-l_p/2\lambda_{\max}}\, dl_p.$$

Then again, for $i \neq k$ and any $d_i \geq 0$, the respective integrals are finite; and to have $\int_0^\infty x^{d_k/2-m+1} e^{-x/2\lambda_{\max}}\, dx < \infty$ for all $d_k \geq 0$ requires $m < 2$. In such a case, we end up with $\mathrm{E}[1/|l_k - l_{k+1}|^m] < \infty$, as desired.

For (iii), we must consider separately the cases $k+1 = b$ and $k+1 \neq b$. In the first case, one can take the change of variables $(l_k, l_b, l_{b+1}) \longrightarrow (l_{b+1} + x + z, l_{b+1} + z, l_{b+1})$. Using that

$$\mathbb{1}\Big[l_1 > \ldots > l_{b+1} + x + z > \ldots > l_{b+1} + z > \ldots > l_{b+1} > \ldots > l_p > 0\Big]$$
$$\leq \mathbb{1}[x > 0]\, \mathbb{1}[z > 0] \prod_{i \neq k,b}^{p} \mathbb{1}[l_i > 0]\,,$$

we then obtain

$$\mathrm{E}\left[\frac{1}{|l_k - l_b|^m |l_b - l_{b+1}|^m}\right] \leq K \int_{\mathbb{R}^p} \frac{1}{|l_k - l_b|^{m-1} |l_b - l_{b+1}|^{m-1}}$$
$$P_3^{1/2}(l_1, \ldots, l_p) \exp\left(-\frac{1}{2\lambda_{\max}} \sum_{i=1}^{p} l_i\right)\, \mathbb{1}[l_1 > \ldots > l_p > 0]\, dl_1 \cdots dl_p$$
$$\leq K \sum_{\substack{d_1+\ldots+d_p \\ \leq D_3}} \sqrt{\left|A^3_{d_1,\ldots,d_p}\right|} \int_0^\infty l_1^{d_1/2} e^{-l_1/2\lambda_{\max}}\, dl_1 \cdots$$
$$\int_0^\infty x^{d_k/2-m+1} e^{-x/2\lambda_{\max}}\, dx \cdots \int_0^\infty z^{d_b/2-m+1} e^{-z/\lambda_{\max}}\, dz \cdots$$
$$\int_0^\infty l_{b+1}^{d_{b+1}/2} e^{-3l_{b+1}/2\lambda_{\max}}\, dl_{b+1} \cdots \int_0^\infty l_p^{d_p/2} e^{-l_p/2\lambda_{\max}}\, dl_p.$$

Again, all the integrals converge as long as $m < 2$, in which case we have $\mathrm{E}[1/|l_k - l_b|^m |l_b - l_{b+1}|^m] < \infty$, as desired.

Finally, for (iii) with $k+1 \neq b$, one can take the change of variables $(l_k, l_{k+1}, l_b, l_{b+1}) \longrightarrow (l_{b+1} + x + y + z, l_{b+1} + y + z, l_{b+1} + z, l_{b+1})$. Then using

that

$$\begin{aligned}\mathbb{1}\Big[&l_1 > ... > l_{b+1} + x + y + z > ... > l_{b+1} + y + z > ... > l_{b+1} + z \\ &> ... > l_{b+1} > ... > l_p > 0\Big] \\ &\leq \mathbb{1}[x > 0]\,\mathbb{1}[y > 0]\,\mathbb{1}[z > 0] \prod_{i \neq k, k+1, b}^{p} \mathbb{1}[l_i > 0]\,,\end{aligned}$$

we obtain

$$\begin{aligned}\mathrm{E}\left[\frac{1}{|l_k - l_{k+1}|^m |l_b - l_{b+1}|^m}\right] &\leq K \int_{\mathbb{R}^p} \frac{1}{|l_k - l_{k+1}|^{m-1} |l_b - l_{b+1}|^{m-1}} \\ &\quad P_4^{1/2}(l_1, ..., l_p) \exp\left(-\frac{1}{2\lambda_{\max}} \sum_{i=1}^{p} l_i\right)\ \mathbb{1}[l_1 > ... > l_p > 0]\, dl_1 \cdots dl_p \\ &\leq K \sum_{\substack{d_1 + ... + d_p \\ \leq D_4}} \sqrt{\left|A^4_{d_1,...,d_p}\right|} \int_0^\infty l_1^{d_1/2} e^{-l_1/2\lambda_{\max}}\, dl_1 \cdots \\ &\quad \int_0^\infty x^{d_k/2 - m + 1} e^{-x/2\lambda_{\max}}\, dx \cdots \int_0^\infty y^{d_k/2 - m + 1} e^{-y/\lambda_{\max}}\, dy \cdots \\ &\quad \int_0^\infty z^{d_b/2 - m + 1} e^{-3z/2\lambda_{\max}}\, dz \cdots \int_0^\infty l_{b+1}^{d_{b+1}/2} e^{-2l_{b+1}/\lambda_{\max}}\, dl_{b+1} \cdots \\ &\quad \int_0^\infty l_p^{d_p/2} e^{-l_p/2\lambda_{\max}}\, dl_p.\end{aligned}$$

All the integrals converge as long as $m < 2$, in which case we have $\mathrm{E}[1/|l_k - l_{b+1}|^m |l_b - l_{b+1}|^m] < \infty$, as desired. □

Proof of Theorem 6. By independence, it is clear that

$$\begin{aligned}\mathrm{E}\Big[L_H(\hat{\Sigma}, \Sigma)\Big] &= \mathrm{E}_{\hat{\rho}}\left[\mathrm{E}_S\left[L_H(\hat{\Sigma}, \Sigma) \,|\, \hat{\rho}\right]\right] \\ &= \mathrm{E}_{\hat{\rho}}\left[\mathrm{E}_S\left[L_H(\hat{\Sigma}, \Sigma)\right]\right],\end{aligned}$$

so we can treat $\hat{\rho}$ as a constant throughout the calculations, without loss of generality. Define the auxiliary terms $\psi_k = \hat{\gamma}_k + \hat{\sigma}^2$ and

$$\psi_k^* = \frac{n - p - 1}{n} \frac{\psi_k^2}{l_k} + 4 \frac{\psi_k}{n} \frac{\partial \psi_k}{\partial l_k} + 2 \frac{\psi_k}{n} \sum_{b \neq k}^{p} \frac{\psi_k - \psi_b}{l_k - l_b}$$

for all $1 \leq k \leq p$, and consider:

$$
\begin{aligned}
R_1 &= \sum_{k=1}^{p} \frac{n-p-1}{n} \frac{\psi_k^*}{l_k} + \frac{2}{n} \sum_{k=1}^{p} \frac{\partial \psi_k^*}{\partial l_k} + \frac{1}{n} \sum_{k \neq b}^{p} \frac{\psi_k^* - \psi_b^*}{l_k - l_b} \\
&= \left\{ \frac{(n-p-1)^2}{n^2} \sum_{k=1}^{p} \frac{\psi_k^2}{l_k^2} + 4 \frac{n-p-1}{n^2} \sum_{k=1}^{p} \frac{\psi_k}{l_k} \frac{\partial \psi_k}{\partial l_k} \right. \\
&\quad \left. + 2 \frac{n-p-1}{n^2} \sum_{k \neq b=1}^{p} \frac{\psi_k}{l_k} \frac{\psi_k - \psi_b}{l_k - l_b} \right\} + \left\{ 4 \frac{n-p-1}{n^2} \sum_{k=1}^{p} \frac{\psi_k}{l_k} \frac{\partial \psi_k}{\partial l_k} \right. \\
&\quad - 2 \frac{n-p-1}{n^2} \sum_{k=1}^{p} \frac{\psi_k^2}{l_k^2} + \frac{8}{n^2} \sum_{k=1}^{p} \left(\frac{\partial \psi_k}{\partial l_k} \right)^2 + \frac{8}{n^2} \sum_{k=1}^{p} \psi_k \frac{\partial^2 \psi_k}{\partial l_k^2} \\
&\quad + \frac{4}{n^2} \sum_{k \neq b=1}^{p} \frac{\partial \psi_k}{\partial l_k} \frac{\psi_k - \psi_b}{l_k - l_b} + \frac{4}{n^2} \sum_{k \neq b=1}^{p} \psi_k \frac{\frac{\partial \psi_k}{\partial l_k} - \frac{\partial \psi_b}{\partial l_k}}{l_k - l_b} \\
&\quad \left. - \frac{2}{n^2} \sum_{k \neq b=1}^{p} \left(\frac{\psi_k - \psi_b}{l_k - l_b} \right)^2 \right\} + \left\{ \left(2 \frac{n-p-1}{n^2} \sum_{k \neq b=1}^{p} \frac{\psi_k}{l_k} \frac{\psi_k - \psi_b}{l_k - l_b} \right. \right. \\
&\quad \left. - \frac{n-p-1}{n^2} \sum_{k \neq b=1}^{p} \frac{\psi_k}{l_k} \frac{\psi_b}{l_b} \right) + \left(\frac{4}{n^2} \sum_{k \neq b=1}^{p} \psi_k \frac{\frac{\partial \psi_k}{\partial l_k} - \frac{\partial \psi_b}{\partial l_b}}{l_k - l_b} \right. \\
&\quad \left. + \frac{4}{n^2} \sum_{k \neq b=1}^{p} \frac{\partial \psi_k}{\partial l_k} \frac{\psi_k - \psi_b}{l_k - l_b} \right) + \left(\frac{2}{n^2} \sum_{k \neq b \neq e=1}^{p} \frac{\psi_k}{l_k - l_b} \left(\frac{\psi_k - \psi_e}{l_k - l_e} - \frac{\psi_b - \psi_e}{l_b - l_e} \right) \right. \\
&\quad \left. \left. + \frac{2}{n^2} \sum_{k \neq b \neq e=1}^{p} \frac{\psi_k - \psi_b}{l_k - l_b} \frac{\psi_k - \psi_e}{l_k - l_e} + \frac{2}{n^2} \sum_{k \neq b=1}^{p} \left(\frac{\psi_k - \psi_b}{l_k - l_b} \right)^2 \right) \right\} \\
&= \frac{(n-p-1)(n-p-2)}{n^2} \sum_{k=1}^{p} \frac{\psi_k^2}{l_k^2} - \frac{(n-p-1)}{n^2} \left(\sum_{k=1}^{p} \frac{\psi_k}{l_k} \right)^2 \\
&\quad + \frac{8}{n^2} \sum_{k=1}^{p} \left(\frac{\partial \psi_k}{\partial l_k} \right)^2 + \frac{8}{n^2} \sum_{k=1}^{p} \psi_k \frac{\partial^2 \psi_k}{\partial l_k^2} + 8 \frac{n-p-1}{n^2} \sum_{k=1}^{p} \frac{\psi_k}{l_k} \frac{\partial \psi_k}{\partial l_k} \\
&\quad + 4 \frac{n-p-1}{n^2} \sum_{k \neq b=1}^{p} \frac{\psi_k}{l_k} \frac{\psi_k - \psi_b}{l_k - l_b} + \frac{8}{n^2} \sum_{k \neq b=1}^{p} \frac{\partial \psi_k}{\partial l_k} \frac{\psi_k - \psi_b}{l_k - l_b} \\
&\quad + \frac{4}{n^2} \sum_{k \neq b=1}^{p} \psi_k \frac{\frac{\partial \psi_k}{\partial l_k} - \frac{\partial \psi_b}{\partial l_b}}{l_k - l_b} + \frac{4}{n^2} \sum_{k \neq b=1}^{p} \psi_k \frac{\frac{\partial \psi_k}{\partial l_k} - \frac{\partial \psi_b}{\partial l_k}}{l_k - l_b} \\
&\quad + \frac{2}{n^2} \sum_{k \neq b \neq e=1}^{p} \frac{\psi_k}{l_k - l_b} \left(\frac{\psi_k - \psi_e}{l_k - l_e} - \frac{\psi_b - \psi_e}{l_b - l_e} \right) \\
&\quad + \frac{2}{n^2} \sum_{k \neq b \neq e=1}^{p} \frac{\psi_k - \psi_b}{l_k - l_b} \frac{\psi_k - \psi_e}{l_k - l_e}.
\end{aligned}
$$

Now, by Hölder's inequality, we find:

$$
\begin{aligned}
\mathrm{E}\Big[|R_1|\Big] \leq{}& \frac{|n-p-1||n-p-2|}{n^2}\sum_{k=1}^{p}\mathrm{E}\left[\left|\frac{\psi_k}{l_k}\right|^2\right] \\
&+ \frac{|n-p-1|}{n^2}\left(\sum_{k=1}^{p}\mathrm{E}\left[\left|\frac{\psi_k}{l_k}\right|^2\right]^{\frac{1}{2}}\right)^2 + \frac{8}{n^2}\sum_{k=1}^{p}\mathrm{E}\left[\left|\frac{\partial\psi_k}{\partial l_k}\right|^2\right] \\
&+ \frac{8}{n^2}\sum_{k=1}^{p}\mathrm{E}\left[\left|\psi_k\frac{\partial^2\psi_k}{\partial l_k^2}\right|\right] + 8\frac{|n-p-1|}{n^2}\sum_{k=1}^{p}\mathrm{E}\left[\left|\frac{\psi_k}{l_k}\right|^2\right]^{\frac{1}{2}}\mathrm{E}\left[\left|\frac{\partial\psi_k}{\partial l_k}\right|^2\right]^{\frac{1}{2}} \\
&+ 4\frac{|n-p-1|}{n^2}\sum_{k\neq b=1}^{p}\mathrm{E}\left[\left|\frac{\psi_k}{l_k}\right|^{4.5}\right]^{\frac{1}{4.5}} \\
&\qquad\qquad\Big(\mathrm{E}\big[|\psi_k|^{4.5}\big]^{\frac{1}{4.5}} + \mathrm{E}\big[|\psi_b|^{4.5}\big]^{\frac{1}{4.5}}\Big)\,\mathrm{E}\left[\frac{1}{|l_k-l_b|^{1.8}}\right]^{\frac{1}{1.8}} \\
&+ \frac{8}{n^2}\sum_{k\neq b=1}^{p}\left(\mathrm{E}\left[\left|\frac{\partial\psi_k}{\partial l_k}\right|^{4.5}\right]^{\frac{1}{4.5}}\mathrm{E}\big[|\psi_k|^{4.5}\big]^{\frac{1}{4.5}}\right. \\
&\qquad\qquad\left.+\,\mathrm{E}\left[\left|\frac{\partial\psi_k}{\partial l_k}\right|^{4.5}\right]^{\frac{1}{4.5}}\mathrm{E}\big[|\psi_b|^{4.5}\big]^{\frac{1}{4.5}}\right)\mathrm{E}\left[\frac{1}{|l_k-l_b|^{1.8}}\right]^{\frac{1}{1.8}} \\
&+ \frac{4}{n^2}\sum_{k\neq b=1}^{p}\mathrm{E}\big[|\psi_k|^{4.5}\big]^{\frac{1}{4.5}}\left(\mathrm{E}\left[\left|\frac{\partial\psi_k}{\partial l_k}\right|^{4.5}\right]^{\frac{1}{4.5}}\right. \\
&\qquad\qquad\left.+\,\mathrm{E}\left[\left|\frac{\partial\psi_b}{\partial l_b}\right|^{4.5}\right]^{\frac{1}{4.5}}\right)\mathrm{E}\left[\frac{1}{|l_k-l_b|^{1.8}}\right]^{\frac{1}{1.8}} \\
&+ \frac{4}{n^2}\sum_{k\neq b=1}^{p}\mathrm{E}\big[|\psi_k|^{4.5}\big]^{\frac{1}{4.5}}\left(\mathrm{E}\left[\left|\frac{\partial\psi_k}{\partial l_k}\right|^{4.5}\right]^{\frac{1}{4.5}}\right. \\
&\qquad\qquad\left.+\,\mathrm{E}\left[\left|\frac{\partial\psi_b}{\partial l_k}\right|^{4.5}\right]^{\frac{1}{4.5}}\right)\mathrm{E}\left[\frac{1}{|l_k-l_b|^{1.8}}\right]^{\frac{1}{1.8}} \\
&+ \frac{2}{n^2}\sum_{k\neq b\neq e=1}^{p}\Big(\mathrm{E}\big[|\psi_k|^{4.5}\big]^{\frac{1}{4.5}} + \mathrm{E}\big[|\psi_b|^{4.5}\big]^{\frac{1}{4.5}}\Big) \\
&\qquad\qquad\Big(\mathrm{E}\big[|\psi_k|^{4.5}\big]^{\frac{1}{4.5}} + \mathrm{E}\big[|\psi_e|^{4.5}\big]^{\frac{1}{4.5}}\Big)\,\mathrm{E}\left[\left|\frac{1}{(l_k-l_b)(l_k-l_e)}\right|^{1.8}\right]^{\frac{1}{1.8}}
\end{aligned}
$$

$$+\frac{2}{n^2}\sum_{k\neq b\neq e=1}^{p}\Big(\mathrm{E}\big[|\psi_k|^{4.5}\big]^{\frac{1}{4.5}}+\mathrm{E}\big[|\psi_b|^{4.5}\big]^{\frac{1}{4.5}}\Big)$$
$$\Big(\mathrm{E}\big[|\psi_k|^{4.5}\big]^{\frac{1}{4.5}}+\mathrm{E}\big[|\psi_e|^{4.5}\big]^{\frac{1}{4.5}}\Big)\,\mathrm{E}\left[\left|\frac{1}{(l_k-l_b)(l_k-l_e)}\right|^{1.8}\right]^{\frac{1}{1.8}}.$$

Similarly, consider

$$R_2=\frac{n-p-1}{n}\sum_{k=1}^{p}\frac{\psi_k}{l_k}+\frac{2}{n}\sum_{k=1}^{p}\frac{\partial\psi_k}{\partial l_k}+\frac{1}{n}\sum_{k\neq b=1}^{p}\frac{\psi_k-\psi_b}{l_k-l_b},$$

so that

$$\mathrm{E}\Big[|R_2|\Big]\leq\frac{|n-p-1|}{n}\sum_{k=1}^{p}\mathrm{E}\left[\left|\frac{\psi_k}{l_k}\right|\right]+\frac{2}{n}\sum_{k=1}^{p}\mathrm{E}\left[\left|\frac{\partial\psi_k}{\partial l_k}\right|\right]$$
$$+\frac{1}{n}\sum_{k\neq b=1}^{p}\Big(\mathrm{E}\big[|\psi_k|^{2.25}\big]^{\frac{1}{2.25}}+\mathrm{E}\big[|\psi_b|^{2.25}\big]^{\frac{1}{2.25}}\Big)\,\mathrm{E}\left[\frac{1}{|l_k-l_b|^{1.8}}\right]^{\frac{1}{1.8}},$$

Moreover, for any $\epsilon>0$,

$$\mathrm{E}\left[\sum_{k=1}^{p}\left|\frac{\psi_k^*}{l_k}\right|\right]\leq\frac{|n-p-1|}{n}\sum_{k=1}^{p}\mathrm{E}\left[\left|\frac{\psi_k}{l_k}\right|^{2(1+\epsilon)}\right]^{\frac{1}{1+\epsilon}}+\frac{4}{n}\sum_{k=1}^{p}\mathrm{E}\left[\left|\frac{\psi_k}{l_k}\right|^{2(1+\epsilon)}\right]$$
$$\cdot\,\mathrm{E}\left[\left|\frac{\partial\psi_k}{\partial l_k}\right|^{2(1+\epsilon}\right]^{\frac{1}{2(1+\epsilon)}}+\frac{2}{n}\sum_{k\neq b}^{p}\mathrm{E}\left[\frac{1}{|l_k-l_b|^{1.8(1+\epsilon)}}\right]^{\frac{1}{1.8(1+\epsilon)}}$$
$$\cdot\,\mathrm{E}\left[\left|\frac{\psi_k}{l_k}\right|^{4.5(1+\epsilon)}\right]^{\frac{1}{4.5(1+\epsilon)}}\left(\mathrm{E}\Big[|\psi_k|^{4.5(1+\epsilon)}\Big]^{\frac{1}{4.5(1+\epsilon)}}+\mathrm{E}\Big[|\psi_b|^{4.5(1+\epsilon)}\Big]^{\frac{1}{4.5(1+\epsilon)}}\right).$$

Note that for any $1\leq k\leq p$ and $\epsilon>0$,

$$\mathrm{E}\big[|\psi_k|^{4.5}\big]^{\frac{1}{4.5}}\leq\mathrm{E}\left[\left|\frac{\psi_k}{l_k}\right|^{9(1+\epsilon)}\right]^{\frac{1}{9(1+\epsilon)}}\mathrm{E}\Big[|l_k|^{\frac{9\epsilon}{1+\epsilon}}\Big]^{\frac{1+\epsilon}{9\epsilon}},$$

and for any $m>1$, $\mathrm{E}[|l_k|^m]^2\leq\mathrm{E}[\mathrm{tr}(S^{2m})]\leq\mathrm{tr}(\Sigma)^{2m}\,\mathrm{E}\Big[\mathrm{tr}(S\Sigma^{-1})^{2m}\Big]$ $=\mathrm{tr}(\Sigma)^{2m}\,\mathrm{E}\big[(\chi^2_{np})^{2m}\big]<\infty$. Now consider that, for any $m>1$,

$$|\psi_k|^m\leq 2^{m-1}(|\hat\gamma_k|^m+\left|\hat\sigma^2\right|^m)$$
$$\left|\frac{\partial\psi_k}{\partial l_k}\right|^m\leq 2^{m-1}(\left|\frac{\partial\hat\gamma_k}{\partial l_k}\right|^m+\left|\frac{\partial\hat\sigma^2}{\partial l_k}\right|^m)$$
$$\left|\psi_k\frac{\partial^2\psi_k}{\partial l_k^2}\right|^m=\left|\hat\gamma_k+\hat\sigma^2\right|\left|\frac{\partial^2\hat\gamma_k}{\partial l_k^2}+\frac{\partial^2\hat\sigma^2}{\partial l_k^2}\right|^m.$$

Therefore, since $\hat{\Gamma}$ satisfies the weak regularity conditions and $\hat{\Sigma} \in V_p(\hat{\Gamma})$, we obtain $\mathrm{E}[|R_1|] < \infty$, $\mathrm{E}[|R_2|] < \infty$ and $\mathrm{E}\Big[\sum_{k=1}^{p} \Big|\frac{\psi_k^*}{l_k}\Big|\Big] < \infty$.

Therefore, all the regularity conditions of Lemmas 6 and 7 are satisfied, and we have for $\Psi = \mathrm{diag}(\psi_1, ..., \psi_p)$

$$
\begin{aligned}
\mathrm{E}\Big[L\left(\hat{\Sigma}, \Sigma\right)\Big] &= \frac{1}{p}\,\mathrm{E}\Big[\mathrm{tr}\Big(\big[\Sigma^{-1}O\Psi O'\big]^2\Big) - 2\,\mathrm{tr}\big(\Sigma^{-1}O\Psi O'\big) + p\Big] \\
&= \frac{1}{p}\,\mathrm{E}[R_1 - 2R_2 + p] \\
&= \mathrm{E}\Bigg[\frac{(n-p-1)(n-p-2)}{n^2p}\sum_{k=1}^{p}\frac{\psi_k^2}{l_k^2} - \frac{(n-p-1)}{n^2p}\left(\sum_{k=1}^{p}\frac{\psi_k}{l_k}\right)^2 \\
&\quad + \frac{8}{n^2p}\sum_{k=1}^{p}\left(\frac{\partial\psi_k}{\partial l_k}\right)^2 + \frac{8}{n^2p}\sum_{k=1}^{p}\psi_k\frac{\partial^2\psi_k}{\partial l_k^2} + 8\frac{n-p-1}{n^2p}\sum_{k=1}^{p}\frac{\psi_k}{l_k}\frac{\partial\psi_k}{\partial l_k} \\
&\quad + 4\frac{n-p-1}{n^2p}\sum_{k\neq b=1}^{p}\frac{\psi_k}{l_k}\frac{\psi_k-\psi_b}{l_k-l_b} + \frac{8}{n^2p}\sum_{k\neq b=1}^{p}\frac{\partial\psi_k^n}{\partial l_k}\frac{\psi_k-\psi_b}{l_k-l_b} \\
&\quad + \frac{4}{n^2p}\sum_{k\neq b=1}^{p}\psi_k^n\frac{\frac{\partial\psi_k}{\partial l_k}-\frac{\partial\psi_b}{\partial l_b}}{l_k-l_b} + \frac{4}{n^2p}\sum_{k\neq b=1}^{p}\psi_k^n\frac{\frac{\partial\psi_k}{\partial l_k}-\frac{\partial\psi_b}{\partial l_k}}{l_k-l_b} \\
&\quad + \frac{2}{n^2p}\sum_{k\neq b\neq e=1}^{p}\frac{\psi_k}{l_k-l_b}\left(\frac{\psi_k-\psi_e}{l_k-l_e} - \frac{\psi_b-\psi_e}{l_b-l_e}\right) \\
&\quad + \frac{2}{n^2p}\sum_{k\neq b\neq e=1}^{p}\frac{\psi_k-\psi_b}{l_k-l_b}\frac{\psi_k-\psi_e}{l_k-l_e} - 2\frac{n-p-1}{np}\sum_{k=1}^{p}\frac{\psi_k}{l_k} \\
&\quad - \frac{4}{np}\sum_{k=1}^{p}\frac{\partial\psi_k}{\partial l_k} - \frac{2}{np}\sum_{k\neq b=1}^{p}\frac{\psi_k-\psi_b}{l_k-l_b} + 1\Bigg].
\end{aligned}
$$

We can now collect the terms of order 1 and $1/p$, defining

$$
\begin{aligned}
F(\hat{\Sigma}) &= \frac{(n-p-1)(n-p-2)}{n^2p}\sum_{k=1}^{p}\frac{\psi_k^2}{l_k^2} - \frac{n-p-1}{n^2p}\left(\sum_{k=1}^{p}\frac{\psi_k}{l_k}\right)^2 \\
&\quad + 4\frac{n-p-1}{n^2p}\sum_{k\neq b=1}^{p}\frac{\psi_k}{l_k}\frac{\psi_k-\psi_b}{l_k-l_b} + \frac{2}{n^2p}\sum_{k\neq b\neq e=1}^{p}\frac{\psi_k-\psi_b}{l_k-l_b}\frac{\psi_k-\psi_e}{l_k-l_e} \\
&\quad + \frac{2}{n^2p}\sum_{k\neq b\neq e=1}^{p}\frac{\psi_k}{l_k-l_b}\left(\frac{\psi_k-\psi_e}{l_k-l_e} - \frac{\psi_b-\psi_e}{l_b-l_e}\right) \\
&\quad - 2\frac{n-p-1}{np}\sum_{k=1}^{p}\frac{\psi_k}{l_k} - \frac{2}{np}\sum_{k\neq b=1}^{p}\frac{\psi_k-\psi_b}{l_k-l_b} + 1
\end{aligned}
$$

and

$$G(\hat{\Sigma}) = \frac{8}{n^2 p}\sum_{k=1}^{p}\left(\frac{\partial \psi_k}{\partial l_k}\right)^2 + \frac{8}{n^2 p}\sum_{k=1}^{p}\psi_k\frac{\partial^2 \psi_k}{\partial l_k^2} + 8\frac{n-p-1}{n^2 p}\sum_{k=1}^{p}\frac{\psi_k}{l_k}\frac{\partial \psi_k}{\partial l_k}$$
$$+ \frac{8}{n^2 p}\sum_{k\neq b=1}^{p}\frac{\partial \psi_k}{\partial l_k}\frac{\psi_k - \psi_b}{l_k - l_b} + \frac{4}{n^2 p}\sum_{k\neq b=1}^{p}\psi_k\frac{\frac{\partial \psi_k}{\partial l_k} - \frac{\partial \psi_b}{\partial l_b}}{l_k - l_b}$$
$$+ \frac{4}{n^2 p}\sum_{k\neq b=1}^{p}\psi_k\frac{\frac{\partial \psi_k}{\partial l_k} - \frac{\partial \psi_b}{\partial l_k}}{l_k - l_b} - \frac{4}{np}\sum_{k=1}^{p}\frac{\partial \psi_k}{\partial l_k}$$

so that $\mathrm{E}\Big[L(\hat{\Sigma}, \Sigma)\Big] = \mathrm{E}\Big[F(\hat{\Sigma}) + G(\hat{\Sigma})\Big]$, with $\mathrm{E}\Big[\Big|F(\hat{\Sigma}) + G(\hat{\Sigma})\Big|\Big] \leq \mathrm{E}[|R_1|] + 2\,\mathrm{E}[|R_2|] + p < \infty$, as desired. Plugging in $\psi_k = \hat{\gamma}_k + \hat{\sigma}^2$ yields, after a bit of algebra:

$$F(\hat{\Gamma} + \hat{\sigma}^2 I) = \frac{(n-p-1)(n-p-2)}{n^2 p}\sum_{k=1}^{\rho}\frac{\hat{\gamma}_k^2}{l_k^2}$$
$$+ 2\frac{(n-p-1)(n-p-2)}{n^2 p}\sum_{k=1}^{\rho}\frac{\hat{\gamma}_k\hat{\sigma}^2}{l_k^2} + \frac{(n-p-1)(n-p-2)}{n^2 p}\sum_{c=1}^{p}\frac{\hat{\sigma}_\rho^4}{l_c^2}$$
$$- \frac{n-p-1}{n^2 p}\left(\sum_{k=1}^{\rho}\frac{\hat{\gamma}_k}{l_k}\right)^2 - 2\frac{n-p-1}{n^2 p}\sum_{c=1}^{p}\frac{\hat{\sigma}^2}{l_c}\sum_{k=1}^{\rho}\frac{\hat{\gamma}_k}{l_k}$$
$$- \frac{n-p-1}{n^2 p}\left(\sum_{c=1}^{p}\frac{\hat{\sigma}^2}{l_c}\right)^2 + 4\frac{n-p-1}{n^2 p}\sum_{k\neq b}^{\rho}\frac{\hat{\gamma}_k}{l_k}\frac{\hat{\gamma}_k - \hat{\gamma}_b}{l_k - l_b}$$
$$+ 4\frac{n-p-1}{n^2 p}\sum_{k=1}^{\rho}\sum_{c=\rho+1}^{p}\frac{\hat{\gamma}_k}{l_k}\frac{\hat{\gamma}_k}{l_k - l_c} + 4\frac{n-p-1}{n^2 p}\sum_{k=1}^{\rho}\sum_{c=\rho+1}^{p}\frac{\hat{\sigma}^2}{l_c}\frac{\hat{\gamma}_k}{l_k - l_c}$$
$$+ \frac{2}{n^2 p}\sum_{k\neq b\neq e=1}^{\rho}\frac{\hat{\gamma}_k - \hat{\gamma}_b}{l_k - l_b}\frac{\hat{\gamma}_k - \hat{\gamma}_e}{l_k - l_e} + \frac{2}{n^2 p}\sum_{k\neq b=1}^{\rho}\sum_{c=\rho+1}^{p}\frac{\hat{\gamma}_k}{l_k - l_c}\frac{\hat{\gamma}_b}{l_b - l_c}$$
$$- \frac{6}{n^2 p}\sum_{k\neq b}^{\rho}\sum_{c=\rho+1}^{p}\frac{\hat{\gamma}_k + \hat{\sigma}^2}{l_k - l_c}\frac{\hat{\gamma}_b}{l_b - l_c} + \frac{6}{n^2 p}\sum_{k=1}^{\rho}\sum_{c\neq d}^{p}\frac{\hat{\gamma}_k + \hat{\sigma}^2}{l_k - l_c}\frac{\hat{\gamma}_k}{l_k - l_d}$$
$$- \frac{2}{n^2 p}\sum_{k=1}^{\rho}\sum_{c\neq d}^{p}\frac{\hat{\gamma}_k}{l_k - l_c}\frac{\hat{\gamma}_k}{l_k - l_d} + \frac{2}{n^2 p}\sum_{k\neq b\neq e=1}^{\rho}\frac{\hat{\gamma}_k}{l_k - l_b}\left(\frac{\hat{\gamma}_k - \hat{\gamma}_e}{l_k - l_e} - \frac{\hat{\gamma}_b - \hat{\gamma}_e}{l_b - l_e}\right)$$
$$+ \frac{4}{n^2 p}\sum_{k\neq b=1}^{p}\sum_{c=\rho+1}^{p}\frac{\hat{\gamma}_k - \hat{\gamma}_b}{l_k - l_b}\frac{\hat{\gamma}_k}{l_k - l_c} + \frac{6}{n^2 p}\sum_{k\neq b}^{\rho}\sum_{c=\rho+1}^{p}\frac{\hat{\gamma}_k - \hat{\gamma}_b}{l_k - l_b}\frac{\hat{\gamma}_k + \hat{\sigma}^2}{l_k - l_c}$$
$$- 2\frac{n-p-1}{np}\sum_{k=1}^{\rho}\frac{\hat{\gamma}_k}{l_k} - 2\frac{n-p-1}{np}\sum_{c=1}^{p}\frac{\hat{\sigma}^2}{l_c} - \frac{2}{np}\sum_{k\neq b=1}^{\rho}\frac{\hat{\gamma}_k - \hat{\gamma}_k}{l_k - l_b}$$

$$- \frac{4}{np} \sum_{k=1}^{\rho} \sum_{c=\rho+1}^{p} \frac{\hat{\gamma}_k}{l_k - l_c} + 1 \tag{4.12}$$

and

$$\begin{aligned}
G(\hat{\Gamma} + \hat{\sigma}^2 I) = {} & \frac{8}{n^2 p} \sum_{k=1}^{\rho} \left(\frac{\partial \hat{\gamma}_k}{\partial l_k} \right)^2 + \frac{16}{n^2 p} \sum_{k=1}^{\rho} \frac{\partial \hat{\gamma}_k}{\partial l_k} \frac{\partial \hat{\sigma}^2}{\partial l_k} + \frac{8}{n^2 p} \sum_{k=1}^{p} \left(\frac{\partial \hat{\sigma}^2}{\partial l_k} \right)^2 \\
& + \frac{8}{n^2 p} \sum_{k=1}^{\rho} \hat{\gamma}_k \frac{\partial^2 \hat{\gamma}_k}{\partial l_k^2} + \frac{8}{n^2 p} \sum_{k=1}^{p} \hat{\sigma}^2 \frac{\partial^2 \hat{\sigma}^2}{\partial l_k^2} + 8 \frac{n-p-1}{n^2 p} \sum_{k=1}^{\rho} \frac{\hat{\gamma}_k}{l_k} \frac{\partial \hat{\gamma}_k}{\partial l_k} \\
& + 8 \frac{n-p-1}{n^2 p} \sum_{k=1}^{\rho} \frac{\hat{\gamma}_k}{l_k} \frac{\partial \hat{\sigma}^2}{\partial l_k} + 8 \frac{n-p-1}{n^2 p} \sum_{k=1}^{\rho} \frac{\hat{\sigma}^2}{l_k} \frac{\partial \hat{\gamma}_k}{\partial l_k} \\
& + 8 \frac{n-p-1}{n^2 p} \sum_{k=1}^{p} \frac{\hat{\sigma}^2}{l_k} \frac{\partial \hat{\sigma}^2}{\partial l_k} + \frac{8}{n^2 p} \sum_{k \neq b=1}^{\rho} \frac{\partial \hat{\gamma}_k}{\partial l_k} \frac{\hat{\gamma}_k - \hat{\gamma}_k}{l_k - l_b} \\
& + \frac{8}{n^2 p} \sum_{k=1}^{\rho} \sum_{c=\rho+1}^{p} \frac{\partial \hat{\gamma}_k}{\partial l_k} \frac{\hat{\gamma}_k}{l_k - l_c} + \frac{8}{n^2 p} \sum_{k=1}^{\rho} \sum_{c=\rho+1}^{p} \frac{\partial \hat{\sigma}^2}{\partial l_c} \frac{\hat{\gamma}_k}{l_k - l_c} \\
& + \frac{4}{n^2 p} \sum_{k \neq b=1}^{\rho} \hat{\gamma}_k \frac{\frac{\partial \hat{\gamma}_k}{\partial l_k} - \frac{\partial \hat{\gamma}_b}{\partial l_b}}{l_k - l_b} + \frac{4}{n^2 p} \sum_{k=1}^{\rho} \sum_{c=\rho+1}^{p} \hat{\gamma}_k \frac{\frac{\partial \hat{\gamma}_k}{\partial l_k} - \frac{\partial \hat{\sigma}^2}{\partial l_c}}{l_k - l_c} \\
& + \frac{4}{n^2 p} \sum_{k=1}^{\rho} \sum_{c=\rho+1}^{p} \hat{\sigma}^2 \frac{\frac{\partial \hat{\gamma}_k}{\partial l_k} - \frac{\partial \hat{\sigma}^2}{\partial l_c}}{l_k - l_c} + \frac{4}{n^2 p} \sum_{c \neq d=\rho+1}^{p} \hat{\sigma}^2 \frac{\frac{\partial \hat{\sigma}^2}{\partial l_c} - \frac{\partial \hat{\sigma}^2}{\partial l_d}}{l_c - l_d} \\
& + \frac{4}{n^2 p} \sum_{k \neq b=1}^{\rho} \hat{\gamma}_k \frac{\frac{\partial \hat{\gamma}_k}{\partial l_k} - \frac{\partial \hat{\gamma}_b}{\partial l_k}}{l_k - l_b} + \frac{4}{n^2 p} \sum_{k=1}^{\rho} \sum_{c=\rho+1}^{p} \hat{\gamma}_k \frac{\frac{\partial \hat{\gamma}_k}{\partial l_k} - \frac{\partial \hat{\sigma}^2}{\partial l_k}}{l_k - l_c} \\
& + \frac{4}{n^2 p} \sum_{k=1}^{\rho} \sum_{c=\rho+1}^{p} \hat{\sigma}^2 \frac{\frac{\partial \hat{\gamma}_k}{\partial l_c} - \frac{\partial \hat{\sigma}^2}{\partial l_c}}{l_k - l_c} + \frac{4}{n^2 p} \sum_{c \neq d=\rho+1}^{p} \hat{\sigma}^2 \frac{\frac{\partial \hat{\sigma}^2}{\partial l_c} - \frac{\partial \hat{\sigma}^2}{\partial l_c}}{l_c - l_d} \\
& - \frac{4}{np} \sum_{k=1}^{\rho} \frac{\partial \hat{\gamma}_k}{\partial l_k} - \frac{4}{np} \sum_{k=1}^{p} \frac{\partial \hat{\sigma}^2}{\partial l_k}.
\end{aligned} \tag{4.13}$$

For the second part of the theorem, we see that

$$\begin{aligned}
\mathrm{E}\Big[\Big|F(\hat{\Sigma})\Big|\Big] \leq {} & 1 + \frac{|n-p-1||n-p-2|}{n^2} \mathrm{E}\left[\sup_{p \in \mathbb{N}^*} \max_{1 \leq k \leq p} \left| \frac{\psi_k}{l_k} \right|^2 \right] \\
& + \frac{|n-p-1|p}{n^2} \mathrm{E}\left[\sup_{p \in \mathbb{N}^*} \max_{1 \leq k \leq p} \left| \frac{\psi_k}{l_k} \right|^2 \right] \\
& + 4 \frac{|n-p-1|(p-1)}{n^2} \mathrm{E}\left[\sup_{p \in \mathbb{N}^*} \max_{1 \leq k \leq p} \left| \frac{\psi_k}{l_k} \right|^2 \right]^{\frac{1}{2}} \mathrm{E}\left[\sup_{p \in \mathbb{N}^*} \max_{1 \leq k \neq b \leq p} \left| \frac{\psi_k - \psi_b}{l_k - l_b} \right|^2 \right]^{\frac{1}{2}}
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{2(p-1)(p-2)}{n^2}\,\mathrm{E}\left[\sup_{p\in\mathbb{N}^*}\max_{1\le k\neq b\le p}\left|\frac{\psi_k-\psi_b}{l_k-l_b}\right|^2\right]\\
&+ \frac{2(p-1)(p-2)}{n^2}\,\mathrm{E}\left[\sup_{p\in\mathbb{N}^*}\max_{1\le k\le p}\left|\frac{\psi_k}{l_k}\right|^2\right]^{\frac{1}{2}}\\
&\qquad\cdot\,\mathrm{E}\left[\sup_{p\in\mathbb{N}^*}\max_{1\le k\neq b\neq e\le p}\left|\frac{l_k}{l_k-l_b}\left(\frac{\psi_k-\psi_e}{l_k-l_e}-\frac{\psi_b-\psi_e}{l_b-l_e}\right)\right|^2\right]^{\frac{1}{2}}\\
&- 2\frac{|n-p-1|}{n}\,\mathrm{E}\left[\sup_{p\in\mathbb{N}^*}\max_{1\le k\le p}\left|\frac{\psi_k}{l_k}\right|\right]-\frac{2(p-1)}{n}\,\mathrm{E}\left[\sup_{p\in\mathbb{N}^*}\max_{1\le k\neq b\le p}\left|\frac{\psi_k-\psi_b}{l_k-l_b}\right|\right]
\end{aligned}
$$

and

$$
\begin{aligned}
p\,\mathrm{E}\Big[\Big|G(\hat{\Sigma})\Big|\Big] \le{}& \frac{8p}{n^2}\,\mathrm{E}\left[\sup_{p\in\mathbb{N}^*}\max_{1\le k\le p}\left|\frac{\partial\psi_k}{\partial l_k}\right|^2\right]+\frac{8p}{n^2}\,\mathrm{E}\left[\sup_{p\in\mathbb{N}^*}\max_{1\le k\le p}\left|\psi_k\frac{\partial^2\psi_k}{\partial l_k^2}\right|\right]\\
&+ 8\frac{|n-p-1|p}{n}\,\mathrm{E}\left[\sup_{p\in\mathbb{N}^*}\max_{1\le k\le p}\left|\frac{\psi_k}{l_k}\right|^2\right]^{\frac{1}{2}}\mathrm{E}\left[\sup_{p\in\mathbb{N}^*}\max_{1\le k\le p}\left|\frac{\partial\psi_k}{\partial l_k}\right|^2\right]^{\frac{1}{2}}\\
&+ \frac{8(p-1)p}{n^2}\,\mathrm{E}\left[\sup_{p\in\mathbb{N}^*}\max_{1\le k\le p}\left|\frac{\partial\psi_k}{\partial l_k}\right|^2\right]^{\frac{1}{2}}\mathrm{E}\left[\sup_{p\in\mathbb{N}^*}\max_{1\le k\neq b\le p}\left|\frac{\psi_k-\psi_b}{l_k-l_b}\right|^2\right]^{\frac{1}{2}}\\
&+ \frac{4(p-1)p}{n^2}\,\mathrm{E}\left[\sup_{p\in\mathbb{N}^*}\max_{1\le k\le p}\left|\frac{\psi_k}{l_k}\right|^2\right]^{\frac{1}{2}}\mathrm{E}\left[\sup_{p\in\mathbb{N}^*}\max_{\substack{1\le k\\ \neq b\le p}}\left|\frac{l_k}{l_k-l_b}\left(\frac{\partial\psi_k}{\partial l_k}-\frac{\partial\psi_b}{\partial l_b}\right)\right|^2\right]^{\frac{1}{2}}\\
&+ \frac{4(p-1)p}{n^2}\,\mathrm{E}\left[\sup_{p\in\mathbb{N}^*}\max_{1\le k\le p}\left|\frac{\psi_k}{l_k}\right|^2\right]^{\frac{1}{2}}\mathrm{E}\left[\sup_{p\in\mathbb{N}^*}\max_{\substack{1\le k\\ \neq b\le p}}\left|\frac{l_k}{l_k-l_b}\left(\frac{\partial\psi_k}{\partial l_k}-\frac{\partial\psi_b}{\partial l_k}\right)\right|^2\right]^{\frac{1}{2}}\\
&- \frac{4p}{n}\,\mathrm{E}\left[\sup_{p\in\mathbb{N}^*}\max_{1\le k\le p}\left|\frac{\partial\psi_k}{\partial l_k}\right|\right].
\end{aligned}
$$

Again, one can proceed like in the weak case to see that if $\hat{\Gamma}$ satisfies its strong regularity conditions and $\hat{\Sigma}\in\tilde{V}_p(\hat{\Gamma})$, we get $\lim_{n\to\infty}\mathrm{E}\Big[\Big|F(\hat{\Sigma})\Big|\Big]<\infty$ and $\lim_{n\to\infty}p\,\mathrm{E}\Big[\Big|G(\hat{\Sigma})\Big|\Big]<\infty$ as $\lim_{n\to\infty}\frac{p_n}{n}\in(0,1)$, as desired. □

Proof of Proposition 8. First note any element of $C_c^\infty(H_p;\mathbb{R})$, the space of smooth, compactly supported functions from H_p to $\mathbb{R}$, satisfies the weak regularity conditions of Definition 2 for any weak $\hat{\Gamma}$. Now, if $\tilde{\Sigma}$ is a minimum over $V_p(\hat{\Gamma})$, then for any $\eta\in C_c^\infty(H_+^p;\mathbb{R})$ and any $t\in\mathbb{R}$, $\tilde{\sigma}^2+t\eta$ satisfies the $\hat{\Gamma}$-weak

regularity conditions too and $\epsilon \to \mathrm{E}\Big[F(\hat{\Gamma} + [\tilde{\sigma}^2 + t\eta]I)\Big]$ is smooth over $\mathbb{R}$ with a minimum at $t = 0$. But we find that the first variation satisfies

$$\begin{aligned}
\frac{\partial}{\partial t}\,\mathrm{E}\Big[F(\hat{\Gamma} + [\tilde{\sigma}^2 + t\eta]I)\Big]\,\Big|_{t=0} &= \mathrm{E}\Bigg[\eta \cdot \Bigg(2\frac{(n-p-1)(n-p-2)}{n^2 p}\sum_{k=1}^{\hat{\rho}}\frac{\hat{\gamma}_k}{l_k^2} \\
&+2\frac{(n-p-1)(n-p-2)}{n^2 p}\sum_{c=1}^{p}\frac{\tilde{\sigma}^2}{l_c^2} - 2\frac{n-p-1}{n^2 p}\sum_{c=1}^{p}\frac{1}{l_c}\sum_{k=1}^{\hat{\rho}}\frac{\hat{\gamma}_k}{l_k} \\
&-2\frac{n-p-1}{n^2 p}\sum_{c=1}^{p}\frac{1}{l_c}\sum_{c=1}^{p}\frac{\tilde{\sigma}^2}{l_c} + 4\frac{n-p-1}{n^2 p}\sum_{k=1}^{\hat{\rho}}\sum_{c=\hat{\rho}+1}^{p}\frac{1}{l_c}\frac{\hat{\gamma}_k}{l_k - l_c} \\
&-\frac{6}{n^2 p}\sum_{k\neq b}^{\hat{\rho}}\sum_{c=\hat{\rho}+1}^{p}\frac{1}{l_k - l_c}\frac{\hat{\gamma}_b}{l_b - l_c} + \frac{6}{n^2 p}\sum_{k=1}^{\hat{\rho}}\sum_{c\neq d=\hat{\rho}+1}^{p}\frac{1}{l_k - l_c}\frac{\hat{\gamma}_k}{l_k - l_d} \\
&+\frac{6}{n^2 p}\sum_{k\neq b}^{\hat{\rho}}\sum_{c=\hat{\rho}+1}^{p}\frac{\hat{\gamma}_k - \hat{\gamma}_b}{l_k - l_b}\frac{1}{l_k - l_c} - 2\frac{n-p-1}{np}\sum_{c=1}^{p}\frac{1}{l_k}\Bigg)\Bigg] \\
&= \mathrm{E}\big[\eta \cdot F_1\big[l, \hat{\rho}, \hat{\gamma}, \tilde{\sigma}^2\big]\big] \\
&= \int_{H_p} \eta(l_1, ..., l_p) F_1\big[l, \hat{\rho}, \hat{\gamma}, \tilde{\sigma}^2\big] \cdot f_{l_1,...,l_p}(l_1, ..., l_p) \prod_{i=1}^{p} dl_i, \qquad (4.14)
\end{aligned}$$

where $f_{l_1,...,l_p}(l_1, ..., l_p)$ stands for the p.d.f. of $l_1 > ... > l_p$. Now, if this equals zero for all $\eta \in C_c^\infty(H_+^p; \mathbb{R})$, by the fundamental lemma of calculus of variations (see, say, Giaquinta and Hildebrandt [1996] ch. 2.2) we obtain $F_1\big[l, \hat{\rho}, \hat{\gamma}, \tilde{\sigma}^2\big] \cdot f_{l_1,...,l_p}(l_1, ..., l_p) \equiv 0$, that is, $F_1\big[l, \hat{\rho}, \hat{\gamma}, \tilde{\sigma}^2\big] \equiv 0$. This implies $\tilde{\sigma}^2 = A/B$.

For the second statement, notice that by construction, the space of $\hat{\Gamma}$-weak noise estimators is convex; let $\hat{\sigma}^2$ be some arbitrary element. Define $H : [0, 1] \to \mathbb{R}$ to be the smooth function

$$H(t) = \mathrm{E}\Big[F(\hat{\Gamma} + [\tilde{\sigma}^2 + t(\hat{\sigma}^2 - \tilde{\sigma}^2)]I)\Big].$$

Notice that, for F_1 as in eq. (4.14),

$$H'(0) = \mathrm{E}\big[(\hat{\sigma}^2 - \tilde{\sigma}^2) \cdot F_1\big[l, \hat{\rho}, \hat{\gamma}, \tilde{\sigma}^2\big]\big] = 0,$$

since $\tilde{\sigma}^2 = A/B$. Moreover,

$$\begin{aligned} H''(t) &= \frac{\partial^2}{\partial t^2} \mathrm{E}\Big[F(\hat{\Gamma}_r + [\tilde{\sigma}^2 + t(\hat{\sigma}^2 - \tilde{\sigma}^2)]I)\Big] \\ &= 2\frac{n-p-1}{n^2 p} \mathrm{E}\left[\left(\hat{\sigma}^2 - \tilde{\sigma}^2\right)^2 \left((n-p-2)\sum_{c=1}^{p} \frac{1}{l_c^2} - \left(\sum_{c=1}^{p} \frac{1}{l_c}\right)^2\right)\right] \qquad (4.15) \\ &\geq 2\frac{(n-p-1)(n-2p-2)}{n^2 p^2} \mathrm{E}\left[\left(\hat{\sigma}^2 - \tilde{\sigma}^2\right)^2 \left(\sum_{c=1}^{p} \frac{1}{l_c}\right)^2\right] \quad \begin{pmatrix}\text{Jensen's} \\ \text{inequality}\end{pmatrix} \\ &\geq 0, \end{aligned}$$

for $n \geq 2p+2$. Therefore, by integration by parts

$$\begin{aligned} &\mathrm{E}\Big[F(\hat{\Gamma} + \hat{\sigma}^2 I)\Big] - \mathrm{E}\Big[F(\hat{\Gamma} + \tilde{\sigma}^2 I)\Big] \\ &\qquad = H(1) - H(0) = \int_0^1 (1-t)H''(t)\, dt \ \geq 0. \end{aligned}$$

Since this is true for any $\hat{\Gamma}$-weak noise estimator $\hat{\sigma}^2$, we conclude that $\tilde{\Sigma}$ is a minimum over $V_p(\hat{\Gamma})$, as desired. □

4.6.2 Proofs for Section 4.3

Proof of Lemma 4. To simplify notation in what follows, define $c_\pm = [1 \pm \sqrt{c}]^2$. In the proof of Theorem 2.3 in Nadler [2008], p. 2807, it is remarked that for $\sigma^2 = 1$ and $\rho = 1$, the empirical distribution of $l_2, ..., l_p$ converges a.s. to a Marčenko-Pastur distribution with parameter c. That is, for the truncated empirical spectral measure $d\mu_p = \frac{1}{p-\rho}\sum_{c=\rho+1}^{p} d\delta_{l_i}$ (where the δ are Dirac measures) we have weak convergence $d\mu_p \Rightarrow d\mu_{\mathrm{MP(c)}}$ a.s. where

$$d\mu_{\mathrm{MP(c)}} = \frac{\sqrt{(c_+ - t)(t - c_-)}}{2\pi c t} \mathbb{1}\Big[c_- \leq t \leq c_+\Big]\, dt.$$

As noted by the author, the argument carries on for $\rho \neq 1$, and if $\sigma \neq 1$ we can apply the argument to $l_{\rho+1}/\sigma^2, ..., l_p/\sigma^2$ to obtain $d\mu_p \Rightarrow d\mu_{\sigma^2\mathrm{MP}(c)}$ a.s., where

$$d\mu_{\sigma^2\mathrm{MP}(c)} = \frac{\sqrt{(\sigma^2 c_+ - t)(t - \sigma^2 c_-)}}{2\pi c \sigma^2 t} \mathbb{1}\Big[\sigma^2 c_- \leq t \leq \sigma^2 c_+\Big]\, dt.$$

Part (i) Applying the results of Baik and Silverstein [2006], Theorem 1.1 to l_k/σ^2 and $l_{\rho+1}/\sigma^2$ we obtain:

$$l_k \xrightarrow[n\to\infty]{\text{a.s.}} \frac{(\gamma_k+\sigma^2)(\gamma_k+c\sigma^2)}{\gamma_k}, \qquad\qquad l_{\rho+1} \xrightarrow[n\to\infty]{\text{a.s.}} c_+\sigma^2.$$

We will write $\bar{l}_k = (\gamma_k+\sigma^2)(\gamma_k+c\sigma^2)/\gamma_k$ to simplify notation. Let the underlying sample space be denoted Ω. Since $\gamma_\rho > \sqrt{c}\sigma^2$, we have $\bar{l}_k - c_+\sigma^2 = M$ for some $M > 0$. Therefore, for almost all $\omega \in \Omega$, there exists an $N_1(\omega)$ such that $\forall n > N(\omega)$, $l_k^p(\omega) - l_p^p(\omega) > ... > l_k^p(\omega) - l_{\rho+1}^p(\omega) > M/2$ and $l_p^p(\omega) < ... < l_{\rho+1}^p(\omega) < c_+\sigma^2 + M/2$. Moreover, for any $\epsilon > 0$, there must be an $N_2(\omega)$ such that for all $n > N_2(\omega,\epsilon)$, $|\bar{l}_k - l_k^p(\omega)| < \epsilon$. Notice that we can write, for any $n > N_1(\omega) \vee N_2(\omega,\epsilon)$,

$$\begin{aligned}\frac{1}{p-\rho}\sum_{c=\rho+1}^{p}\frac{l_c^p(\omega)}{l_k^p(\omega)-l_c^p(\omega)} =& \frac{1}{p-\rho}\sum_{c=\rho+1}^{p}\frac{[\bar{l}_k - l_k^p(\omega)]l_c^p(\omega)}{[l_k^p(\omega)-l_c^p(\omega)][\bar{l}_k - l_c^p(\omega)]} \\ &+ \int_{(0,c_+\sigma^2+\frac{M}{2})}\frac{t}{\bar{l}_k - t}d\mu_p(t,\omega),\end{aligned}$$

and

$$\left|\frac{1}{p-\rho}\sum_{c=\rho+1}^{p}\frac{[\bar{l}_k - l_k^p(\omega)]l_c^p(\omega)}{[l_k^p(\omega)-l_c^p(\omega)][\bar{l}_k - l_c^p(\omega)]}\right| < \left[\frac{4}{M^2}c_+\sigma^2 + \frac{2}{M}\right]\epsilon. \tag{4.16}$$

But $0 < t/(\bar{l}_k - t) < 1 + 2c_+\sigma^2/M$ on $t \in (0, c_+\sigma^2 + \frac{M}{2})$, and it is certainly continuous. Therefore, by the portmanteau theorem of weak convergence of measures,

$$\begin{aligned}\lim_{n\to\infty}\int_{(0,c_+\sigma^2+\frac{M}{2})}\frac{t}{\bar{l}_k - t}d\mu_p(t,\omega) &= \int_{(0,c_+\sigma^2+\frac{M}{2})}\frac{t}{\bar{l}_k - t}d\mu_{\sigma^2\text{MP}(c)}(t) \\ &= \int_{c_-\sigma^2}^{c^+\sigma^2}\frac{t}{\bar{l}_k - t}\frac{\sqrt{(\sigma^2c_+ - t)(t-\sigma^2c_-)}}{2\pi c\sigma^2 t}dt \\ &= \frac{2\sigma^2}{\bar{l}_k - (c+1)\sigma^2 + \sqrt{[\bar{l}_k-(c+1)\sigma^2]^2 - 4c\sigma^4}}.\end{aligned} \tag{4.17}$$

But by definition, $\bar{l}_k = (\gamma_k+\sigma^2)(\gamma_k+c\sigma^2)/\gamma_k$ which can be rewritten as a quadratic equation in γ_k,

$$\gamma_k^2 - [(\bar{l}_k - (c+1)\sigma^2]\gamma_k + c\sigma^4 = 0.$$

The roots are

$$\frac{1}{2}[\bar{l}_k - (c+1)\sigma^2] \pm \frac{1}{2}\sqrt{[\bar{l}_k - (c+1)\sigma^2]^2 - 4c\sigma^4},$$

and notice that twice the negative root satisfies

$$\begin{aligned}
&[\bar{l}_k - (c+1)\sigma^2] - \sqrt{[\bar{l}_k - (c+1)\sigma^2]^2 - 4c\sigma^4} \\
&\quad = [\bar{l}_k - (c+1)\sigma^2] \\
&\qquad - \sqrt{\left([\bar{l}_k - (c+1)\sigma^2] - 2\sqrt{c}\sigma^2\right)\left([\bar{l}_k - (c+1)\sigma^2] + 2\sqrt{c}\sigma^2\right)} \\
&\quad \leq [\bar{l}_k - (c+1)\sigma^2] - [\bar{l}_k - (c+1)\sigma^2] + 2\sqrt{c}\sigma^2 \\
&\quad = 2\sqrt{c}\sigma^2.
\end{aligned}$$

Therefore, γ_k cannot equal the negative root, because it would imply $\gamma_k \leq \sqrt{c}\sigma^2$, a contradiction. So γ_k equals the positive root, which, plugged in eq. (4.17), yields

$$\lim_{n\to\infty} \int_{(0, c_+\sigma^2 + \frac{M}{2})} \frac{t}{\bar{l}_k - t} d\mu_p(t, \omega) = \frac{\sigma^2}{\gamma_k}.$$

Hence, for some $N_3(\omega, \epsilon)$, we have for all $n > N_3(\omega, \epsilon)$

$$\left| \int_{(0, c_+\sigma^2 + \frac{M}{2})} \frac{t}{\bar{l}_k - t} d\mu_p(t, \omega) - \frac{\sigma^2}{\gamma_k} \right| < \epsilon.$$

Therefore, from eq. (4.16), we obtain that for $n > N_1(\omega) \vee N_2(\omega, \epsilon) \vee N_3(\omega, \epsilon)$

$$\left| \frac{1}{p-\rho} \sum_{c=\rho+1}^{p} \frac{l_c^p(\omega)}{l_k^p(\omega) - l_c^p(\omega)} - \frac{\sigma^2}{\gamma_k} \right| < \left[\frac{4}{M^2} c_+ \sigma^2 + \frac{2}{M} + 1 \right] \epsilon.$$

Since $\epsilon > 0$ is arbitrary, we conclude that for almost all $\omega \in \Omega$,

$$\lim_{n\to\infty} \frac{1}{p-\rho} \sum_{c=\rho+1}^{p} \frac{l_c^p(\omega)}{l_k^p(\omega) - l_c^p(\omega)} = \frac{\sigma^2}{\gamma_k},$$

as desired.

Part (ii) The proof is similar in spirit to the previous one, but simpler. Applying the results of Baik and Silverstein [2006], Theorem 1.1 to l_p/σ^2 we obtain:

$$l_p \xrightarrow[n\to\infty]{\text{a.s.}} c_-\sigma^2.$$

Therefore, for almost all $\omega \in \Omega$, there is a $N(\omega)$ such that $\forall n > N(\omega)$, $l^p_{\rho+1}(\omega) > ... > l^p_p(\omega) > c_-\sigma^2/2$. Hence, for $n > N(\omega)$,

$$\frac{1}{p-\rho}\sum_{c=\rho+1}^{p}\frac{1}{l_c^{pm}(\omega)} = \int_{\left(\frac{c_-\sigma^2}{2},\infty\right)}\frac{1}{t^m}d\mu_p(t,\omega),$$

Certainly, $0 < 1/t^m < \left(\frac{2}{c_-\sigma^2}\right)^m$ for $t \in \left(\frac{c_-\sigma^2}{2},\infty\right)$, and $1/t^m$ is continuous there. Thus, by the portmanteau theorem of weak convergence of measures, we have for almost all ω

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{p-\rho}\sum_{c=\rho+1}^{p}\frac{1}{l_c^{pm}(\omega)} &= \int_{\left(\frac{c_-\sigma^2}{2},\infty\right)}\frac{1}{t^m}d\mu_{\sigma^2\mathrm{MP}(c)}(t)\\ &= \int_{c_-\sigma^2}^{c^+\sigma^2}\frac{1}{t^m}\frac{\sqrt{(\sigma^2c_+-t)(t-\sigma^2c_-)}}{2\pi c\sigma^2 t}dt\\ &= \frac{1}{(1-c)^{2m-1}}\frac{1}{\sigma^{2m}},\end{aligned}$$

which concludes the proof. □

Proof of Theorem 7. Recall the definition of $\tilde{\sigma}^2$ as A/B from proposition 8.

Part (i) It follows easily from the Lemma 4 that

$$A \xrightarrow[n\to\infty]{\text{a.s.}} \frac{1}{\sigma^2} \qquad \text{and} \qquad B \xrightarrow[n\to\infty]{\text{a.s.}} \frac{1}{\sigma^4}. \tag{4.18}$$

The result then follows immediately.

Part (ii) Denote by $\bar{\gamma}_k$ the a.s. finite limit $\lim_{n\to\infty}\hat{\gamma}_k$. By writing $A = \frac{n-p-1}{np}\sum_{c=1}^{p}\frac{1}{l_c} + E_n$,

we find

$$
\begin{aligned}
& n(\tilde{\sigma}^2 - \sigma^2) \\
& = n\sigma^2 \left[\frac{\frac{n-p-1}{np} \sum_{c=1}^{p} \frac{\sigma^2}{l_c}}{\frac{(n-p-1)(n-p-2)}{n^2 p} \sum_{c=1}^{p} \frac{\sigma^4}{l_c^2} - \frac{(n-p-1)p}{n^2} \left(\frac{1}{p} \sum_{c=1}^{p} \frac{\sigma^2}{l_c} \right)^2} - 1 \right] + \frac{nE_n}{B} \\
& = \frac{\sigma^2}{\sigma^4 B} \left[n \left(\frac{n-p-1}{np} \sum_{c=1}^{p} \frac{\sigma^2}{l_c} - 1 \right) \right. \\
& \qquad + n \left(\frac{1}{1 - \frac{p-\hat{\rho}}{n}} - \frac{(n-p-1)(n-p-2)}{n^2 p} \sum_{c=1}^{p} \frac{\sigma^4}{l_c^2} \right) \\
& \qquad \left. + n \left(\frac{(n-p-1)p}{n^2} \left(\frac{1}{p} \sum_{c=1}^{p} \frac{\sigma^2}{l_c} \right)^2 - \frac{\frac{p-\hat{\rho}}{n}}{1 - \frac{p-\hat{\rho}}{n}} \right) \right] + \frac{nE_n}{B}. \qquad (4.19)
\end{aligned}
$$

First consider nE_n. We know the asymptotic behavior of all terms except $\frac{3}{np} \sum_{k=1}^{\hat{\rho}} \sum_{c \neq d = r+1}^{p} \frac{1}{l_k - l_c} \frac{\hat{\gamma}_k}{l_k - l_d}$, which we can crudely bound as

$$
0 < \frac{3}{np} \sum_{k=1}^{\hat{\rho}} \sum_{c \neq d = \hat{\rho}+1}^{p} \frac{1}{l_k - l_c} \frac{\hat{\gamma}_k}{l_k - l_d} < \frac{3}{np} \sum_{k=1}^{\hat{\rho}} \hat{\gamma}_k \left(\sum_{c=\hat{\rho}+1}^{p} \frac{1}{l_k - l_c} \right)^2.
$$

Therefore, we obtain that

$$
\begin{aligned}
& \frac{(1-c)^2}{c} \sum_{k=1}^{\hat{\rho}} \frac{\gamma_k^2 \bar{\gamma}_k}{(\gamma_k + \sigma^2)^2 (\gamma_k + c\sigma^2)^2} + \frac{1}{\sigma^2} \sum_{k=1}^{\hat{\rho}} \frac{\gamma_k \bar{\gamma}_k}{(\gamma_k + \sigma^2)(\gamma_k + c\sigma^2)} \\
& \qquad - 2 \frac{1}{\sigma^2} \sum_{k=1}^{\hat{\rho}} \frac{\gamma_k \bar{\gamma}_k}{(\gamma_k + c\sigma^2)^2} \\
& \qquad \geq \lim_{n \to \infty} nE_n \geq \\
& \frac{(1-c)^2}{c} \sum_{k=1}^{\hat{\rho}} \frac{\gamma_k^2 \bar{\gamma}_k}{(\gamma_k + \sigma^2)^2 (\gamma_k + c\sigma^2)^2} + \frac{1}{\sigma^2} \sum_{k=1}^{\hat{\rho}} \frac{\gamma_k \bar{\gamma}_k}{(\gamma_k + \sigma^2)(\gamma_k + c\sigma^2)} \\
& \qquad - 2 \frac{1}{\sigma^2} \sum_{k=1}^{\hat{\rho}} \frac{\gamma_k \bar{\gamma}_k}{(\gamma_k + c\sigma^2)^2} - 3c \sum_{k=1}^{\hat{\rho}} \frac{\bar{\gamma}_k}{(\gamma_k + c\sigma^2)^2} \qquad (4.20)
\end{aligned}
$$

almost surely, using the results of Lemma 4 and Baik and Silverstein [2006], The-

orem 1.1. Now notice that

$$
\begin{aligned}
& n\left(\frac{n-p-1}{np}\sum_{c=1}^{p}\frac{\sigma^2}{l_c}-1\right) \\
&= n\left[\frac{n-p-1}{np}\sum_{c=1}^{p}\frac{\sigma^2}{l_c}-\frac{n-p-1}{n(1-\frac{p-\rho}{n})}+\frac{n-p-1}{n(1-\frac{p-\rho}{n})}-1\right] \\
&= \left(1-\frac{p}{n}-\frac{1}{n}\right)n\left[\frac{1}{p}\sum_{c=1}^{p}\frac{\sigma^2}{l_c}-\frac{1}{1-\frac{p-\rho}{n}}\right]-\frac{n(\rho+1)}{n-p}
\end{aligned}
\tag{4.21}
$$

and

$$
\begin{aligned}
& n\left(\frac{1}{1-\frac{p}{n}}-\frac{(n-p-1)(n-p-2)}{n^2p}\sum_{c=1}^{p}\frac{\sigma^4}{l_c^2}\right) \\
&= n\left(\frac{n}{n-p}-\frac{(n-p-1)(n-p-2)}{n^2(1-\frac{p-\rho}{n})^3}\right. \\
&\qquad \left.+\frac{(n-p-1)(n-p-2)}{n^2(1-\frac{p-\rho}{n})^3}-\frac{(n-p-1)(n-p-2)}{n^2p}\sum_{c=1}^{p}\frac{\sigma^4}{l_c^2}\right) \\
&= \left(1-\frac{p}{n}-\frac{1}{n}\right)\left(1-\frac{p}{n}-\frac{2}{n}\right)n\left[\frac{1}{(1-\frac{p-\rho}{n})^3}-\frac{1}{p}\sum_{c=1}^{p}\frac{\sigma^4}{l_c^2}\right] \\
&\qquad +\frac{\rho n^2}{(n-p)(n-p+\rho)}+\frac{(2r+3)n^2}{(n-p+\rho)^2}-\frac{(\rho^2+3r+2)n^2}{(n-p+\rho)^3}.
\end{aligned}
\tag{4.22}
$$

Moreover,

$$
\begin{aligned}
& n\left(\frac{(n-p-1)p}{n^2}\left(\frac{1}{p}\sum_{c=1}^{p}\frac{\sigma^2}{l_c}\right)^2-\frac{\frac{p-\rho}{n}}{1-\frac{p-\rho}{n}}\right) \\
&= n\left(\frac{(n-p-1)p}{n^2}\left(\frac{1}{p}\sum_{c=1}^{p}\frac{\sigma^2}{l_c}\right)^2-\frac{(n-p-1)p}{n^2}\frac{1}{(1-\frac{p-\rho}{n})^2}\right. \\
&\qquad \left.+\frac{(n-p-1)p}{n^2}\frac{1}{(1-\frac{p-\rho}{n})^2}-\frac{\frac{p-\rho}{n}}{1-\frac{p-\rho}{n}}\right) \\
&= \left(1-\frac{p}{n}-\frac{1}{n}\right)\frac{p}{n}n\left[\left(\frac{1}{p}\sum_{c=1}^{p}\frac{\sigma^2}{l_c}\right)^2-\frac{1}{(1-\frac{p-\rho}{n})^2}\right] \\
&\qquad +\frac{\rho n}{n-p+\rho}-\frac{(\rho+1)pn}{(n-p+\rho)^2} \\
&= \left(1-\frac{p}{n}-\frac{1}{n}\right)\frac{p}{n}\left[\frac{1}{p}\sum_{c=1}^{p}\frac{\sigma^2}{l_c}+\frac{1}{1-\frac{p-\rho}{n}}\right]n\left[\frac{1}{p}\sum_{c=1}^{p}\frac{\sigma^2}{l_c}-\frac{1}{1-\frac{p-\rho}{n}}\right]
\end{aligned}
$$

$$+\frac{\rho n}{n-p+\rho}-\frac{(\rho+1)pn}{(n-p+\rho)^2}. \tag{4.23}$$

Now, divide the sample covariance matrix as

$$S_n=\sigma^2\left(\begin{array}{c|c} S_n^{11} & S_n^{12} \\ \hline S_n^{21} & S_n^{22} \end{array}\right)$$

with S_n^{11} $\rho\times\rho$. Then S_n^{22} has a $\mathrm{W}_{p-\rho}(n,I)$ distribution – let $\mu_1>...>\mu_{p-\rho}$ be its eigenvalues and notice that by Cauchy's interlacing theorem, $l_i>\sigma^2\mu_i>l_{i+\rho}$ for all $i=1,...,p-\rho$. Therefore, we have

$$\begin{aligned}
&\frac{p-\rho}{p}n\left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho}\frac{1}{\mu_c}-\frac{1}{1-\frac{p-\rho}{n}}\right]+\rho\frac{n}{p}\frac{\sigma^2}{l_p}-\frac{\rho n^2}{(n-p+\rho)p} \\
&\qquad\geq\ n\left[\frac{1}{p}\sum_{c=1}^{p}\frac{\sigma^2}{l_c}-\frac{1}{1-\frac{p-\rho}{n}}\right]\geq \\
&\frac{p-\rho}{p}n\left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho}\frac{1}{\mu_c}-\frac{1}{1-\frac{p-\rho}{n}}\right]+\frac{n}{p}\sum_{c=1}^{\rho}\frac{\sigma^2}{l_c}-\frac{\rho n^2}{(n-p+\rho)p}
\end{aligned} \tag{4.24}$$

and

$$\begin{aligned}
&\frac{p-\rho}{p}n\left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho}\frac{1}{\mu_c^2}-\frac{1}{(1-\frac{p-\rho}{n})^3}\right]+\rho\frac{n}{p}\frac{\sigma^4}{l_p^2}-\frac{\rho n^4}{(n-p+\rho)^3p} \\
&\qquad\geq\ n\left[\frac{1}{p}\sum_{c=1}^{p}\frac{\sigma^4}{l_c^2}-\frac{1}{(1-\frac{p-\rho}{n})^3}\right]\geq \\
&\frac{p-\rho}{p}n\left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho}\frac{1}{\mu_c^2}-\frac{1}{(1-\frac{p-\rho}{n})^3}\right]+\frac{n}{p}\sum_{c=1}^{\rho}\frac{\sigma^4}{l_c^2}-\frac{\rho n^4}{(n-p+\rho)^3p}.
\end{aligned} \tag{4.25}$$

Consequently, let us study the quantities

$$n\left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho}\frac{1}{\mu_c}-\frac{1}{1-\frac{p-\rho}{n}}\right]\quad\text{and}\quad n\left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho}\frac{1}{\mu_c^2}-\frac{1}{(1-\frac{p-\rho}{n})^3}\right].$$

We use Theorem 1.1 in Bai and Silverstein [2004]. First notice that, in our white Wishart case, what they write $F^{c,H}$ is the c.d.f. of a Marčenko-Pastur distribution

with parameter c. Let $c_n = (p-\rho)/n$ - then

$$\int \frac{1}{x} dG_n(x) = n\left[\int \frac{1}{x} dF^{S_n^{22}}(x) - \int \frac{1}{x} dF^{c_n, F^I}(x)\right]$$

$$= n\left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho}\frac{1}{\mu_c} - \int_{[1-\sqrt{c_n}]^2}^{[1+\sqrt{c_n}]^2} \frac{\sqrt{([1+\sqrt{c_n}]^2 - t)(t - [1-\sqrt{c_n}]^2)}}{2\pi c_n t^2} dt\right]$$

$$= n\left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho}\frac{1}{\mu_c} - \frac{1}{(1-\frac{p-\rho}{n})^2}\right]$$

and

$$\int \frac{1}{x^2} dG_n(x) = n\left[\int \frac{1}{x^2} dF^{S^{22}}(x) - \int \frac{1}{x^2} dF^{c_n, F^I}(x)\right]$$

$$= n\left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho}\frac{1}{\mu_c^2} - \int_{[1-\sqrt{c_n}]^2}^{[1+\sqrt{c_n}]^2} \frac{\sqrt{([1+\sqrt{c_n}]^2 - t)(t - [1-\sqrt{c_n}]^2)}}{2\pi c_n t^3} dt\right]$$

$$= n\left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho}\frac{1}{\mu_c^2} - \frac{1}{(1-\frac{p-\rho}{n})^3}\right].$$

But according to the theorem, as $p_n/n \to c \in (0,1)$

$$\left(\int \frac{1}{x} dG_n(x), \int \frac{1}{x^2} dG_n(x)\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathrm{N}_2\left(\tilde{\mu}, \tilde{\Sigma}\right)$$

where the components of $\tilde{\mu}$ and $\tilde{\Sigma}$ are given by eq. (1.6) and (1.7) in the theorem statement (Bai and Silverstein [2004] p. 558). To compute these, we follow the arguments of Section 5 from the same paper. According to eq. (5.13) from p. 598,

we find

$$\tilde{\mu}_1 = \frac{1}{4}\left[\frac{1}{[1-\sqrt{c}]^2} + \frac{1}{[1+\sqrt{c}]^2}\right] - \frac{1}{2\pi}\int_{[1-\sqrt{c}]^2}^{[1+\sqrt{c}]^2} \frac{dt}{t\sqrt{4c-(t-1-c)^2}},$$
$$\tilde{\mu}_2 = \frac{1}{4}\left[\frac{1}{[1-\sqrt{c}]^4} + \frac{1}{[1+\sqrt{c}]^4}\right] - \frac{1}{2\pi}\int_{[1-\sqrt{c}]^2}^{[1+\sqrt{c}]^2} \frac{dt}{t^2\sqrt{4c-(t-1-c)^2}},$$

and the integrals give, after a Poisson substitution $t = 1 + c - 2\sqrt{c}\cos\theta$,

$$\begin{aligned}
\frac{1}{2\pi}\int_{[1-\sqrt{c}]^2}^{[1+\sqrt{c}]^2} \frac{dt}{t\sqrt{4c-(t-1-c)^2}} &= \frac{1}{2\pi}\int_0^\pi \frac{d\theta}{1+c-2\sqrt{c}\cos\theta} \\
&= \frac{1}{2\pi}\left[\frac{2}{1-c}\arctan\left(\frac{1+\sqrt{c}}{1-\sqrt{c}}\tan(\theta/2)\right)\right]_0^\pi \\
&= \frac{1}{2\pi}\left[\frac{2}{1-c}\frac{\pi}{2} - 0\right] = \frac{1}{2(1-c)}, \\
\frac{1}{2\pi}\int_{[1-\sqrt{c}]^2}^{[1+\sqrt{c}]^2} \frac{dt}{t^2\sqrt{4c-(t-1-c)^2}} &= \frac{1}{2\pi}\int_0^\pi \frac{d\theta}{(1+c-2\sqrt{c}\cos\theta)^2} \\
&= \frac{1}{2\pi}\left[\frac{2(1+c)}{(1-c)^3}\arctan\left(\frac{1+\sqrt{c}}{1-\sqrt{c}}\tan(\theta/2)\right)\right. \\
&\qquad \left. + \frac{1}{(1-c)^2}\frac{2\sqrt{c}\sin\theta}{1+c-2\sqrt{c}\cos\theta}\right]_0^\pi \\
&= \frac{1}{2\pi}\left[\frac{2(1+c)}{(1-c)^3}\frac{\pi}{2} + 0 - 0 - 0\right] = \frac{1+c}{2(1-c)^3}.
\end{aligned}$$

Therefore, we obtain

$$\tilde{\mu}_1 = \frac{1+c}{2[1-\sqrt{c}]^2[1+\sqrt{c}]^2} - \frac{1}{2(1-c)} = \frac{c}{(1-c)^2},$$
$$\tilde{\mu}_2 = \frac{1+6c+c^2}{2[1-\sqrt{c}]^4[1+\sqrt{c}]^4} - \frac{1+c}{2(1-c)^3} = \frac{c(c+3)}{(1-c)^4}.$$

For the variances, according to (1.16), p. 564, we can write

$$\tilde{\Sigma}_{11} = -\frac{1}{2\pi^2}\oint_{C_1}\oint_{C_2} \frac{dm_1 dm_2}{z(m_1)z(m_2)(m_1-m_2)^2},$$
$$\tilde{\Sigma}_{12} = -\frac{1}{2\pi^2}\oint_{C_1}\oint_{C_2} \frac{dm_1 dm_2}{z(m_1)^2 z(m_2)(m_1-m_2)^2},$$
$$\tilde{\Sigma}_{22} = -\frac{1}{2\pi^2}\oint_{C_1}\oint_{C_2} \frac{dm_1 dm_2}{z(m_1)^2 z(m_2)^2(m_1-m_2)^2},$$

where C_1, C_2 are contours that can be chosen counterclockwise, nonintersecting and enclosing $1/(c-1)$ (cf. p. 598), and where $z(m)$ stands for the inverse Stieltjes

transform of the complimentary Marčenko-Pastur distribution, which has closed form

$$z(m) = -\frac{1}{m} + \frac{c}{1+m}.$$

We first find:

$$\begin{aligned}
\oint_{C_1} \frac{dm_1}{z(m_1)(m_1-m_2)^2} &= -\frac{1}{1-c}\oint_{C_1} \frac{m_1(m_1+1)}{(m_1-m_2)^2}\frac{dm_1}{m_1-1/(c-1)} \\
&= -\frac{2\pi c i}{(1-c)^3}\frac{1}{[m_2-1/(c-1)]^2}, \\
\oint_{C_1} \frac{dm_1}{z(m_1)^2(m_1-m_2)^2} &= \frac{1}{(1-c)^2}\oint_{C_1} \frac{m_1^2(m_1+1)^2}{(m_1-m_2)^2}\frac{dm_1}{[m_1-1/(c-1)]^2} \\
&= -4\pi i\frac{c}{(1-c)^5}\frac{1}{[m_2-1/(c-1)]^2}.
\end{aligned}$$

But then,

$$\begin{aligned}
\oint_{C_2} \frac{dm_2}{z(m_2)[m_2-1/(c-1)]^2} &= -\frac{1}{1-c}\oint_{C_2} \frac{m_2(m_2+1)dm_2}{[m_2-1/(c-1)]^3} \\
&= -2\pi i\frac{2}{2!(1-c)} = -2\pi i\frac{1}{1-c} \\
\oint_{C_2} \frac{dm_2}{z(m_2)^2[m_2-1/(c-1)]^2} &= \frac{1}{(1-c)^2}\oint_{C_2} \frac{m_2^2(m_2+1)^2dm_2}{[m_2-1/(c-1)]^4} \\
&= 2\pi i\frac{12(1-2/(1-c))}{3!(1-c)^2} = -4\pi i\frac{1+c}{(1-c)^3}.
\end{aligned}$$

Therefore,

$$\tilde{\Sigma}_{11} = \frac{2c}{(1-c)^4}, \quad \tilde{\Sigma}_{12} = \frac{4c}{(1-c)^6} \quad \text{and} \quad \tilde{\Sigma}_{22} = \frac{8c(1+c)}{(1-c)^8}.$$

In summary,

$$\begin{aligned}
&\left(\int \frac{1}{x}dG_n(x), \int \frac{1}{x^2}dG_n(x)\right) \\
&\qquad \xrightarrow[n\to\infty]{\mathcal{D}} \mathrm{N}_2\left(\begin{bmatrix} \frac{c}{(1-c)^2} \\ \frac{c(c+3)}{(1-c)^4} \end{bmatrix}, \begin{bmatrix} \frac{2c}{(1-c)^4} & \frac{4c}{(1-c)^6} \\ \frac{4c}{(1-c)^6} & \frac{8c(1+c)}{(1-c)^8} \end{bmatrix}\right).
\end{aligned} \tag{4.26}$$

Therefore, going back to eq. (4.19) and ineq. (4.21), (4.22), (4.23), (4.24) and (4.25), we have the upper bound

$$
\begin{aligned}
&n(\tilde{\sigma}^2-\sigma^2)\\
&\quad=\frac{\sigma^2}{\sigma^4 B}\left[\left(1-\frac{p}{n}-\frac{1}{n}\right)n\left[\frac{1}{p}\sum_{c=1}^{p}\frac{\sigma^2}{l_c}-\frac{1}{1-\frac{p-\rho}{n}}\right]\right.\\
&\qquad\qquad-\frac{n(\rho+1)}{n-p}\\
&\qquad\quad-\left(1-\frac{p}{n}-\frac{1}{n}\right)\left(1-\frac{p}{n}-\frac{2}{n}\right)n\left[\frac{1}{p}\sum_{c=1}^{p}\frac{\sigma^4}{l_c^2}-\frac{1}{(1-\frac{p-\rho}{n})^3}\right]\\
&\qquad\qquad+\frac{\rho n^2}{(n-p)(n-p+\rho)}+\frac{(2\rho+3)n^2}{(n-p+\rho)^2}-\frac{(\rho^2+3\rho+2)n^2}{(n-p+\rho)^3}\\
&\qquad+\left(1-\frac{p}{n}-\frac{1}{n}\right)\frac{p}{n}\left[\frac{1}{p}\sum_{c=1}^{p}\frac{\sigma^2}{l_c}+\frac{1}{1-\frac{p-\rho}{n}}\right]n\left[\frac{1}{p}\sum_{c=1}^{p}\frac{\sigma^2}{l_c}-\frac{1}{1-\frac{p-\rho}{n}}\right]\\
&\qquad\qquad\left.+\frac{rn}{n-p+\rho}-\frac{(\rho+1)pn}{(n-p+\rho)^2}\right]+\frac{nE_n}{B}\\
&\le\frac{\sigma^2}{\sigma^4 B}\left[\left(1-\frac{p}{n}-\frac{1}{n}\right)\frac{p-\rho}{p}n\left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho}\frac{1}{\mu_c}-\frac{1}{1-\frac{p-\rho}{n}}\right]-\frac{n(\rho+1)}{n-p}\right.\\
&\qquad\qquad+\left(1-\frac{p}{n}-\frac{1}{n}\right)\rho\frac{n}{p}\frac{\sigma^2}{l_p}-\left(1-\frac{p}{n}-\frac{1}{n}\right)\frac{\rho n^2}{(n-p+\rho)p}\\
&\qquad\quad-\left(1-\frac{p}{n}-\frac{1}{n}\right)\left(1-\frac{p}{n}-\frac{2}{n}\right)\frac{p-\rho}{p}n\left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho}\frac{1}{\mu_c^2}-\frac{1}{(1-\frac{p-\rho}{n})^3}\right]\\
&\qquad\qquad-\left(1-\frac{p}{n}-\frac{1}{n}\right)\left(1-\frac{p}{n}-\frac{2}{n}\right)\frac{n}{p}\sum_{c=1}^{\rho}\frac{\sigma^4}{l_c^2}\\
&\qquad\qquad+\left(1-\frac{p}{n}-\frac{1}{n}\right)\left(1-\frac{p}{n}-\frac{2}{n}\right)\frac{\rho n^4}{(n-p+\rho)^3 p}\\
&\qquad\qquad+\frac{\rho n^2}{(n-p)(n-p+\rho)}+\frac{(2\rho+3)n^2}{(n-p+\rho)^2}-\frac{(\rho^2+3\rho+2)n^2}{(n-p+\rho)^3}\\
&\qquad+\left(1-\frac{p}{n}-\frac{1}{n}\right)\frac{p}{n}\left[\frac{1}{p}\sum_{c=1}^{p}\frac{\sigma^2}{l_c}+\frac{1}{1-\frac{p-\rho}{n}}\right]\cdot\\
&\qquad\qquad\qquad\frac{p-\rho}{p}n\left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho}\frac{1}{\mu_c}-\frac{1}{1-\frac{p-\rho}{n}}\right]\\
&\qquad\qquad+\left(1-\frac{p}{n}-\frac{1}{n}\right)\frac{p}{n}\left[\frac{1}{p}\sum_{c=1}^{p}\frac{\sigma^2}{l_c}+\frac{1}{1-\frac{p-\rho}{n}}\right]\rho\frac{n}{p}\frac{\sigma^2}{l_p}
\end{aligned}
$$

$$
\begin{aligned}
&- \left(1 - \frac{p}{n} - \frac{1}{n}\right) \frac{p}{n} \left[\frac{1}{p}\sum_{c=1}^{p} \frac{\sigma^2}{l_c} + \frac{1}{1 - \frac{p-\rho}{n}}\right] \frac{\rho n^2}{(n-p+\rho)p} \\
&+ \frac{\rho n}{n-p+\rho} - \frac{(\rho+1)pn}{(n-p+\rho)^2}\Bigg] + \frac{nE_n}{B} \\
= \frac{\sigma^2}{\sigma^4 B} &\left(a_1^{(n)} n \left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho} \frac{1}{\mu_c} - \frac{1}{1-c}\right] + a_2^{(n)} n \left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho} \frac{1}{\mu_c^2} - \frac{1}{(1-c)^3}\right] \right. \\
&\left. + b^{(n)} \right) + \frac{nE_n}{B}.
\end{aligned}
$$

Now note that

$$
\begin{aligned}
a_1^{(n)} &\overset{\text{a.s.}}{\to} 1 + c, \qquad a_2^{(n)} \overset{\text{a.s.}}{\to} -(1-c)^2, \\
b^{(n)} &\overset{\text{a.s.}}{\to} \frac{2c(\rho+1)-1}{(1-c)^2} + \frac{(2-c)(1+c)\rho}{(1-\sqrt{c})^2} - \frac{(1-c)^2}{c} \sum_{k=1}^{\rho} \frac{\sigma^4 \gamma_k^2}{(\gamma_k + \sigma^2)^2 (\gamma_k + c\sigma^2)^2}
\end{aligned}
$$

using Lemma 4 and Baik and Silverstein [2006], Theorem 1.1. Therefore, using Slutsky and eq. (4.26),

$$
\begin{aligned}
a_1^{(n)} n \left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho} \frac{1}{\mu_c} - \frac{1}{1-c}\right] &+ a_2^{(n)} n \left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho} \frac{1}{\mu_c^2} - \frac{1}{(1-c)^3}\right] + b^{(n)} \\
&\xrightarrow{\mathcal{D}} \quad \mathrm{N}\left(\mu, \frac{2c(1+c)^2}{(1-c)^4}\right),
\end{aligned}
$$

with

$$
\mu = \frac{(2-c)(1+c)\rho}{(1-\sqrt{c})^2} + \frac{2c\rho - 1}{(1-c)^2} - \frac{(1-c)^2}{c} \sum_{k=1}^{\rho} \frac{\sigma^4 \gamma_k^2}{(\gamma_k + \sigma^2)^2 (\gamma_k + c\sigma^2)^2}.
$$

Therefore, using eq. (4.18) and (4.20) we obtain $n(\tilde{\sigma}^2 - \sigma^2) \leq X_n^+$ with $X_n^+ \xrightarrow[n\to\infty]{\mathcal{D}}$ $\mathrm{N}\left(\mu^+, \frac{2c(1+c)^2\sigma^4}{(1-c)^4}\right)$, where

$$
\begin{aligned}
\mu^+ = \frac{(2c\rho - 1)\sigma^2}{(1-c)^2} &+ \frac{(2-c)(1+c)\rho\sigma^2}{(1-\sqrt{c})^2} - \frac{(1-c)^2}{c} \sum_{k=1}^{\rho} \frac{\sigma^6 \gamma_k^2}{(\gamma_k + \sigma^2)^2 (\gamma_k + c\sigma^2)^2} \\
&+ \frac{(1-c)^2}{c} \sum_{k=1}^{\rho} \frac{\gamma_k^2 \bar{\gamma}_k \sigma^4}{(\gamma_k + \sigma^2)^2 (\gamma_k + c\sigma^2)^2} + \sum_{k=1}^{\rho} \frac{\gamma_k \bar{\gamma}_k \sigma^2}{(\gamma_k + \sigma^2)(\gamma_k + c\sigma^2)} \\
&- 2 \sum_{k=1}^{\rho} \frac{\gamma_k \bar{\gamma}_k \sigma^2}{(\gamma_k + c\sigma^2)^2}. \qquad (4.27)
\end{aligned}
$$

The same argument can be done to obtain a lower bound. Using eq. (4.19), (4.21), (4.22), (4.23), (4.24) and (4.25) again, we get

$$\begin{aligned} n(\tilde{\sigma}^2 - \sigma^2) \leq \; & \frac{\sigma^2}{\sigma^4 B}\left(a_1^{(n)} n \left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho} \frac{1}{\mu_c} - \frac{1}{1-c} \right] \right. \\ & \left. + a_2^{(n)} n \left[\frac{1}{p-\rho}\sum_{c=1}^{p-\rho} \frac{1}{\mu_c^2} - \frac{1}{(1-c)^3} \right] + b^{(n)} \right) + \frac{nE_n}{B} \end{aligned}$$

where

$$\begin{aligned} & a_1^{(n)} \overset{\text{a.s.}}{\to} 1+c, \qquad a_2^{(n)} \overset{\text{a.s.}}{\to} -(1-c)^2, \\ & b^{(n)} \overset{\text{a.s.}}{\to} \frac{2c(\rho+1)-1}{(1-c)^2} - \frac{(1-c)^2\rho}{c(1-\sqrt{c})^2} + \frac{1+c}{c}\sum_{k=1}^{\rho} \frac{\sigma^2\gamma_k}{(\gamma_k+\sigma^2)(\gamma_k+c\sigma^2)} \end{aligned}$$

Therefore, again using eq. (4.18) and (4.20) we obtain that $n(\tilde{\sigma}^2 - \sigma^2) \geq X_n^-$ with $X_n^- \xrightarrow[n\to\infty]{\mathcal{D}} \mathrm{N}\left(\mu^-, \frac{2c(1+c)^2\sigma^4}{(1-c)^4}\right)$, where

$$\begin{aligned} \mu^- = \; & \frac{(2c\rho-1)\sigma^2}{(1-c)^2} - \frac{(1-c)^2\rho\sigma^2}{c(1-\sqrt{c})^2} + \frac{1+c}{c}\sum_{k=1}^{\rho} \frac{\sigma^4\gamma_k}{(\gamma_k+\sigma^2)(\gamma_k+c\sigma^2)} \\ & + \frac{(1-c)^2}{c}\sum_{k=1}^{\rho} \frac{\gamma_k^2\bar{\gamma}_k\sigma^4}{(\gamma_k+\sigma^2)^2(\gamma_k+c\sigma^2)^2} + \sum_{k=1}^{\rho} \frac{\gamma_k\bar{\gamma}_k\sigma^2}{(\gamma_k+\sigma^2)(\gamma_k+c\sigma^2)} \\ & - 2\sum_{k=1}^{\rho} \frac{\gamma_k\bar{\gamma}_k\sigma^2}{(\gamma_k+c\sigma^2)^2} - 3c\sum_{k=1}^{\rho} \frac{\bar{\gamma}_k\sigma^4}{(\gamma_k+c\sigma^2)^2}. \end{aligned} \tag{4.28}$$

This concludes the proof. □

Proof of Lemma 5. Let $\Sigma' \in \mathrm{B}_r(\Sigma, 2M)$, and write $\lambda_i = \lambda_i(\Sigma)$, $\lambda_i' = \lambda_i'(\Sigma_p')$ to simplify notation. Since the sequences are spiked, there are a finite number of different eigenvalues as $n \to \infty$. We can decompose

$$\mathrm{N}(0, \Sigma_p)^n = \bigotimes_{i=1}^{p} \mathrm{N}(0, \lambda_i)^n, \qquad \mathrm{N}(0, \Sigma_p')^n = \bigotimes_{i=1}^{p} \mathrm{N}(0, \lambda_i')^n.$$

Let λ be the Lebesgue measure on $\mathbb{R}$. Writing the Hellinger affinity between two densities f, g as $\alpha(f,g) = \int \sqrt{fg}d\lambda$, we have by Cauchy-Schwarz

$$\begin{aligned} \delta_{TV}\big(f,g\big)^2 = 4(\int |f-g|d\lambda)^2 \leq \; & \frac{1}{4}\big(1-\alpha(f,g)\big)\big(1+\alpha(f,g)\big) \\ = \; & (1-\alpha(f,g)^2) \end{aligned}$$

Now for two normals, we have

$$\alpha\big(\mathrm{N}(0,a^2),\mathrm{N}(0,b^2)\big)^2 = \frac{2ab}{a^2+b^2}$$

and since Hellinger affinity distributes over a product of independent densities, we get

$$\delta_{TV}\left(\mathrm{N}(0,\Sigma_p)^n,\mathrm{N}(0,\Sigma_p')^n\right)^2 \le 1-\prod_{i=1}^{p}\left(\frac{2\sqrt{\lambda_i\lambda_i'}}{\lambda_i+\lambda_i'}\right)^n$$

Now note that $\frac{2\sqrt{\lambda_i\lambda_i'}}{\lambda_i+\lambda_i'} > \frac{\sqrt{1-2M/n^r}}{1-M/n^r}$ if and only if $\lambda_i(1-\frac{2M}{n^r}) < x < \lambda_i(1+\frac{2M}{n^r(1-2M/n^r)})$. By definition, $\Sigma' \in \mathrm{B}_r(\Sigma,2M)$ means $|\lambda_i-\lambda_i'| < \frac{2M\lambda_i}{n^r}$ for all i, so we have the bound

$$< 1-\left(\frac{\sqrt{1-2M/n^r}}{1-M/n^r}\right)^{np} = 1-\left(1-\frac{M^2}{n^{2r}(1-M/n^r)^2}\right)^{np/2}$$

over all $\Sigma' \in \mathrm{B}_r(\Sigma,2M)$. Taking a supremum and a limit yields

$$\lim_{n\to\infty}\sup_{\Sigma'\in\mathrm{B}_1(\Sigma,2M)}\delta_{TV}\left(\mathrm{N}(0,\Sigma_p)^n,\mathrm{N}(0,\Sigma_p')^n\right)^2$$
$$\le \lim_{n\to\infty}\left[1-\left(1-\frac{M^2}{n^2(1-M/n)^2}\right)^{np/2}\right] = 1-e^{-cM^2/2}$$

and

$$\lim_{n\to\infty}\sup_{\Sigma'\in\mathrm{B}_r(\Sigma,2M)}\delta_{TV}\left(\mathrm{N}(0,\Sigma_p)^n,\mathrm{N}(0,\Sigma_p')^n\right)^2 = 0,$$

as desired. □

Proof of Theorem 8. Define $\Sigma_p' = (1-\frac{2M_\epsilon}{n})\Sigma_p$. We have

$$|\sigma^2-\sigma^{2\prime}| = 2\sigma^2\frac{M_\epsilon}{n}. \tag{4.29}$$

Moreover, according to Lemma 5, we have

$$\lim_{n\to\infty}\delta_{TV}\left(\mathrm{N}(0,\Sigma_p)^n,\mathrm{N}(0,\Sigma_{p,M}')^n\right) \le \sqrt{1-\exp\left(-\frac{cM_\epsilon^2}{2}\right)} = 1-4\epsilon,$$

so for some N_ϵ we have $\delta_{TV} < 1 - 2\epsilon$ for all $n \geq N_\epsilon$. Now note that $\Sigma'_p \in \mathrm{B}_1(\Sigma_p, \frac{2M_\epsilon}{n}\|\Sigma_p\|_2)$. Say for some $\hat{\sigma}^2$ we have

$$\sup_{\Sigma' \in \mathrm{B}_1(\Sigma, 2M_\epsilon)} \mathrm{P}_{\Sigma'_p}\left[|\hat{\sigma}^2 - \sigma^{2\prime}| \geq \sigma^2 \frac{M_\epsilon}{n}\right] < \epsilon. \tag{4.30}$$

Define the event $A = \left[|\hat{\sigma}^2 - \sigma^2| \leq \sigma^2 \frac{M_\epsilon}{n}\right]$ - then by (4.29) and (4.30) we find $\mathrm{P}_{\Sigma_p}[A] \geq 1 - \epsilon$ and $\mathrm{P}_{\Sigma'_p}[A] < \epsilon$. Therefore,

$$\delta_{TV}\left(\mathrm{N}(0, \Sigma_p)^n, \mathrm{N}(0, \Sigma'_{p,M})^n\right) \geq \mathrm{P}_{\Sigma_p}[A] - \mathrm{P}_{\Sigma'_p}[A] \geq 1 - 2\epsilon$$

which contradicts $\delta_{TV} < 1 - 2\epsilon$ for all $n \geq N_\epsilon$ - therefore,

$$\sup_{\Sigma' \in \mathrm{B}_1(\Sigma, 2M_\epsilon)} \mathrm{P}_{\Sigma'_p}\left[|\hat{\sigma}^2 - \sigma^{2\prime}| \geq \sigma^2 \frac{M_\epsilon}{n^r}\right] \geq \epsilon$$

for all $\hat{\sigma}^2$ and $n \geq N_\epsilon$, yielding the desired result. □

Proof of Proposition 9. We have:

$$\begin{aligned}
\mathrm{P}_{\Sigma'_p}\left[|\tilde{\sigma}^2 - \sigma^{2\prime}| \geq \sigma^2 \frac{M}{n^r}\right] &\leq \mathrm{P}_{\Sigma_p}\left[|\tilde{\sigma}^2 - \sigma^{2\prime}| \geq \sigma^2 \frac{M}{n^r}\right] \\
&\quad + \left|\mathrm{P}_{\Sigma_p}\left[|\tilde{\sigma}^2 - \sigma^{2\prime}| \geq \sigma^2 \frac{M}{n^r}\right] - \mathrm{P}_{\Sigma'_p}\left[|\tilde{\sigma}^2 - \sigma^{2\prime}| \geq \sigma^2 \frac{M}{n^r}\right]\right| \\
&= A_p + B_p.
\end{aligned}$$

Choose any $\Sigma' \in \mathrm{B}_r(\Sigma, 2M)$ and write $\lambda_i = \lambda_i(\Sigma)$, $\lambda'_i = \lambda'_i(\Sigma'_p)$. Using Lemma 5, we find for the second term

$$\begin{aligned}
\lim_{n\to\infty} \sup_{\Sigma \in \mathrm{B}_r(\Sigma, 2M)} B_p &\leq \lim_{n\to\infty} \sup_{\Sigma' \in \mathrm{B}_r(\Sigma, 2M)} \delta_{TV}\left(\mathrm{N}(0, \Sigma_p)^n, \mathrm{N}(0, \Sigma'_p)^n\right) \\
&= 0.
\end{aligned}$$

For the first term, let us use Theorem 7. Since $|\sigma^{2\prime} - \sigma^2| < \sigma^2 M/n^r$, we have

$$\begin{aligned}
\sup_{\Sigma \in \mathrm{B}(\Sigma, 2M)} A_p \leq\ & \mathrm{P}_{\Sigma_p}\left[n\left|\tilde{\sigma}^2 - \sigma^2\right| > \sigma^2 \frac{M}{n^{r-1}}\right] \\
& \leq \mathrm{P}_{\Sigma_p}\left[\frac{(1-c)^2}{\sqrt{2c}(1+c)\sigma^2}\left(X_n^+ - \mu^+\right) > \frac{(1-c)^2(-2\sigma^2\mu^+ + M/n^{r-1})}{2\sqrt{2c}(1+c)}\right] \\
& \quad + \mathrm{P}_{\Sigma_p}\left[\frac{(1-c)^2}{\sqrt{2c}(1+c)\sigma^2}\left(X_n^- - \mu^-\right) < \frac{(1-c)^2(-2\sigma^2\mu^- - M/n^{r-1})}{2\sqrt{2c}(1+c)}\right] \\
& \xrightarrow{n\to\infty} 1 - \Phi(\infty) + \Phi(-\infty) = 0.
\end{aligned}$$

This concludes the proof. □

4.6.3 Proofs for Section 4.4

Proof of proposition 10. We first notice that, from Lemma 4 and Theorem 7, for any $1 \leq k \leq \rho$ we have

$$\hat{\gamma}_{\rho,k} \xrightarrow[n\to\infty]{\text{a.s.}} \gamma_k, \qquad \tilde{\sigma}_\rho^2 \xrightarrow[n\to\infty]{\text{a.s.}} \sigma^2. \tag{4.31}$$

Denote as usual the unbiased risk estimator of Theorem 6 at $\hat{\Gamma}_r + \tilde{\sigma}_r^2 I_p$ by $F_r + G_r$. By their definitions and Lemma 4, we see that

$$F_\rho + G_\rho \xrightarrow[n\to\infty]{\text{a.s.}} 0.$$

Thus there exists a N_1 (random) such that $\rho \in \left\{r \,|\, |F_r + G_r| \leq \frac{p+1}{n}\right\}$ for all $n \geq N_1$. Now note that there is a N_2 (also random) such that for all $n \geq N_2$, $\left\{\frac{\mathbb{1}[r<p]}{l_{r+1}} \frac{(1+\sqrt{p/n})^2}{p-r} \sum_{c=r+1}^{p} l_c \geq 1\right\} = \{\rho\}$. Thus for all $n \geq N_1 \vee N_2$,

$$\hat{\rho} = \min\{\rho\} = \rho,$$

so in particular $\hat{\rho} \xrightarrow[n\to\infty]{\text{a.s.}} \rho$. Thus, by eq. (4.31) and any $\epsilon > 0$ there exists a N_3 random such that for all $n \geq N_1 \vee N_2 \vee N_3$, $k < p$ and $|\hat{\gamma}_{\hat{\rho},k} - \gamma_k| = |\hat{\gamma}_{\rho,k} - \gamma_k| < \epsilon$. That is, $\gamma_{\hat{\rho},k} \xrightarrow[n\to\infty]{\text{a.s.}} \gamma_k$. Finally, again by eq. (4.31) for all $\epsilon > 0$, there exists a N_3

random such that for all $n \geq N_1 \vee N_2 \vee N_3$, $|\tilde{\sigma}^2 - \sigma^2| = |\tilde{\sigma}^2_{\hat{\rho}} - \sigma^2| = |\tilde{\sigma}^2_{\rho} - \sigma^2| < \epsilon$, i.e. $\tilde{\sigma}^2 \xrightarrow[n\to\infty]{\text{a.s.}} \sigma^2$, as desired. $\square$

BIBLIOGRAPHY

A. Aspremont, O. Banerjee, and L.E. Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30: 56–66, 2008.

Z.D. Bai and J.W. Silverstein. CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Annals of Probability*, 32:553–605, 2004.

J. Baik and J.W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97:1382–1408, 2006.

J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, pages 1643–1697, 2005.

A.J. Baranchik. A family of minimax estimators of the mean of a multivariate normal distribution. *Annals of Mathematical Statistics*, 41:642–656, 1970.

J.O. Berger and M.E. Bock. Combining independent normal mean estimation problems with unknown variances. *Annals of Statistics*, 4:642–648, 1976.

J.O. Berger and L.R. Haff. A class of minimax estimators of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. *Statistics & Decisions*, 1:105–130, 1983.

J.O. Berger, M.E. Bock, L.D. Brown, G. Casella, and L. Gleser. Minimax estimation of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. *Annals of Statistics*, 5:763–771, 1977.

P.J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Annals of Statistics*, 36:190–227, 2008a.

P.J. Bickel and E. Levina. Covariance regularization by thresholding. *Annals of Statistics*, pages 2577–2604, 2008b.

J. Bien, F. Bunea, and L. Xiao. Convex banding of the covariance matrix. *arXiv preprint arXiv:1405.6210*, 2014.

L. Brown. Estimation with incomplete specified loss functions (the case with several location parameters). *Journal of the American Statistical Association*, 70:417–427, 1975.

L.D. Brown, H. Nie, and X. Xie. Ensemble minimax estimation for multivariate normal means. *Annals of Statistics*, page To appear, 2012.

F. Bunea and L. Xiao. On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA. *arXiv preprint arXiv:1212.5321*, 2012.

T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106:672–684, 2011.

T. Cai, W. Liu, and X. Luo. A constrained $\ell 1$ minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106:594–607, 2011.

G. Casella. Minimax ridge regression estimation. *Annals of Statistics*, 8:1036–1056, 1980.

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

D. Chételat and M.T. Wells. Noise estimation in the spiked covariance model. *arXiv preprint arXiv:1408.6440*, 2014.

D.K. Dey. Improved estimation of a multinormal precision matrix. *Statistics & probability letters*, 6(2):125–128, 1987.

D.K. Dey and C. Srinivasan. On estimation of discriminant coefficients. *Statistics & probability letters*, 11(3):189–193, 1991.

L.D. Donoho, M. Gavish, and I. Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *arXiv preprint arXiv:1311.0851*, 2014.

B. Efron and C. Morris. Multivariate empirical Bayes and estimation of covariance matrices. *The Annals of Statistics*, pages 22–32, 1976.

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

N. El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Annals of Statistics*, pages 2717–2756, 2008b.

K.T. Fang, S. Kotz, and K.W. Ng. *Symmetric Multivariate and Related Distributions*. Chapman and Hall, 1990.

D. Fourdrinier and W.E. Strawderman. On Bayes and unbiased estimators of loss. *Annals of the Institute of Statistical Mathematics*, 55:803–816, 2003.

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

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2008.

M. Giaquinta and S. Hildebrandt. *Calculus of Variations*, volume I. Springer, Berlin, 1996.

L.J. Gleser. Minimax estimation of a normal mean vector when the covariance matrix is unknown. *Annals of Statistics*, 7:838–846, 1979.

L.J. Gleser. Minimax estimators of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. *Annals of Statistics*, 14:1625–1633, 1986.

G.H. Golub and V. Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal of Numerical Mathematics*, 10:413–432, 1973.

L.R. Haff. Minimax estimators for a multinormal precision matrix. *Journal of Multivariate Analysis*, 7:374–385, 1977.

L.R. Haff. An identity for the Wishart distribution with applications. *Journal of Multivariate Analysis*, 9:531–544, 1979a.

L.R. Haff. Estimation of the inverse covariance matrix: Random mixtures of the inverse Wishart matrix and the identity. *The Annals of Statistics*, pages 1264–1276, 1979b.

L.R. Haff. Empirical Bayes estimation of the multivariate normal covariance matrix. *Annals of Statistics*, 8:586–597, 1980.

L.R. Haff. On linear log-odds and estimation of discriminant coefficients. *Communications in Statistics-Theory and Methods*, 15(7):2131–2144, 1986.

L.R. Haff. The variational form of certain Bayes estimators. *Annals of Statistics*, 19:1109–1680, 1991.

D.A. Harville. *Matrix algebra from a statistician's perspective*, volume 157. Springer, 1997.

W. James and C. Stein. Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, volume 1, pages 361–380. Berkeley, University of California Press, 1961.

I.M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, pages 295–327, 2001.

I.M. Johnstone and A.Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104, 2009.

T. Kollo and D. von Rosen. *Advanced multivariate statistics with matrices*, volume 579. Springer Science & Business Media, 2006.

Y. Konno. Shrinkage estimators for large covariance matrices in multivariate real and complex normal distributions under an invariant quadratic loss. *Journal of Multivariate Analysis*, 100:2237–2253, 2009.

T. Kubokawa and M.S. Srivastava. Robust improvement in estimation of a mean matrix in an elliptically contoured distribution. *Journal of Multivariate Analysis*, 76:138–152, 2001.

T. Kubokawa and M.S. Srivastava. Estimation of the precision matrix of a singular Wishart distribution and its application in high-dimensional data. *Journal of Multivariate Analysis*, 99:1906–1928, 2008.

L. Le Cam. Convergence of estimates under dimensionality restrictions. *Annals of Statistics*, 1:38–53, 1973.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411, 2004.

O. Ledoit and M. Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics*, 40:1024–1060, 2012.

E. Levina, A. Rothman, and J. Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. *Annals of Applied Statistics*, 2:245–263, 2008.

S.P. Lin and M.D. Perlman. A Monte Carlo comparison of four estimators of a covariance matrix. In P.R. Krishnaiah, editor, *Multivariate Analysis VI*, pages 411–429. Elsevier Science Publishers B.V., Amsterdam, 1985.

J.R. Magnus. On certain moments relating to ratios of quadratic forms in normal variables: further results. *Sankhyā B*, 52:1–13, 1990.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, pages 1436–1462, 2006.

C. Morris and M. Lysy. Shrinkage estimation in multi-level normal models. *Statistical Science*, 27:115–134, 2009.

R.J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley, New York, 1982.

B. Nadler. Finite sample approximation results for principal component analysis: a matrix perturbation approach. *Annals of Statistics*, 36:2791–2817, 2008.

D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617, 2007.

P. Ravikumar, M.J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing $\ell 1$-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

A. J. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104:177–186, 2009.

A.J. Rothman, P.J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Y. Sheena. Unbiased estimator of risk for an orthogonally invariant estimator of a covariance matrix. *Journal of the Japan Statistical Society*, 25:35–48, 1995.

M.S. Srivastava. Stein estimation under elliptical distributions. *Journal of Multivariate Analysis*, 28:247–259, 1989.

M.S. Srivastava. Singular Wishart and multivariate beta distributions. *Annals of Statistics*, 31:1537–1560, 2003.

M.S. Srivastava. Multivariate theory for analyzing high dimensional data. *Journal of the Japanese Statistical Society*, 37:53–86, 2007.

M.S. Srivastava and Y. Fujikoshi. Multivariate analysis of variance with fewer observations than the dimension. *Journal of Multivariate Analysis*, 97:1927–1940, 2006.

M.S. Srivastava and C.G. Khatri. *An Introduction to Multivariate Statistics*. North-Holland, New York, 1979.

C. Stein. Estimation of a covariance matrix. In *Rietz Lecture, 39th Annual Meeting of the IMS*. Atlanta, Georgia, 1975.

C. Stein. Lectures on the theory of estimation of many parameters. In *Studies in the Statistical Theory of Estimation, Part I.*, volume 74, pages 4–65. Proc. Scientific Seminars Steklov Institute, Leningrad Division, 1977. (In Russian).

C. Stein. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9:1135–1151, 1981.

C. Stein. Lectures on the theory of estimation of many parameters. *Journal of Soviet Mathematics*, 34:1373–1403, 1986.

Y. Tian and S. Cheng. Some identities for Moore-Penrose inverses of matrix products. *Linear and Multilinear Algebra*, 52:405–420, 2004.

H. Tsukuma and T. Kubokawa. Estimation of the mean vector in a singular multivariate normal distribution. Draft, 2014.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.