A WELL-CONDITIONED ESTIMATOR FOR LARGE DIMENSIONAL COVARIANCE MATRICES

Olivier Ledoit
Michael Wolf

00-76

WORKING PAPERS

Universidad Carlos III de Madrid

Working Paper 00-76
Statistics and Econometrics Series 35
November 2000

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

A WELL-CONDITIONED ESTIMATOR FOR LARGE DIMENSIONAL COVARIANCE MATRICES

Olivier Ledoit and Michael Wolf *

Abstract

Many economic problems require a covariance matrix estimator that is not only invertible, but also well-conditioned (that is, inverting it does not amplify estimation error). For large-dimensional covariance matrices, the usual estimator –the sample covariance matrix- is typically not well-conditioned and may not even be invertible. This paper introduces an estimator that is both well-conditioned and more accurate than the sample covariance matrix asymptotically. This estimator is distribution-free and has a simple explicit formula that is easy to compute and interpret. It is the asymptotically optimal convex linear combination of the sample covariance matrix with the identity matrix. Optimality is meant with respect to a quadratic loss function, asymptotically as the number of observations and the number of variables go to infinity together. Extensive Monte-Carlo confirm that the asymptotic results tend to hold well in finite sample.

Keywords: Condition number; Covariance matrix estimation; Empirical Bayes; General asymptotics; Shrinkage.

*Ledoit, Anderson Graduate School of Management, UCLA, USA; Wolf, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, C/ Madrid 126 28903 Getafe, Madrid, Spain, e-mail: mwolf@est-econ.uc3m.es; Phone: 34-91-6249893. We wish to thank Petr Adamek, Dimitri Bertsimas, Minh Chau, Whitney Newey, Firooz Partovi, Pedro Santa-Clara and specially Tim Crack and Jack Silverstein for their valuable feedback. Also, the paper has benefited from seminar participants at MIT, the NBER, UCLA, Washington University in Saint Louis, Yale, Chicago, Wharton and UBC. Errors are still our own. Research of the second author partly funded by the Spanish "Dirección General de Enseñanza Superior" (DGES), reference number PB98-0025.

1 Introduction

Many problems of variance minimization in Finance and Economics are solved by inverting a covariance matrix. Sometimes the matrix dimension can be large. Examples include selecting a mean-variance efficient portfolio from a large universe of stocks (Markowitz, 1952), running Generalized Least Squares (GLS) regressions in large cross-sections (see e.g. Kandel and Stambaugh, 1995), and choosing an optimal weighting matrix in the General Method of Moments (GMM; see Hansen, 1982) when the number of moment restrictions is large. In such situations the usual estimator – the sample covariance matrix – is known to perform poorly. When the matrix dimension p is larger than the number of observations available n, the sample covariance matrix is not even invertible. When the ratio p/n is less than one but not negligible, the sample covariance matrix is invertible, but numerically ill-conditioned, which means that inverting it amplifies estimation error dramatically.[1] For large p, it is difficult to find enough observations to make p/n negligible. Therefore it is important to develop a well-conditioned estimator for large-dimensional covariance matrices.

If we wanted a well-conditioned estimator at any cost, we could always impose some ad-hoc structure on the covariance matrix to force it to be well-conditioned, such as diagonality or a factor model. But, in the absence of prior information about the true structure of the matrix, this ad-hoc structure will in general be misspecified. The resulting estimator can be so biased that it may bear little resemblance to the true covariance matrix. To the best of our knowledge, no existing estimator is both well-conditioned *and* more accurate than the sample covariance matrix. The contribution of this paper is to propose an estimator that possesses both these properties asymptotically.

One way to get a well-conditioned structured estimator is to impose that all variances are the same and all covariances are zero. The estimator that we recommend is a weighted average of this structured estimator with the sample covariance matrix. The average inherits the good conditioning of the structured estimator. By choosing the weight optimally according to a quadratic loss function, we can ensure that our weighted average of the sample covariance matrix and the structured estimator is more accurate than either of them. The only difficulty is that the true optimal weight depends on the true covariance matrix, which is unobservable. We solve this difficulty by finding a consistent estimator for the optimal weight. We also show that replacing the true optimal weight with a consistent estimator makes no difference asymptotically.

Standard asymptotics assume that the number of variables p is finite, while the number of observations n goes to infinity. Under standard asymptotics, the sample covariance matrix is well-conditioned (in the limit), and has some appealing optimality properties (e.g., maximum likelihood for normally distributed data). However, this is a bad approximation of many real-world situations where the number of variables p is of the same order of magnitude as the number of observations n, and possibly larger. We introduce a different framework, called *general* asymptotics, where we allow the number of variables p to go to infinity too. The

[1]The *condition number* is defined as the ratio of the largest to the smallest eigenvalue. It measures how invertible a matrix is. A matrix with low condition number can be safely inverted, and is called well-conditioned. A matrix with high condition number is almost not invertible, and is called ill-conditioned. Jobson and Korkie (1980) and Michaud (1989) show that it is difficult to use the sample covariance matrix for portfolio selection because it is typically ill-conditioned.

only constraint is that the ratio p/n must remain bounded. We see standard asymptotics as a special case in which it is optimal to (asymptotically) put all the weight on the sample covariance matrix and none on the structured estimator. In the general case, however, our estimator is different from the sample covariance matrix, substantially more accurate, and of course well-conditioned.

Extensive Monte-Carlo simulations indicate that: (i) the new estimator is more accurate than the sample covariance matrix, even for very small numbers of observations and variables, and usually by a lot; (ii) it is essentially as accurate or substantially more accurate than some estimators proposed in finite sample decision theory, as soon as there are at least ten variables and observations; (iii) it is better-conditioned than the true covariance matrix; and (iv) general asymptotics are a good approximation of finite sample behavior when there are at least twenty observations and variables.

The next section characterizes in finite sample the linear combination of the identity matrix and the sample covariance matrix with minimum quadratic risk. Section 3 develops a linear shrinkage estimator with uniformly minimum quadratic risk in its class asymptotically as the number of observations and the number of variables go to infinity together. In Section 4, Monte-Carlo simulations indicate that this estimator behaves well in finite sample. The conclusions suggest directions for future research.

2 Analysis in Finite Sample

The easiest way to explain what we do is to first analyze in detail the finite sample case. Let X denote a $p \times n$ matrix of n independent and identically distributed (iid) observations on a system of p random variables with mean zero and covariance matrix Σ. Following the lead of Muirhead and Leung (1987), we consider the Frobenius norm: $\|A\| = \sqrt{\mathrm{tr}(AA^t)/p}$.[2] Our goal is to find the linear combination $\Sigma^* = \rho_1 I + \rho_2 S$ of the identity matrix I and the sample covariance matrix $S = XX^t/n$ whose expected quadratic loss $\mathrm{E}[\|\Sigma^* - \Sigma\|^2]$ is minimum. Haff (1980) studied this class of linear shrinkage estimators, but did not get any optimality results. The optimality result that we obtain in finite sample will come at a price: Σ^* will not be a *bona fide* estimator, because it will require hindsight knowledge of four scalar functions of the true (and unobservable) covariance matrix Σ. This would seem like a high price to pay but, interestingly, it is not: In the next section, we are able to develop a *bona fide* estimator S^* with the same properties as Σ^* *asymptotically as the number of observations and the number of variables go to infinity together.* Furthermore, extensive Monte-Carlo simulations will indicate that twenty observations and variables are enough for the asymptotic approximations to typically hold well in finite sample. Even the formulas for Σ^* and S^* will look the same and will have the same interpretations. This is why we study the properties of Σ^* in finite sample "as if" it was a *bona fide* estimator.

[2] Dividing by the dimension p is not standard, but it does not matter in this section because p remains finite. The advantages of this convention are that the norm of the identity matrix is simply one, and that it will be consistent with Definition 2 below.

2.1 Optimal Linear Shrinkage

The squared Frobenius norm $\|\cdot\|^2$ is a quadratic form whose associated inner product is: $A_1 \circ A_2 = \text{tr}(A_1 A_2^t)/p$. Four scalars play a central role in the analysis: $\mu = \Sigma \circ I$, $\alpha^2 = \|\Sigma - \mu I\|^2$, $\beta^2 = \text{E}[\|S - \Sigma\|^2]$, and $\delta^2 = \text{E}[\|S - \mu I\|^2]$. We do not need to assume that the random variables in X follow a specific distribution, but we do need to assume that they have finite fourth moments, so that β^2 and δ^2 are finite. The following relationship holds.

Lemma 2.1 $\alpha^2 + \beta^2 = \delta^2$.

Proof of Lemma 2.1

$$\begin{aligned} \text{E}[\|S - \mu I\|^2] &= \text{E}[\|S - \Sigma + \Sigma - \mu I\|^2] && (1) \\ &= \text{E}[\|S - \Sigma\|^2] + \text{E}[\|\Sigma - \mu I\|^2] + 2\,\text{E}[(S - \Sigma) \circ (\Sigma - \mu I)] && (2) \\ &= \text{E}[\|S - \Sigma\|^2] + \|\Sigma - \mu I\|^2 + 2\,\text{E}[S - \Sigma] \circ (\Sigma - \mu I) && (3) \end{aligned}$$

Notice that $\text{E}[S] = \Sigma$, therefore the third term on the right hand side of Equation (3) is equal to zero. This completes the proof of Theorem 2.1. □

The optimal linear combination $\Sigma^* = \rho_1 I + \rho_2 S$ of the identity matrix I and the sample covariance matrix S is the standard solution to a simple quadratic programming problem under linear equality constraint.

Theorem 2.1 *Consider the optimization problem:*

$$\begin{aligned} &\min_{\rho_1, \rho_2} \text{E}[\|\Sigma^* - \Sigma\|^2] \\ &\text{s.t.} \quad \Sigma^* = \rho_1 I + \rho_2 S \end{aligned} \tag{4}$$

where the coefficients ρ_1 and ρ_2 are nonrandom. Its solution verifies:

$$\Sigma^* = \frac{\beta^2}{\delta^2}\mu I + \frac{\alpha^2}{\delta^2} S \tag{5}$$

$$\text{E}[\|\Sigma^* - \Sigma\|^2] = \frac{\alpha^2 \beta^2}{\delta^2}. \tag{6}$$

Proof of Theorem 2.1 By a change of variables, Problem (4) can be rewritten as:

$$\begin{aligned} &\min_{\rho, \nu} \text{E}[\|\Sigma^* - \Sigma\|^2] \\ &\text{s.t.} \quad \Sigma^* = \rho\nu I + (1 - \rho) S. \end{aligned} \tag{7}$$

With a little algebra, and using $\text{E}[S] = \Sigma$ as in the proof of Lemma 2.1, we can rewrite the objective as:

$$\text{E}[\|\Sigma^* - \Sigma\|^2] = \rho^2 \|\Sigma - \nu I\|^2 + (1 - \rho)^2 \text{E}[\|S - \Sigma\|^2]. \tag{8}$$

Therefore the optimal value of ν can be obtained as the solution to a reduced problem that does not depend on ρ: $\min_\nu \|\Sigma - \nu I\|^2$. Remember that the norm of the identity is one by convention, so the objective of this problem can be rewritten as: $\|\Sigma - \nu I\|^2 = \|\Sigma\|^2 - 2\nu\Sigma \circ I + \nu^2$. The first order condition is: $-2\Sigma \circ I + 2\nu = 0$. The solution is: $\nu = \Sigma \circ I = \mu$.

Replacing ν by its optimal value μ in Equation (8), we can rewrite the objective of the original problem as: $\mathrm{E}[\|\Sigma^* - \Sigma\|^2] = \rho^2\alpha^2 + (1-\rho)^2\beta^2$. The first order condition is: $2\rho\alpha^2 - 2(1-\rho)\beta^2 = 0$. The solution is: $\rho = \beta^2/(\alpha^2+\beta^2) = \beta^2/\delta^2$. Note that $1-\rho = \alpha^2/\delta^2$. At the optimum, the objective is equal to: $(\beta^2/\delta^2)^2\alpha^2 + (\alpha^2/\delta^2)^2\beta^2 = \alpha^2\beta^2/\delta^2$. This completes the proof. □

Note that μI can be interpreted as a shrinkage target and the weight β^2/δ^2 placed on μI as a shrinkage intensity. The Percentage Relative Improvement in Average Loss (PRIAL) over the sample covariance matrix is equal to:

$$\frac{\mathrm{E}[\|S-\Sigma\|^2] - \mathrm{E}[\|\Sigma^*-\Sigma\|^2]}{\mathrm{E}[\|S-\Sigma\|^2]} = \frac{\beta^2}{\delta^2}, \tag{9}$$

same as the shrinkage intensity. Therefore everything is controlled by the ratio β^2/δ^2, which is a properly normalized measure of the error of the sample covariance matrix S. Intuitively, if S is relatively accurate, then you should not shrink it too much, and shrinking it will not help you much either; if S is relatively inaccurate, then you should shrink it a lot, and you also stand to gain a lot from shrinking.

2.2 Interpretations

The mathematics underlying Theorem 2.1 are so rich that we are able to provide four complementary interpretations of it. One is geometric and the others echo some of the most important ideas in finite sample multivariate statistics.

First, we can see Theorem 2.1 as a projection theorem in Hilbert space. The appropriate Hilbert space is the space of p-dimensional symmetric random matrices A such that $\mathrm{E}[\|A\|^2] < \infty$. The associated norm is, of course, $\sqrt{\mathrm{E}[\|\cdot\|^2]}$, and the inner product of two random matrices A_1 and A_2 is $\mathrm{E}[A_1 \circ A_2]$. With this structure, Lemma 2.1 is just a rewriting of the Pythagorean Theorem. Furthermore, Formula (5) can be justified as follows: In order to project the true covariance matrix Σ onto the space spanned by the identity matrix I and the sample covariance matrix S, we first project it onto the line spanned by the identity, which yields the shrinkage target μI; then we project Σ onto the line joining the shrinkage target μI to the sample covariance matrix S. Whether the projection Σ^* ends up closer to one end of the line (μI) or to the other (S) depends on which one of them Σ was closer to. Figure 1 provides a geometrical illustration.

The second way to interpret Theorem 2.1 is as a trade-off between bias and variance. We seek to minimize mean squared error, which can be decomposed into variance and squared bias:

$$\mathrm{E}[\|\Sigma^* - \Sigma\|^2] = \mathrm{E}[\|\Sigma^* - \mathrm{E}[\Sigma^*]\|^2] + \|\mathrm{E}[\Sigma^*] - \Sigma\|^2. \tag{10}$$

The mean squared error of the shrinkage target μI is all bias and no variance, while for the sample covariance matrix S it is exactly the opposite: all variance and no bias. Σ^* represents the optimal trade-off between error due to bias and error due to variance. See Figure 2 for an illustration. The idea of a trade-off between bias and variance was already central to the original James-Stein (1961) shrinkage technique.

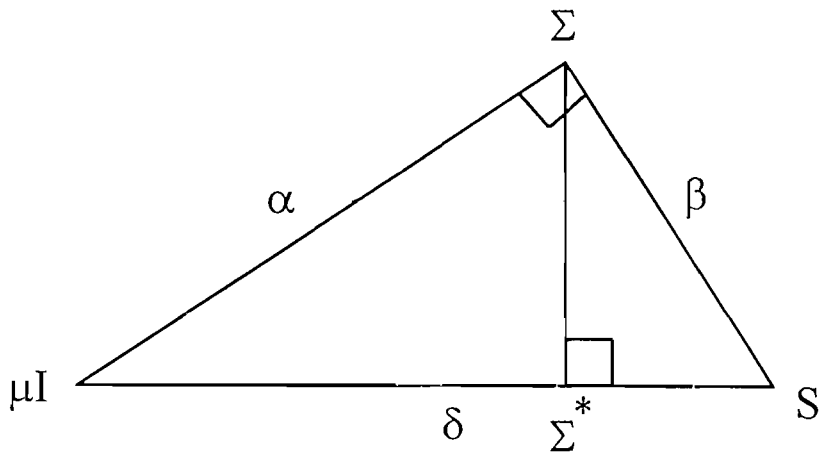

Figure 1: Theorem 2.1 Interpreted as a Projection in Hilbert Space.

The third interpretation is Bayesian. Σ^* can be seen as the combination of two signals: prior information and sample information. Prior information states that the true covariance matrix Σ lies on the sphere centered around the shrinkage target μI with radius α. Sample information states that Σ lies on another sphere, centered around the sample covariance matrix S with radius β. Bringing together prior and sample information, Σ must lie on the intersection of the two spheres, which is a circle. At the center of this circle stands Σ^*. The relative importance given to prior vs. sample information in determining Σ^* depends on which one is more accurate.[3] See Figure 3 for an illustration. The idea of drawing inspiration from the Bayesian perspective to obtain an improved estimator of the covariance matrix was used by Haff (1980).

The fourth and last interpretation involves the cross-sectional dispersion of covariance matrix eigenvalues. Let $\lambda_1, \ldots, \lambda_p$ denote the eigenvalues of the true covariance matrix Σ, and $l_1, \ldots, l_p$ those of the sample covariance matrix S. We can exploit the Frobenius norm's elegant relationship to eigenvalues. Note that

$$\mu = \frac{1}{p}\sum_{i=1}^{p}\lambda_i = \mathrm{E}\left[\frac{1}{p}\sum_{i=1}^{p} l_i\right] \tag{11}$$

represents the grand mean of both true and sample eigenvalues. Then Lemma 2.1 can be rewritten as:

$$\frac{1}{p}\mathrm{E}\left[\sum_{i=1}^{p}(l_i-\mu)^2\right] = \frac{1}{p}\sum_{i=1}^{p}(\lambda_i-\mu)^2 + \mathrm{E}[\|S-\Sigma\|^2]. \tag{12}$$

In words, sample eigenvalues are more dispersed around their grand mean than true ones, and the excess dispersion is equal to the error of the sample covariance matrix. Excess dispersion implies that the largest sample eigenvalues are biased upwards, and the smallest ones

[3]Strictly speaking, a full Bayesian approach would specify not only the support of the distribution of Σ, but also the distribution itself. We could assume that Σ is uniformly distributed on the sphere, but it might be difficult to justify. Thus, Σ^* should not be thought of as the expectation of the posterior distribution, as is traditional, but rather as the center of mass of its support.

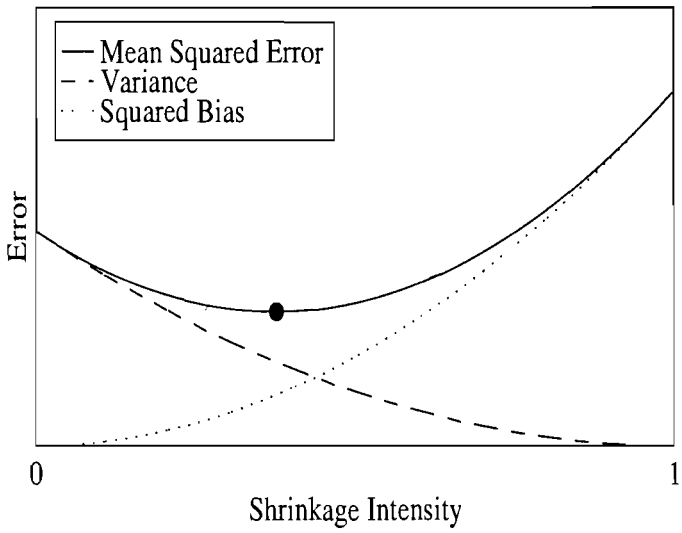

Figure 2: Theorem 2.1 Interpreted as a Trade-off Between Bias and Variance. Shrinkage intensity zero corresponds to the sample covariance matrix S. Shrinkage intensity one corresponds to the shrinkage target μI. Optimal shrinkage intensity (represented by $\bullet$) corresponds to the minimum expected loss combination Σ^*.

downwards. Therefore we can improve upon the sample covariance matrix by shrinking its eigenvalues towards their grand mean, as in:

$$\forall i = 1, \ldots, p \qquad \lambda_i^* = \frac{\beta^2}{\delta^2}\mu + \frac{\alpha^2}{\delta^2} l_i. \tag{13}$$

Note that $\lambda_1^*, \ldots, \lambda_p^*$ defined by Equation (13) are precisely the eigenvalues of Σ^*. Surprisingly, their dispersion $\mathrm{E}[\sum_{i=1}^p (\lambda_i^* - \mu)^2]/p = \alpha^2/\delta$ is even below the dispersion of true eigenvalues. For the interested reader, the next subsection explains why. The idea that shrinking sample eigenvalues towards their grand mean yields an improved estimator of the covariance matrix was highlighted in Muirhead's (1987) review paper.

2.3 Further Results on Sample Eigenvalues

The following paragraphs contain additional insights about the eigenvalues of the sample covariance matrix, but the reader can skip them and go directly to Section 3 if he or she so wishes. We discuss: 1) why the eigenvalues of the sample covariance matrix are *more* dispersed than those of the true covariance matrix (Equation (12)); 2) how important this effect is in practice; and 3) why we should use instead an estimator whose eigenvalues are *less* dispersed than those of the true covariance matrix (Equation (13)). The explanation relies on a result from matrix algebra.

Theorem 2.2 *The eigenvalues are the most dispersed diagonal elements that can be obtained by rotation.*

Proof of Theorem 2.2 Let R denote a p-dimensional symmetric matrix and V a p-dimensional rotation matrix: $VV' = V'V = I$. First, note that $(1/p)\mathrm{tr}(V'RV) = (1/p)\mathrm{tr}(R)$. The average

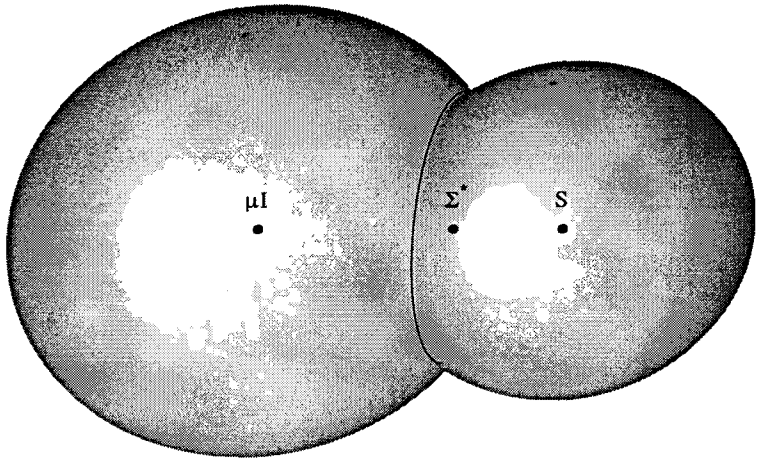

Figure 3: Bayesian Interpretation. The left sphere has center μI and radius α and represents prior information. The right sphere has center S and radius β. The distance between sphere centers is δ and represents sample information. If all we knew was that the true covariance matrix Σ lies on the left sphere, our best guess would be its center: the shrinkage target μI. If all we knew was that the true covariance matrix Σ lies on the right sphere, our best guess would be its center: the sample covariance matrix S. Putting together both pieces of information, the true covariance matrix Σ must lie on the circle where the two spheres intersect, therefore our best guess is its center: the optimal linear shrinkage Σ^*.

of the diagonal elements is invariant by rotation. Call it r. Let v_i denote the i^{th} column of V. The dispersion of the diagonal elements of $V'RV$ is $(1/p)\sum_{i=1}^{p}(v_i'Rv_i - r)^2$. Note that $\sum_{i=1}^{p}(v_i'Rv_i - r)^2 + \sum_{i=1}^{p}\sum_{\substack{j=1 \\ j\neq i}}^{p}(v_i'Rv_j)^2 = \text{tr}[(V'RV - rI)^2] = \text{tr}[(R - rI)^2]$ is invariant by rotation. Therefore the rotation V maximizes the dispersion of the diagonal elements of $V'RV$ if and only if it minimizes $\sum_{i=1}^{p}\sum_{\substack{j=1 \\ j\neq i}}^{p}(v_i'Rv_j)^2$. This is achieved by setting $v_i'Rv_j$ to zero for all $i \neq j$. In this case, $V'RV$ is a diagonal matrix, call it D. $V'RV = D$ is equivalent to $R = VRV'$. Since V is a rotation and D is diagonal, the columns of V must contain the eigenvectors of R and the diagonal of D its eigenvalues. Therefore the dispersion of the diagonal elements of $V'RV$ is maximized when these diagonal elements are equal to the eigenvalues of R. This completes the proof of Theorem 2.2. □

Decompose the true covariance matrix into eigenvalues and eigenvectors: $\Sigma = \Gamma'\Lambda\Gamma$, where Λ is a diagonal matrix, and Γ is a rotation matrix. The diagonal elements of Λ are the eigenvalues $\lambda_1, \ldots, \lambda_p$, and the columns of Γ are the eigenvectors $\gamma_1, \ldots, \gamma_p$. Similarly, decompose the sample covariance matrix into eigenvalues and eigenvectors: $S = G'LG$, where L is a diagonal matrix, and G is a rotation matrix. The diagonal elements of L are the eigenvalues $l_1, \ldots, l_p$, and the columns of G are the eigenvectors $g_1, \ldots, g_p$.

Since S is unbiased and Γ is nonstochastic, $\Gamma'S\Gamma$ is an unbiased estimator of $\Lambda = \Gamma'\Sigma\Gamma$. The diagonal elements of $\Gamma'S\Gamma$ are approximately as dispersed as the ones of $\Gamma'\Sigma\Gamma$. For convenience, let us speak as if they were exactly as dispersed. By contrast, $L = G'SG$ is not at all an unbiased estimator of $\Gamma'\Sigma\Gamma$. This is because the errors of G and S interact.

Theorem 2.2 shows us the effect of this interaction: the diagonal elements of $G'SG$ are more dispersed than those of $\Gamma'S\Gamma$ (and hence than those of $\Gamma'\Sigma\Gamma$). This is why sample eigenvalues are more dispersed than true ones. See Table 1 for a summary.

$\Gamma'S\Gamma$	$\prec$	$G'SG$
$\wr\!\!\wr$		
$\Gamma'\Sigma\Gamma$	$\succ$	$G'\Sigma G$

Table 1: Dispersion of Diagonal Elements
This table compares the dispersion of the diagonal elements of certain products of matrices. The symbols $\prec$, $\approx$, and $\succ$ pertain to diagonal elements, and mean less dispersed than, approximately as dispersed as, and more dispersed than, respectively.

We illustrate how important this effect is in a particular case: when the true covariance matrix is the identity matrix. Let us sort the eigenvalues of the sample covariance matrix from largest to smallest, and plot them against their rank. The shape of the plot depends on the ratio p/n, but does not depend on the particular realization of the sample covariance matrix, at least approximately when p and n are very large. Figure 4 shows the distribution of sample eigenvalues for various values of the ratio p/n. This figure is based on the asymptotic formula proven by Marčenko and Pastur (1967). We notice that the largest sample eigenvalues are severely biased upwards, and the smallest ones downwards. The bias increases in p/n. This phenomenon is very general and is not limited to the identity case. It is similar to the effect observed by Brown (1989) in Monte-Carlo simulations.

Finally, let us remark that the sample eigenvalues $l_i = g_i'Sg_i$ should not be compared to the true eigenvalues $\lambda_i = \gamma_i'\Sigma\gamma_i$, but to $g_i'\Sigma g_i$. We should compare estimated vs. true variance associated with vector g_i. By Theorem 2.2 again, the diagonal elements of $G'\Sigma G$ are even less dispersed than those of $\Gamma'\Sigma\Gamma$. Not only are sample eigenvalues more dispersed than true ones, but they should be less dispersed. This effect is attributable to error in the sample eigenvectors. Intuitively: *Statisticians should shy away from taking a strong stance on extremely small and extremely large eigenvalues, because they know that they have the wrong eigenvectors.* The sample covariance matrix is guilty of taking an unjustifiably strong stance. The optimal linear shrinkage Σ^* corrects for that.

3 Analysis under General Asymptotics

In the previous section, we have shown that Σ^* has an appealing optimality property and fits well in the existing literature. It has only one drawback: it is not a *bona fide* estimator, since it requires hindsight knowledge of four scalar functions of the true (and unobservable) covariance matrix Σ: μ, α^2, β^2 and δ^2. We now address this problem. The idea is that, asymptotically, there exists consistent estimators for μ, α^2, β^2 and δ^2, hence for Σ^* too. At this point we need to choose an appropriate asymptotic framework. Standard asymptotics consider p fixed while n tends to infinity, implying that the optimal shrinkage intensity vanishes in the limit. This would be reasonable for situations where p is very small in comparison to n. However, in the problems of interest us p tends to be of the same order as n and can even be larger. Hence, we consider it more appropriate to use a framework that reflects this condition.

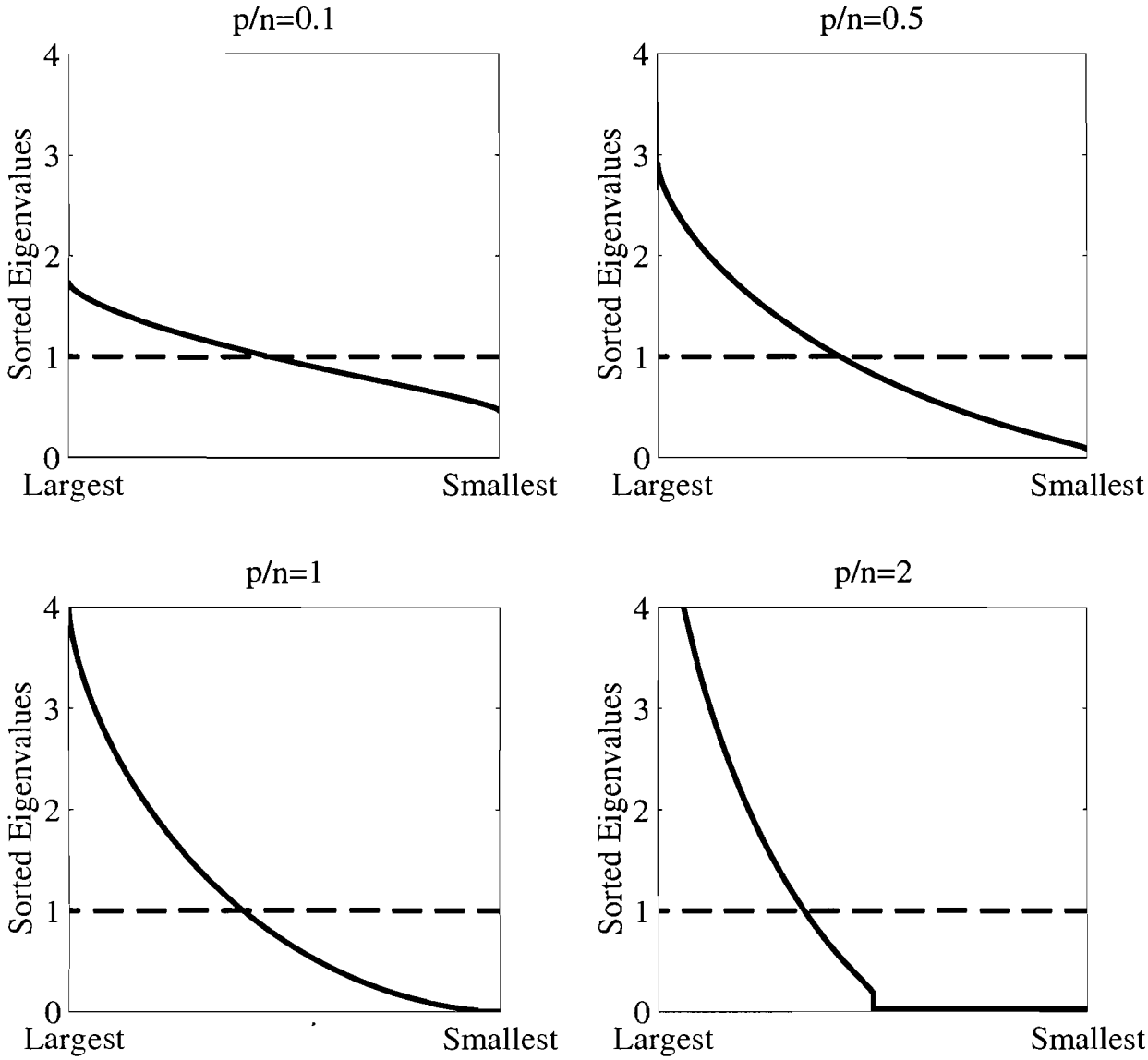

Figure 4: Sample vs. True Eigenvalues. The solid line represents the distribution of the eigenvalues of the sample covariance matrix. Eigenvalues are sorted from largest to smallest, then plotted against their rank. In this case, the true covariance matrix is the identity, that is, the true eigenvalues are all equal to one. The distribution of true eigenvalues is plotted as a dashed horizontal line at one. Distributions are obtained in the limit as the number of observations n and the number of variables p both go to infinity with the ratio p/n converging to a finite positive limit. The four plots correspond to different values of this limit.

This is achieved by allowing the number of variables p to go to infinity at the same speed as the number of observations n. It is called *general* asymptotics.[4] In this framework, the optimal shrinkage intensity generally does not vanish asymptotically but rather it tends to a limiting constant that we will be able to estimate consistently. The idea then is to use the estimated shrinkage intensity in order to arrive at a *bona fide* estimator.

3.1 General Asymptotics

Let $n = 1, 2, \ldots$ index a sequence of statistical models. For every n, X_n is a $p_n \times n$ matrix of n iid observations on a system of p_n random variables with mean zero and covariance matrix Σ_n. The number of variables p_n can change and even go to infinity with the number of observations n, but not too fast.

Assumption 1 *There exists a constant K_1 independent of n such that $p_n/n \leq K_1$.*

Assumption 1 is very weak. It does not require p_n to change and go to infinity, therefore standard asymptotics are included as a particular case. It is not even necessary for the ratio p_n/n to converge to any limit.

Decompose the covariance matrix into eigenvalues and eigenvectors: $\Sigma_n = \Gamma_n \Lambda_n \Gamma_n^{\mathrm{t}}$, where Λ_n is a diagonal matrix, and Γ_n a rotation matrix. The diagonal elements of Λ_n are the eigenvalues $\lambda_1^n, \ldots, \lambda_{p_n}^n$, and the columns of Γ_n are the eigenvectors $\gamma_1^n, \ldots, \gamma_{p_n}^n$. $Y_n = \Gamma_n^{\mathrm{t}} X_n$ is a $p_n \times n$ matrix of n iid observations on a system of p_n uncorrelated random variables that spans the same space as the original system. We impose restrictions on the higher moments of Y_n. Let $(y_{11}^n, \ldots, y_{p_n 1}^n)^{\mathrm{t}}$ denote the first column of the matrix Y_n.

Assumption 2 *There exists a constant K_2 independent of n such that* $\dfrac{1}{p_n} \displaystyle\sum_{i=1}^{p_n} \mathrm{E}[(y_{i1}^n)^8] \leq K_2$.

Assumption 3

$$\lim_{n \to \infty} \frac{p_n^2}{n^2} \times \frac{\sum_{(i,j,k,l) \in Q_n} (\mathrm{Cov}[y_{i1}^n y_{j1}^n, y_{k1}^n y_{l1}^n])^2}{\text{Cardinal of } Q_n} = 0,$$

where Q_n denotes the set of all the quadruples that are made of four distinct integers between 1 and p_n.

Assumption 2 states that the eighth moment is bounded (on average). Assumption 3 states that products of uncorrelated random variables are themselves uncorrelated (on average, in the limit). In the case where general asymptotics degenerate into standard asymptotics ($p_n/n \to 0$), Assumption 3 is trivially verified as a consequence of Assumption 2. Assumption 3 is verified when random variables are normally or even elliptically distributed, but it is much weaker than that. Assumptions 1-3 are implicit throughout the paper.

Our matrix norm is based on the Frobenius norm.

[4] To the best of our knowledge, the framework of general asymptotics has not been used before to improve over the sample covariance matrix, but only to characterize the distribution of its eigenvalues, as in Silverstein (1994).

Definition 1 *The norm of the p_n-dimensional matrix A is: $\|A\|_n^2 = f(p_n)\,\mathrm{tr}(AA^t)$, where $f(p_n)$ is a scalar function of the dimension.*

It defines a quadratic form on the linear space of p_n-dimensional symmetric matrices whose associated inner product is: $A_1 \circ_n A_2 = f(p_n)\mathrm{tr}(A_1 A_2^t)$.

The behavior of $\|\cdot\|_n$ across dimensions is controlled by the function $f(\cdot)$. The norm $\|\cdot\|_n$ is used mainly to define a notion of consistency. A given estimator will be called consistent if the norm of its difference with the true covariance matrix goes to zero (in quadratic mean) as n goes to infinity. If p_n remains bounded, then all positive functions $f(\cdot)$ generate equivalent notions of consistency. But this particular case similar to standard asymptotics is not very representative. If p_n (or a subsequence) goes to infinity, then the choice of $f(\cdot)$ becomes much more important. If $f(p_n)$ is too large (small) as p_n goes to infinity, then it will define too strong (weak) a notion of consistency. $f(\cdot)$ must define the notion of consistency that is "just right" under general asymptotics.

Our solution is to define a *relative* norm. The norm of a p_n-dimensional matrix is divided by the norm of a benchmark matrix of the same dimension p_n. The benchmark must be chosen carefully. For lack of any other attractive candidate, we take the identity matrix as benchmark. Therefore, by convention, the identity matrix has norm one in every dimension. This determines the function $f(\cdot)$ uniquely as follows.

Definition 2 *The scalar coefficient left unspecified in Definition 1 is: $f(p_n) = 1/p_n$.*

Intuitively, it seems that the norm of the identity matrix should remain bounded away from zero and from infinity as its dimension goes to infinity. All choices of $f(\cdot)$ satisfying this property would define equivalent notions of consistency. Therefore our particular norm is equivalent to any norm that would make sense under general asymptotics.

An example might help familiarize the reader with Definitions 1-2. Let A_n be the $p_n \times p_n$ matrix with one in its top left entry and zeros everywhere else. Let Z_n be the $p_n \times p_n$ matrix with zeros everywhere (i.e. the null matrix). A_n and Z_n differ in a way that is independent of p_n: the top left entry is not the same. Yet their squared distance $\|A_n - Z_n\|^2 = 1/p_n$ depends on p_n. This apparent paradox has an intuitive resolution. A_n and Z_n disagree on the first dimension, but they agree on the $p_n - 1$ others. The importance of their disagreement is relative to the extent of their agreement. If $p_n = 1$, then A_n and Z_n have nothing in common, and their distance is 1. If $p_n \to \infty$, then A_n and Z_n have almost everything in common, and their distance goes to 0. Thus, disagreeing on one entry can either be important (if this entry is the only one) or negligible (if this entry is just one among a large number of others).

3.2 The Behavior of the Sample Covariance Matrix

Define the sample covariance matrix $S_n = X_n X_n^t/n$. We follow the notation of Section 2, except that we add the subscript n to signal that all results hold asymptotically. Thus, we have: $\mu_n = \Sigma_n \circ_n I_n$, $\alpha_n^2 = \|\Sigma_n - \mu_n I_n\|_n^2$, $\beta_n^2 = \mathrm{E}[\|S_n - \Sigma_n\|_n^2]$, and $\delta_n^2 = \mathrm{E}[\|S_n - \mu_n I_n\|_n^2]$. These four scalars are well behaved asymptotically.

Lemma 3.1 *μ_n, α_n^2, β_n^2 and δ_n^2 remain bounded as $n \to \infty$.*

They can go to zero in special cases, but in general they do not, in spite of the division by p_n in the definition of the norm. The proofs of all the technical results of Section 3 are in Appendix A.

The most basic question is whether the sample covariance matrix is consistent under general asymptotics. Specifically, we ask whether S_n converges in quadratic mean to the true covariance matrix, that is, whether β_n^2 vanishes. In general, the answer is no, as shown below.[5]

Theorem 3.1 *Define* $\theta_n^2 = \mathrm{Var}\left[\frac{1}{p_n}\sum_{i=1}^{p_n}(y_{i1}^n)^2\right]$. *$\theta_n^2$ is bounded as $n \to \infty$, and we have:* $\lim_{n\to\infty} \mathrm{E}[\|S_n - \Sigma_n\|_n^2] - \frac{p_n}{n}(\mu_n^2 + \theta_n^2) = 0$.

Theorem 3.1 shows that the expected loss of the sample covariance matrix $\mathrm{E}[\|S_n - \Sigma_n\|_n^2]$ is bounded, but it is at least of the order of $\frac{p_n}{n}\mu_n^2$, which does not usually vanish. Therefore the sample covariance matrix is not consistent under general asymptotics, except in special cases.

The first special case is when $p_n/n \to 0$. For example, under standard asymptotics, p_n is fixed, and it is well-known that the sample covariance matrix is consistent. Theorem 3.1 shows that consistency extends to cases where p_n is not fixed, not even necessarily bounded, as long as it is of order $o(n)$. The second special case is when $\mu_n^2 \to 0$ and $\theta_n^2 \to 0$. $\mu_n^2 \to 0$ implies that most of the p_n random variables have vanishing variances, i.e. they are asymptotically degenerate. The number of random variables escaping degeneracy must be negligible with respect to n. This is like the previous case, except that the $o(n)$ nondegenerate random variables can now be augmented with $O(n)$ degenerate ones. Overall, a loose condition for the consistency of the sample covariance matrix under general asymptotics is that the number of nondegenerate random variables be negligible with respect to the number of observations.

If the sample covariance matrix is not consistent under general asymptotics, it is because of its off-diagonal elements. Granted, the error on each one of them vanishes, but their number grows too fast. The accumulation of a large number of small errors off the diagonal prevents the sample covariance matrix from being consistent. By contrast, the contribution of the errors on the diagonal is negligible. This is apparent from the proof of Theorem 3.1. After all, it should in general not be possible to consistently estimate $p_n(p_n+1)/2$ parameters from a data set of $n\,p_n$ random realizations if these two numbers are of the same order of magnitude. For this reason, we believe that there does not exist any consistent estimator of the covariance matrix under general asymptotics.

Theorem 3.1 also shows what factors determine the error of S_n. The first factor is the ratio p_n/n. It measures deviation from standard asymptotics. People often figure out whether they can use asymptotics by checking whether they have enough observations, but in this case it would be unwise: it is the ratio of observations to variables that needs to be big. 200 observations might seem like a lot, but it is not nearly enough if there are 100 variables: it

[5]The results stated in Theorem 3.1 and Lemmata 3.2 and 3.3 are related to special cases of a general result proven by Yin (1986). But we work under weaker assumptions than he does. Also, his goal is to find the distribution of the eigenvalues of the sample covariance matrix, while ours is to find an improved estimator of the covariance matrix.

would be about as bad as using 2 observations to estimate the variance of 1 random variable! The second factor μ_n^2 simply gives the scale of the problem. The third factor θ_n^2 measures covariance between the squared variables over and above what is implied by covariance between the variables themselves. For example, θ_n^2 is zero in the normal case, but usually positive in the elliptic case. Intuitively, a "cross-sectional" law of large numbers could make the variance of $p_n^{-1} \sum_{i=1}^{p_n} y_{i1}^2$ vanish asymptotically as $p_n \to \infty$ if the y_{i1}^2's were sufficiently uncorrelated with one another. But Assumption 3 is too weak to ensure that, so in general θ_n^2 is not negligible, which might be more realistic sometimes.

This analysis enables us to answer another basic question: When does shrinkage matter? Remember that $\beta_n^2 = \mathrm{E}[\|S_n - \Sigma_n\|_n^2]$ denotes the error of the sample covariance matrix, and that $\delta_n^2 = \mathrm{E}[p_n^{-1} \sum_{i=1}^{p_n} (l_i^n - \mu_n)^2]$ denotes the cross-sectional dispersion of the sample eigenvalues $l_1^n, \ldots, l_{p_n}^n$ around the expectation of their grand mean $\mu_n = \mathrm{E}[\sum_{i=1}^{p_n} l_i^n / p_n]$. Theorem 2.1 states that shrinkage matters unless the ratio β_n^2/δ_n^2 is negligible, but this answer is rather abstract. Theorem 3.1 enables us to rewrite it in more intuitive terms. Ignoring the presence of θ_n^2, the error of the sample covariance matrix β_n^2 is asymptotically close to $\frac{p_n}{n}\mu_n^2$. Therefore *shrinkage matters unless the ratio of variables to observations p_n/n is negligible with respect to δ_n^2/μ_n^2, which is a scale-free measure of cross-sectional dispersion of sample eigenvalues.* Figure 5 provides a graphical illustration. This constitutes an easy diagnostic test to reveal whether

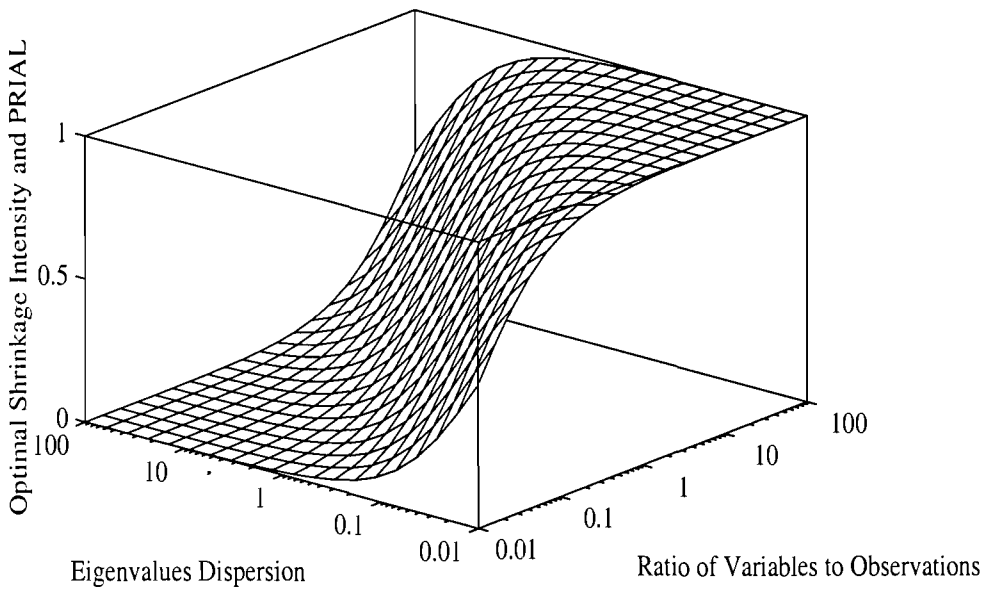

Figure 5: Optimal Shrinkage Intensity and PRIAL as Function of Eigenvalues Dispersion and the Ratio of Variables to Observations. Note that eigenvalues dispersion is measured by the scale-free ratio δ_n^2/μ_n^2.

our shrinkage method can substantially improve upon the sample covariance matrix. In our opinion, there are many important practical situations where shrinkage does matter according to this criterion. Also, it is rather exceptional for gains from shrinkage to be as large as p_n/n, because most of the time (for example in estimation of the mean) they are of the much smaller order $1/n$.

3.3 A Consistent Estimator for Σ_n^*

Σ_n^* is not a *bona fide* estimator because it depends on the true covariance matrix Σ_n, which is unobservable. Fortunately, computing Σ_n^* does not require knowledge of the whole matrix Σ_n, but only of four scalar functions of Σ_n: μ_n, α_n^2, β_n^2 and δ_n^2. Given the size of the data set ($p_n \times n$), we cannot estimate all of Σ_n consistently, but we can estimate the optimal shrinkage target, the optimal shrinkage intensity, and even Σ_n^* itself consistently.

For μ_n, a consistent estimator is its sample counterpart.

Lemma 3.2 *Define $m_n = S_n \circ_n I_n$. Then $\mathrm{E}[m_n] = \mu_n$ for all n, and $m_n - \mu_n$ converges to zero in quartic mean (fourth moment) as n goes to infinity.*

It implies that $m_n^2 - \mu_n^2 \overset{\text{q.m.}}{\to} 0$ and $m_n - \mu_n \overset{\text{q.m.}}{\to} 0$, where $\overset{\text{q.m.}}{\to}$ denotes convergence in *quadratic* mean as $n \to \infty$. A consistent estimator for $\delta_n^2 = \mathrm{E}[\|S_n - \mu_n I_n\|_n^2]$ is also its sample counterpart.

Lemma 3.3 *Define $d_n^2 = \|S_n - m_n I_n\|_n^2$. Then $d_n^2 - \delta_n^2 \overset{\text{q.m.}}{\to} 0$.*

Now let the $p_n \times 1$ vector $x_{\cdot k}^n$ denote the k^{th} column of the observations matrix X_n, for $k = 1, \ldots, n$. $S_n = n^{-1} X_n X_n^{\mathrm{t}}$ can be rewritten as $S_n = n^{-1} \sum_{k=1}^n x_{\cdot k}^n (x_{\cdot k}^n)^{\mathrm{t}}$. S_n is the average of the matrices $x_{\cdot k}^n (x_{\cdot k}^n)^{\mathrm{t}}$. Since the matrices $x_{\cdot k}^n (x_{\cdot k}^n)^{\mathrm{t}}$ are iid across k, we can estimate the error $\beta_n^2 = \mathrm{E}[\|S_n - \Sigma_n\|_n^2]$ of their average by seeing how far each one of them deviates from the average.

Lemma 3.4 *Define $\bar{b}_n^2 = \dfrac{1}{n^2} \sum_{k=1}^n \|x_{\cdot k}^n (x_{\cdot k}^n)^{\mathrm{t}} - S_n\|_n^2$ and $b_n^2 = \min(\bar{b}_n^2, d_n^2)$. Then $\bar{b}_n^2 - \beta_n^2 \overset{\text{q.m.}}{\to} 0$ and $b_n^2 - \beta_n^2 \overset{\text{q.m.}}{\to} 0$.*

We introduce the constrained estimator b_n^2 because $\beta_n^2 \leq \delta_n^2$ by Lemma 2.1. In general, this constraint is rarely binding. But it ensures that the following estimator of α_n^2 is nonnegative.

Lemma 3.5 *Define $a_n^2 = d_n^2 - b_n^2$. Then $a_n^2 - \alpha_n^2 \overset{\text{q.m.}}{\to} 0$.*

The next step of the strategy is to replace the unobservable scalars in the formula defining Σ_n^* with consistent estimators, and to show that the asymptotic properties are unchanged. This yields our *bona fide* estimator of the covariance matrix:

$$\boxed{S_n^* = \frac{b_n^2}{d_n^2} m_n I_n + \frac{a_n^2}{d_n^2} S_n.} \tag{14}$$

The next theorem shows that S_n^* has the same asymptotic properties as Σ_n^*. Thus, we can neglect the error that we introduce when we replace the unobservable parameters μ_n, α_n^2, β_n^2 and δ_n^2 by estimators.

Theorem 3.2 *S_n^* is a consistent estimator of Σ_n^*, i.e. $\|S_n^* - \Sigma_n^*\|_n \overset{\text{q.m.}}{\to} 0$. As a consequence, S_n^* has the same asymptotic expected loss (or risk) as Σ_n^*, i.e. $\mathrm{E}[\|S_n^* - \Sigma_n\|_n^2] - \mathrm{E}[\|\Sigma_n^* - \Sigma_n\|_n^2] \to 0$.*

This justifies our studying the properties of Σ_n^* in Section 2 "as if" it was a *bona fide* estimator.

It is interesting to recall the Bayesian interpretation of Σ_n^* (see Section 2.2). From this point of view, S_n^* is an *empirical* Bayesian estimator. Empirical Bayesians (see e.g. Frost and Savarino, 1986) ignore the fact that their prior contains estimation error because it comes from the data. Usually, this is done without any rigorous justification, and it requires sophisticated "judgement" to pick an empirical Bayesian prior whose estimation error is "not too" damaging. Here, we treat this issue rigorously instead: We give a set of conditions (Assumptions 1-3) under which it is legitimate to neglect the estimation error of our empirical Bayesian prior.

Finally, it is possible to estimate the expected quadratic loss of Σ_n^* and S_n^* consistently.

Lemma 3.6 $\mathrm{E}\left[\left|\frac{a_n^2 b_n^2}{d_n^2} - \frac{\alpha_n^2 \beta_n^2}{\delta_n^2}\right|\right] \to 0$

3.4 Optimality Property of the Estimator S_n^*

The final step of our strategy is to demonstrate that S_n^*, which we obtained as a consistent estimator for Σ_n^*, possesses an important optimality property. We already know that Σ_n^* (hence S_n^* in the limit) is optimal among the linear combinations of the identity and the sample covariance matrix with *nonrandom* coefficients (see Theorem 2.1). This is interesting, but only mildly so, because it excludes the other linear shrinkage estimators with random coefficients. In this section, we show that S_n^* is still optimal within a bigger class: the linear combinations of I_n and S_n with *random* coefficients. This class includes both the linear combinations that represent *bona fide* estimators, and those with coefficients that require hindsight knowledge of the true (and unobservable) covariance matrix Σ_n.

Let Σ_n^{**} denote the linear combination of I_n and S_n with minimum quadratic loss. It solves:

$$\begin{aligned} &\min_{\rho_1,\rho_2} \|\Sigma_n^{**} - \Sigma_n\|_n^2 \\ &\text{s.t.} \quad \Sigma_n^{**} = \rho_1 I_n + \rho_2 S_n. \end{aligned} \tag{15}$$

In contrast to the optimization problem in Theorem 2.1 with solution Σ_n^*, here we minimize the loss instead of the expected loss, and we allow the coefficients ρ_1 and ρ_2 to be random. It turns out that the formula for Σ_n^{**} is a function of Σ_n, therefore Σ_n^{**} does not constitute a *bona fide* estimator. By construction, Σ_n^{**} has lower loss than Σ_n^* and S_n^* a.s., but asymptotically it makes no difference.

Theorem 3.3 *S_n^* is a consistent estimator of Σ_n^{**}, i.e. $\|S_n^* - \Sigma_n^{**}\|_n \overset{\text{q.m.}}{\to} 0$. As a consequence, S_n^* has the same asymptotic expected loss (or risk) as Σ_n^{**}, that is, $\mathrm{E}[\|S_n^* - \Sigma_n\|_n^2] - \mathrm{E}[\|\Sigma_n^{**} - \Sigma_n\|_n^2] \to 0$.*

Both Σ_n^* and Σ_n^{**} have the same asymptotic properties as S_n^*, therefore they also have the same asymptotic properties as each other. The most important result of this paper is the following: The *bona fide* estimator S_n^* has uniformly minimum quadratic risk asymptotically among all the linear combinations of the identity with the sample covariance matrix, including those that are *bona fide* estimators, and even those that use hindsight knowledge of the true covariance matrix.

Theorem 3.4 *For any sequence of linear combinations $\widehat{\Sigma}_n$ of the identity and the sample covariance matrix, the estimator S_n^* defined in Equation (14) verifies:*

$$\lim_{N\to\infty} \inf_{n\geq N} (\mathrm{E}[\|\widehat{\Sigma}_n - \Sigma_n\|_n^2] - \mathrm{E}[\|S_n^* - \Sigma_n\|_n^2]) \geq 0. \tag{16}$$

In addition, every $\widehat{\Sigma}_n$ that performs as well as S_n^* is identical to S_n^* in the limit:

$$\lim_{n\to\infty} (\mathrm{E}[\|\widehat{\Sigma}_n - \Sigma_n\|_n^2] - \mathrm{E}[\|S_n^* - \Sigma_n\|_n^2]) = 0 \iff \|\widehat{\Sigma}_n - S_n^*\|_n \stackrel{\text{q.m.}}{\to} 0. \tag{17}$$

Thus it is legitimate to say that S_n^* is an asymptotically optimal linear shrinkage estimator of the covariance matrix with respect to quadratic loss under general asymptotics. Typically, only maximum likelihood estimators have such a sweeping optimality property, so we believe that this result is unique in shrinkage theory.

Yet another distinctive feature of S_n^* is that, to the best of our knowledge, it is the only estimator of the covariance matrix to retain a rigorous justification when the number of variables p_n exceeds the number of observations n. Not only that, but S_n^* is guaranteed to be always invertible, even in the case $p_n > n$, where rank deficiency makes the sample covariance matrix singular. Estimating the inverse covariance matrix when variables outnumber observations is sometimes dismissed as impossible, but the existence of $(S_n^*)^{-1}$ certainly proves otherwise. The following theorem shows that S_n^* is usually well-conditioned.

Theorem 3.5 *Assume that the condition number of the true covariance matrix Σ_n is bounded, and that the normalized variables $y_{i1}/\sqrt{\lambda_i}$ are iid across $i = 1, \ldots, n$. Then the condition number of the estimator S_n^* is bounded in probability.*

This result follows from powerful results proven recently by probabilists (Bai and Yin, 1993). If the cross-sectional iid assumption is violated, it does not mean that the condition number goes to infinity, but rather that it is technically too difficult to find out anything about it.

Interestingly, there is one case where the estimator S_n^* is even better-conditioned than the true covariance matrix Σ_n: if the ill-conditioning of Σ_n comes from eigenvalues close to zero (multicollinearity in the variables) and the ratio of variables to observations p_n/n is not negligible. In this case, S_n^* is well-conditioned because the sample observations do not provide enough information to update our prior belief that there is no multicollinearity.

4 Monte-Carlo Simulations

The goal is to compare the expected loss (or risk) of various estimators across a wide range of situations. The benchmark is the expected loss of the sample covariance matrix. We report the Percentage Relative Improvement in Average Loss of S^*, defined as: $\mathrm{PRIAL}(S^*) = (\mathrm{E}[\|S - \Sigma\|^2] - \mathrm{E}[\|S^* - \Sigma\|^2])/\mathrm{E}[\|S - \Sigma\|^2] \times 100$. The subscript n is omitted for brevity, since no confusion is possible. If the PRIAL is positive (negative), then S^* performs better (worse) than S. The PRIAL of the sample covariance matrix S is zero by definition. The PRIAL cannot exceed 100%. We compare the PRIAL of S^* to the PRIAL of other estimators from finite sample decision theory. There are many estimators worthy of investigation, and we cannot possibly study all the interesting ones.

4.1 Other Estimators

Haff (1980) introduces an estimator with an empirical Bayesian inspiration. Like S^*, it is a linear combination of the sample covariance matrix and the identity. The difference lies in the coefficients of the combination. Haff's coefficients do not depend on the observations X, only on p and n. If the criterion is the mean squared error, Haff's approach suggests:

$$\widehat{S}_{\text{EB}} = \frac{pn - 2n - 2}{pn^2} m_{\text{EB}} I + \frac{n}{n+1} S \tag{18}$$

with $m_{\text{EB}} = [\det(S)]^{1/p}$. When $p > n$ we take $m_{\text{EB}} = m$ because the regular formula would yield zero. The initials EB stand for empirical Bayesian.

Stein (1975) proposes an estimator that keeps the eigenvectors of the sample covariance matrix and replaces its eigenvalues $l_1, \ldots, l_p$ by:

$$nl_i \Bigg/ \left(n - p + 1 + 2l_i \sum_{\substack{j=1 \\ j \neq i}}^{p} \frac{1}{l_i - l_j} \right) \qquad i = 1, \ldots, p. \tag{19}$$

These corrected eigenvalues need neither be positive nor in the same order as sample eigenvalues. To prevent this from happening, an ad-hoc procedure called isotonic regression is applied before recombining corrected eigenvalues with sample eigenvectors.[6] Haff (1982) independently obtains a closely related estimator. In any given simulation, we call $\widehat{S}_{\text{SH}}$ the better performing estimator of the two. The other one is not reported. The initials SH stand for Stein and Haff.[7]

Stein (1982) and Dey and Srinivasan (1985) both derive the same estimator. Under a certain loss function, it is minimax, which means that no other estimator has lower worst-case error. The minimax criterion is sometimes criticized as overly pessimistic, since it looks at the worst case only. This estimator preserves sample eigenvectors and replaces sample eigenvalues by:

$$\frac{n}{n + p + 1 - 2i} \widetilde{\lambda}_i, \tag{20}$$

where sample eigenvalues $l_1, \ldots, l_p$ are sorted in descending order. We call this estimator $\widehat{S}_{\text{MX}}$, where the initials MX stand for minimax.

When the number of variables p is very large, S^* and S take much less time to compute than $\widehat{S}_{\text{EB}}$, $\widehat{S}_{\text{SH}}$ and $\widehat{S}_{\text{MX}}$, because they do not need eigenvalues and determinants. Indeed the number and nature of operations needed to compute S^* are of the same order as for S. It can be an enormous advantage in practice. The only seemingly slow step is the estimation of β^2, but it can be accelerated by writing:

$$b^2 = \frac{1}{pn} \sum_{i=1}^{p} \sum_{j=1}^{p} \left[\frac{1}{n} \left(X^{\wedge 2} \right) \left(X^{\wedge 2} \right)^{\text{t}} \right]_{ij} - \frac{1}{pn} \sum_{i=1}^{p} \sum_{j=1}^{p} \left[\left(\frac{1}{n} X X^{\text{t}} \right)^{\wedge 2} \right]_{ij} \tag{21}$$

where $[\cdot]_{ij}$ denotes the entry (i, j) of a matrix and the symbol $^{\wedge}$ denotes elementwise exponentiation, i.e. $[A^{\wedge 2}]_{ij} = ([A]_{ij})^2$ for any matrix A.

[6] Intuitively, isotonic regression restores the ordering by assigning the same value to a subsequence of corrected eigenvalues that would violate it. Lin and Perlman (1985) explain it in detail.

[7] When $p > n$ some of the terms $\widetilde{\lambda}_i - \widetilde{\lambda}_j$ in formula (19) result in a division by zero. We just ignore them. Nonetheless, when p is too large compared to n, the isotonic regression does not converge. In this case $\widehat{S}_{\text{SH}}$ does not exist.

4.2 Results

The random variables used in the simulations are normally distributed. The true covariance matrix Σ is diagonal without loss of generality. Its eigenvalues are drawn according to a log-normal distribution. Their grand mean μ is set equal to one without loss of generality. We let their cross-sectional dispersion α^2 vary around the central value $1/2$. We let the ratio p/n vary around the central value $1/2$. Finally, we let the product pn vary around the central value 800. We study the influence of α^2, p/n and pn separately. When one parameter moves, the other two remain fixed at their central values.

The asymptotic PRIAL of S^* implied by Theorems 2.1, 3.1 and 3.2 is $\frac{p/n}{p/n+\alpha^2} \times 100$. This is the "speed of light" that we would attain if we knew the true parameters μ, α^2, β^2, δ^2, instead of having to estimate them. When all three parameters are fixed at their repective central values, we get the results in Table 2. "Risk" means the average loss over

Estimator	S	S^*	$\widehat{S}_{\text{EB}}$	$\widehat{S}_{\text{SH}}$	$\widehat{S}_{\text{MX}}$
Risk	0.5372	0.2723	0.5120	0.3076	0.3222
Standard Error on Risk	(0.0033)	(0.0013)	(0.0031)	(0.0014)	(0.0014)
PRIAL	0.0%	49.3%	4.7%	42.7%	40.0%

Table 2: Result of 1,000 Monte-Carlo Simulations for Central Parameter Values.

1,000 simulations. For the central values of the parameters, the asymptotic PRIAL of S^* is equal to 50%, and its simulated PRIAL is 49.3%. Therefore asymptotic behavior is almost attained in this case for $p = 20$ and $n = 40$. S^* improves substantially over S and $\widehat{S}_{\text{EB}}$, and moderately over $\widehat{S}_{\text{SH}}$ and $\widehat{S}_{\text{MX}}$.

When we increase p/n from zero to infinity, the asymptotic PRIAL of S^* increases from 0% to 100% with an "S" shape. Figure 6 confirms this.[8] S^* always has lower risk than S and $\widehat{S}_{\text{EB}}$. It usually has slightly lower risk than $\widehat{S}_{\text{SH}}$ and $\widehat{S}_{\text{MX}}$. $\widehat{S}_{\text{SH}}$ is not defined for high values of p/n. $\widehat{S}_{\text{MX}}$ performs slightly better than S^* for the highest values of p/n. This may be due to the fact that S^* is not close to its asymptotic performance for values of n below 10.

When we increase α^2 from zero to infinity, the asymptotic PRIAL of S^* decreases from 100% to 0% with a reverse "S" shape. Figure 7 confirms this. S^* has lower mean squared error than S always, and than $\widehat{S}_{\text{EB}}$ almost always. S^* always has lower mean squared error than $\widehat{S}_{\text{SH}}$ and $\widehat{S}_{\text{MX}}$. When α^2 gets too large, $\widehat{S}_{\text{SH}}$ and $\widehat{S}_{\text{MX}}$ perform worse than the sample covariance matrix. The reason is that true eigenvalues are very dispersed, and they shrink sample eigenvalues together too much. This may be due to the fact that $\widehat{S}_{\text{SH}}$ and $\widehat{S}_{\text{MX}}$ were originally derived under another loss function than the Frobenius norm. It is very reassuring that, even in a case where some of its competitors perform much worse than S, S^* performs at least as well as S.

When we increase pn from zero to infinity, we should see the PRIAL of S^* converge to its asymptotic value of 50%. Figure 8 confirms this. S^* always has lower risk than S and $\widehat{S}_{\text{EB}}$. It

[8] Corresponding tables of results are available from the authors upon request. Standard errors on simulated risk have the same order of magnitude as in Table 2.

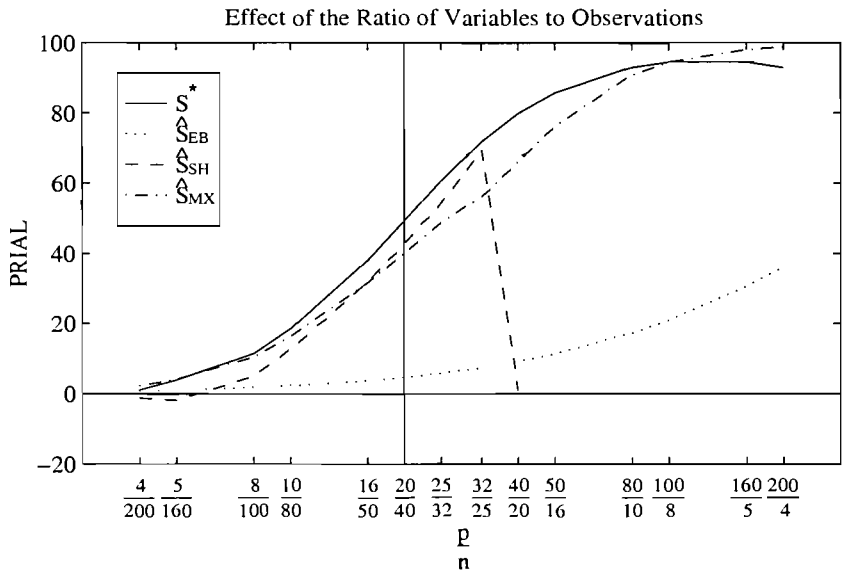

Figure 6: Effect of the Ratio of Number of Variables to Number of Observations on the PRIAL. $\widehat{S}_{\mathrm{SH}}$ is not defined when $p/n > 2$ because the isotonic regression does not converge.

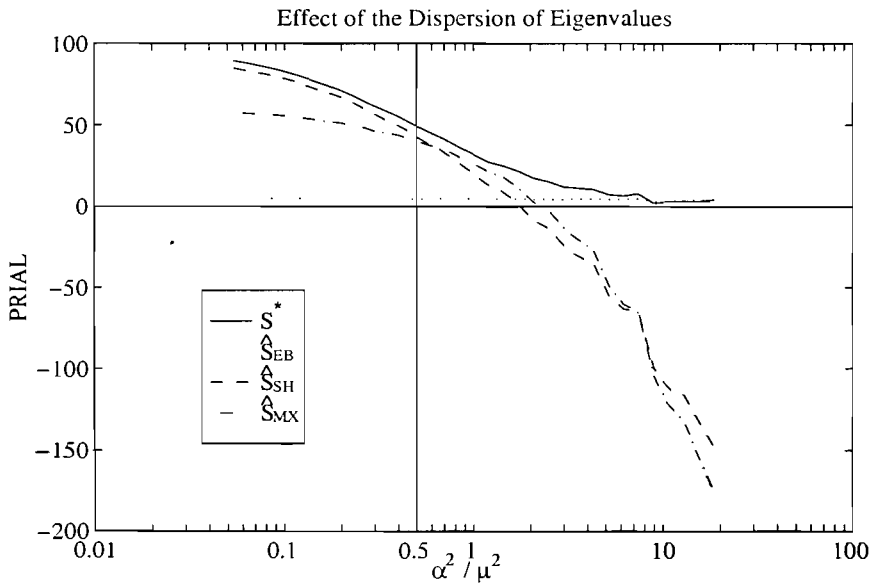

Figure 7: Effect of the Dispersion of Eigenvalues on the PRIAL.

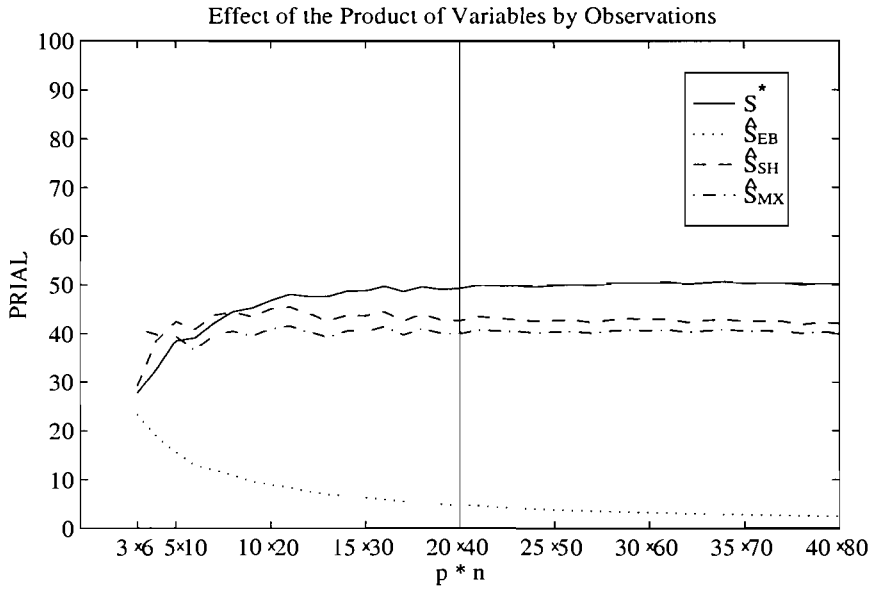

Figure 8: Effect of the Product of Variables by Observations on the PRIAL.

has moderately lower risk than $\widehat{S}_{\text{SH}}$ and $\widehat{S}_{\text{MX}}$, except when n is below 20. When n is below 20, S^* performs slightly worse than $\widehat{S}_{\text{SH}}$ and moderately worse than $\widehat{S}_{\text{MX}}$, but still substantially better than S and $\widehat{S}_{\text{EB}}$.

Simulations not reported here study departures from normality. These departures have very little impact on the above results. In relative terms, S and $\widehat{S}_{\text{EB}}$ appear to suffer the most; then $\widehat{S}_{\text{SH}}$ and $\widehat{S}_{\text{MX}}$; and S^* appears to suffer the least.

We draw the following conclusions from these simulations. The asymptotic theory developed in Section 3 approximates finite sample behavior well, as soon as n and p become of the order of twenty. S^* improves over the sample covariance matrix in every one of the situations simulated, and usually by a lot. It also improves over $\widehat{S}_{\text{EB}}$ in almost every situation simulated, and usually by a lot too. S^* never performs substantially worse than $\widehat{S}_{\text{SH}}$ and $\widehat{S}_{\text{MX}}$, often performs about as well or slightly better, and in some cases does substantially better. In the cases where $\widehat{S}_{\text{SH}}$ or $\widehat{S}_{\text{MX}}$ do better, it is attributable to small sample size (less than ten).[9]

4.3 Condition Number

This section studies the condition number of the estimator S^* in finite sample. The procedure for the Monte-Carlo simulations is the same as in Section 4.2, except that we do not compute the other estimators $\widehat{S}_{\text{EB}}$, $\widehat{S}_{\text{SH}}$ and $\widehat{S}_{\text{MX}}$. Figures 9, 10 and 11 plot the behavior of the condition number when p/n varies, when α^2/μ^2 varies, and when pn varies, respectively. The graphs show the average condition number over 1,000 replications for the sample covariance matrix S and for the improved estimator S^*. They also show the condition number of the true covariance matrix for comparison. We can see that the sample covariance matrix

[9]We acknowledge that $\widehat{S}_{\text{SH}}$ and $\widehat{S}_{\text{MX}}$ were designed with another criterion than the Frobenius norm in mind. Our conclusions say nothing about performance under any other criterion. Nonetheless, the Frobenius norm is an important criterion.

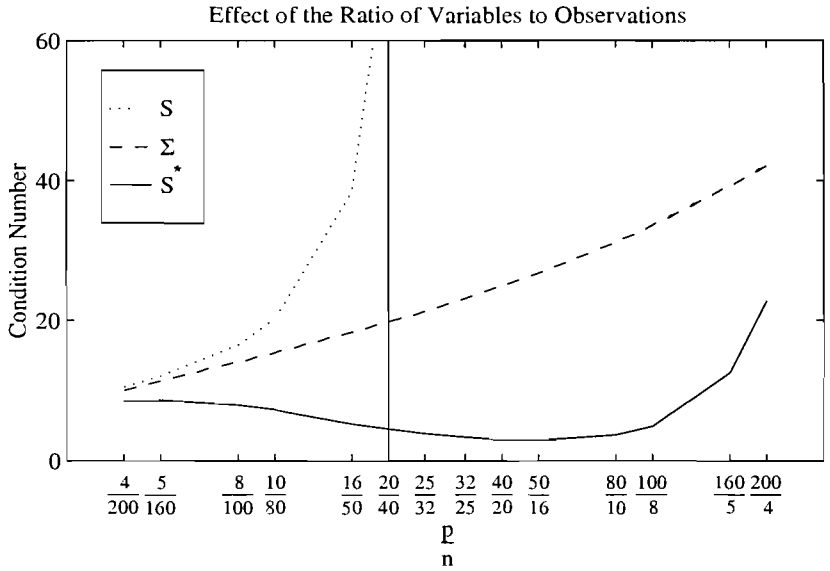

Figure 9: Effect of the Ratio of Number of Variables to Number of Observations on the Condition Number.

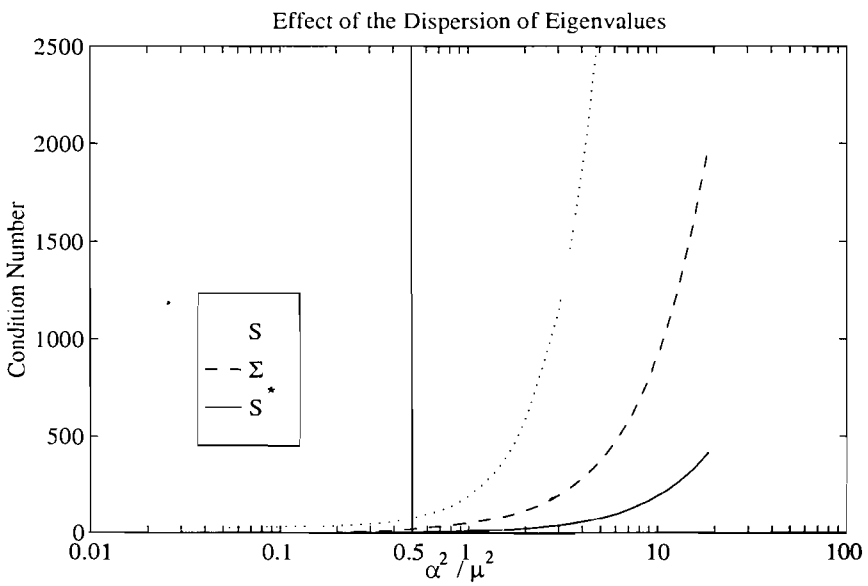

Figure 10: Effect of the Dispersion of Eigenvalues on the Condition Number. Even for small dispersions, the condition number of S is 3 to 10 times bigger than the true condition number, while the condition number of S^* is 2 to 5 times smaller than the true one.

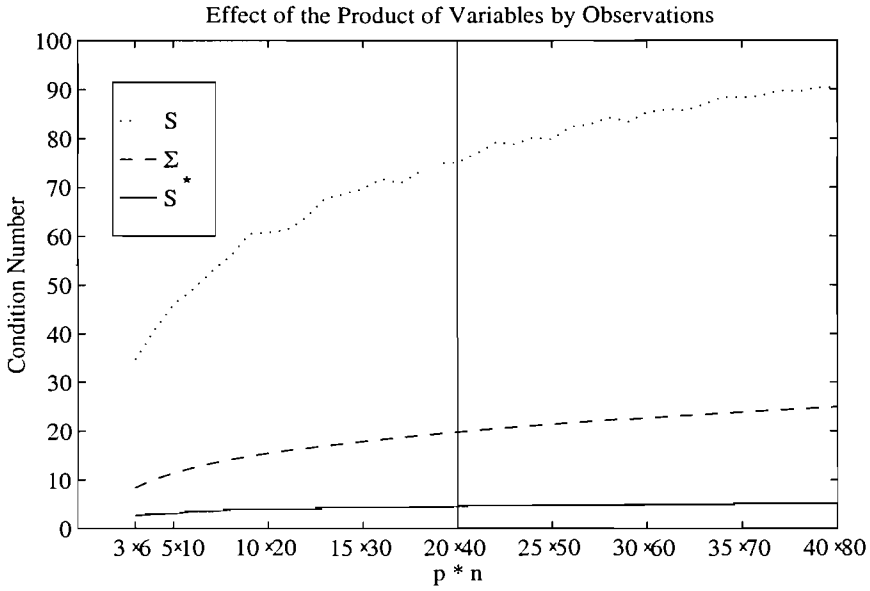

Figure 11: Effect of the Product of Variables by Observations on the Condition Number.

is always worse-conditioned than the true covariance matrix, while our estimator is always better-conditioned. This suggests that the asymptotic result proven in Theorem 3.5 holds well in finite sample.

5 Conclusions

In this paper we have discussed the estimation of large dimensional covarinance matrices where the number of (iid) variables is not small compared to the sample size. It is well-known that in such situations the usual estimator, the sample covariance matrix, is ill-conditioned and may not even be invertible. The approach suggested is to shrink the sample covariance matrix towards the identity matrix, which means to consider a convex linear combination of these two matrices.

The practical problem is to determine the shrinkage intensity, that is, the amount of shrinkage of the sample covariance matrix towards the identity matrix. To solve this problem, we considered a general asymptotics framework where the number of variables is allowed to tend to infinity with the sample size. It was seen that under mild conditions the optimal shrinkage intensity then tends to a limiting constant; here, optimality is meant with respect to a quadratic loss function based on the Frobenius norm. It was shown that the asymptotically optimal shrinkage intensity can be estimated consistently, which leads to a feasible estimator.

Both the asymptotic results and the extensive Monte-Carlo simulations presented in this paper indicate that the suggested shrinkage estimator can serve as an all-purpose alternative to the sample covariance matrix. It has smaller risk *and* is better-conditioned. This is especially true when the dimension of the covariance matrix is large.

Directions for future research include: applying this technique to portfolio selection, Generalized Least Squares (GLS) regressions, and the General Method of Moments (GMM); finding

the optimal *non*-linear way to shrink sample eigenvalues towards their grand mean; taking other shrinkage targets than the identity; minimizing alternative loss functions; relaxing the assumption that observations are iid across time; characterizing the limiting distribution of this estimator and its inverse; applying this technique to the estimation of the vector of means or the vector of variances. The authors are currently investigating some of these issues.

Appendix

A Proofs of the Technical Results in Section 3

For brevity we omit the subscript n, but it is understood that everything depends on n.

The notation is as follows. The elements of the true covariance matrix Σ are called σ_{ij}. Σ can be decomposed into $\Sigma = \Gamma\Lambda\Gamma'$, where Λ is a diagonal matrix, and Γ is a rotation matrix. We denote the elements of Λ by λ_{ij}, thus $\lambda_{ij} = 0$ for $i \neq j$, and the eigenvalues of Σ are called λ_{ii}. This differs from the body of the paper, where the eigenvalues are called λ_i instead, but no confusion should be possible. We use the matrix U to rotate the data: $Y = U^t X$ is a $p \times n$ matrix of n iid observations on a system of p random variables with mean zero and covariance matrix Λ.

A.1 Proof of Lemma 3.1

Since the Frobenius norm is invariant by rotation, we have:

$$\|\Sigma\|^2 = \|\Lambda\|^2 = \frac{1}{p}\sum_{i=1}^{p} \mathrm{E}[y_{i1}^2]^2 \leq \frac{1}{p}\sum_{i=1}^{p} \mathrm{E}[y_{i1}^4] \leq \sqrt{\frac{1}{p}\sum_{i=1}^{p} \mathrm{E}[y_{i1}^8]} \leq \sqrt{K_2},$$

where the constant K_2 is defined by Assumption 2. Therefore the norm of the true covariance matrix remains bounded as n goes to infinity. This implies that $\mu = \Sigma \circ I \leq \|\Sigma\|$ is bounded too (remember that Definition 2 assigns norm one to the identity). Also, $\alpha^2 = \|\Sigma - \mu I\|^2 = \|\Sigma\|^2 - \mu^2$ remains bounded as n goes to infinity. Furthermore, we have:

$$\begin{aligned}
\mathrm{E}[\|S - \Sigma\|^2] &= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p} \mathrm{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n} x_{ik}x_{jk} - \sigma_{ij}\right)^2\right] \\
&= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p} \mathrm{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n} y_{ik}y_{jk} - \lambda_{ij}\right)^2\right] \\
&= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p} \mathrm{Var}\left[\frac{1}{n}\sum_{k=1}^{n} y_{ik}y_{jk}\right] \\
&= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p} \frac{1}{n}\mathrm{Var}\left[y_{i1}y_{j1}\right] \\
&\leq \frac{1}{pn}\sum_{i=1}^{p}\sum_{j=1}^{p} \mathrm{E}[y_{i1}^2 y_{j1}^2] \\
&\leq \frac{1}{pn}\sum_{i=1}^{p}\sum_{j=1}^{p} \sqrt{\mathrm{E}[y_{i1}^4]}\sqrt{\mathrm{E}[y_{j1}^4]} \\
&\leq \frac{p}{n}\left(\frac{1}{p}\sum_{i=1}^{p} \sqrt{\mathrm{E}[y_{i1}^4]}\right)^2 \\
&\leq \frac{p}{n}\left(\frac{1}{p}\sum_{i=1}^{p} \mathrm{E}[y_{i1}^4]\right)
\end{aligned}$$

$$\leq \frac{p}{n}\sqrt{\frac{1}{p}\sum_{i=1}^{p}\mathrm{E}[y_{i1}^8]}$$
$$\leq K_1\sqrt{K_2},$$

where the constants K_1 and K_2 are defined by Assumptions 1 and 2 respectively. It shows that β^2 remains bounded as n goes to infinity. Finally, by Lemma 2.1, $\delta^2 = \alpha^2 + \beta^2$ also remains bounded as n goes to infinity. □

A.2 Proof of Theorem 3.1

We have:

$$\begin{aligned}
\mu^2 + \theta^2 &= \left(\mathrm{E}\left[\frac{1}{p}\sum_{i=1}^{p} y_{i1}^2\right]\right)^2 + \mathrm{Var}\left[\frac{1}{p}\sum_{i=1}^{p} y_{i1}^2\right] \\
&= \mathrm{E}\left[\left(\frac{1}{p}\sum_{i=1}^{p} y_{i1}^2\right)^2\right] \\
&= \frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{E}[y_{i1}^2 y_{j1}^2] \\
&\leq \frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{p}\sqrt{\mathrm{E}[y_{i1}^4]}\sqrt{\mathrm{E}[y_{j1}^4]} \\
&\leq \left(\frac{1}{p}\sum_{i=1}^{p}\sqrt{\mathrm{E}[y_{i1}^4]}\right)^2 \\
&\leq \frac{1}{p}\sum_{i=1}^{p}\mathrm{E}[y_{i1}^4] \\
&\leq \sqrt{\frac{1}{p}\sum_{i=1}^{p}\mathrm{E}[y_{i1}^8]} \\
&\leq \sqrt{K_2}.
\end{aligned}$$

Therefore θ^2 remains bounded as n goes to infinity. We can rewrite the expected quadratic loss of the sample covariance matrix as:

$$\begin{aligned}
\beta^2 &= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n} y_{ik}y_{jk} - \lambda_{ij}\right)^2\right] \\
&= \frac{1}{pn}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{E}[(y_{i1}y_{j1} - \lambda_{ij})^2] \\
&= \frac{1}{pn}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{E}[y_{i1}^2 y_{j1}^2] - \frac{1}{pn}\sum_{i=1}^{p}\sum_{j=1}^{p}\lambda_{ij}^2 \\
&= \frac{p}{n}(\mu^2 + \theta^2) - \frac{1}{pn}\sum_{i=1}^{p}\lambda_{ii}^2.
\end{aligned}$$

The last term on the right hand side of the last equation verifies:

$$\frac{1}{pn}\sum_{i=1}^{p}\lambda_{ii}^2 = \frac{1}{n}\left(\frac{1}{p}\sum_{i=1}^{p}\mathrm{E}[y_{i1}^2]^2\right) \leq \frac{1}{n}\left(\frac{1}{p}\sum_{i=1}^{p}\mathrm{E}[y_{i1}^4]\right) \leq \frac{1}{n}\sqrt{\frac{1}{p}\sum_{i=1}^{p}\mathrm{E}[y_{i1}^8]} \leq \frac{1}{n}\sqrt{K_2},$$

therefore the difference $\beta^2 - \frac{p}{n}(\mu^2 + \theta^2)$ converges to zero as n goes to infinity. □

A.3 Proof of Lemma 3.2

The proof of the first statement is:

$$\mathrm{E}[m] = \mathrm{E}[S \circ I] = \mathrm{E}[S] \circ I = \Sigma \circ I = \mu.$$

Consider the second statement.

$$\begin{aligned}
\mathrm{E}[(m-\mu)^4] &= \mathrm{E}\left[\left\{\frac{1}{p}\sum_{i=1}^{p}\frac{1}{n}\sum_{k=1}^{n}(y_{ik}^2 - \lambda_{ii})\right\}^4\right] \\
&= \mathrm{E}\left[\left\{\frac{1}{n}\sum_{k=1}^{n}\frac{1}{p}\sum_{i=1}^{p}(y_{ik}^2 - \lambda_{ii})\right\}^4\right] \\
&= \frac{1}{n^4}\sum_{k_1=1}^{n}\sum_{k_2=1}^{n}\sum_{k_3=1}^{n}\sum_{k_4=1}^{n}\mathrm{E}\left[\left\{\frac{1}{p}\sum_{i=1}^{p}(y_{ik_1}^2 - \lambda_{ii})\right\}\left\{\frac{1}{p}\sum_{i=1}^{p}(y_{ik_2}^2 - \lambda_{ii})\right\}\right. \\
&\qquad \left.\times\left\{\frac{1}{p}\sum_{i=1}^{p}(y_{ik_3}^2 - \lambda_{ii})\right\}\left\{\frac{1}{p}\sum_{i=1}^{p}(y_{ik_4}^2 - \lambda_{ii})\right\}\right] \quad (22)
\end{aligned}$$

In the summation on the right hand side of Equation (22), the expectation is nonzero only if $k_1 = k_2$ or $k_1 = k_3$ or $k_1 = k_4$ or $k_2 = k_3$ or $k_2 = k_4$ or $k_3 = k_4$. Since these six conditions are symmetric, we have:

$$\begin{aligned}
&\mathrm{E}[(m-\mu)^4] \\
&\leq \frac{6}{n^4}\sum_{k_1=1}^{n}\sum_{k_3=1}^{n}\sum_{k_4=1}^{n}\left|\mathrm{E}\left[\left\{\frac{1}{p}\sum_{i=1}^{p}(y_{ik_1}^2 - \lambda_{ii})\right\}^2\left\{\frac{1}{p}\sum_{i=1}^{p}(y_{ik_3}^2 - \lambda_{ii})\right\}\left\{\frac{1}{p}\sum_{i=1}^{p}(y_{ik_4}^2 - \lambda_{ii})\right\}\right]\right| \\
&\leq \frac{6}{n^4}\sum_{k_1=1}^{n}\sum_{k_3=1}^{n}\sum_{k_4=1}^{n}\sqrt{\mathrm{E}\left[\left\{\frac{1}{p}\sum_{i=1}^{p}(y_{ik_1}^2 - \lambda_{ii})\right\}^4\right]} \\
&\qquad \times\sqrt{\mathrm{E}\left[\left\{\frac{1}{p}\sum_{i=1}^{p}(y_{ik_3}^2 - \lambda_{ii})\right\}^2\left\{\frac{1}{p}\sum_{i=1}^{p}(y_{ik_4}^2 - \lambda_{ii})\right\}^2\right]} \\
&\leq \frac{6}{n^4}\sum_{k_1=1}^{n}\sum_{k_3=1}^{n}\sum_{k_4=1}^{n}\sqrt{\mathrm{E}\left[\left\{\frac{1}{p}\sum_{i=1}^{p}(y_{ik_1}^2 - \lambda_{ii})\right\}^4\right]} \\
&\qquad \times\sqrt[4]{\mathrm{E}\left[\left\{\frac{1}{p}\sum_{i=1}^{p}(y_{ik_3}^2 - \lambda_{ii})\right\}^4\right]}\sqrt[4]{\mathrm{E}\left[\left\{\frac{1}{p}\sum_{i=1}^{p}(y_{ik_4}^2 - \lambda_{ii})\right\}^4\right]} \\
&\leq \frac{6}{n}\mathrm{E}\left[\left\{\frac{1}{p}\sum_{i=1}^{p}(y_{i1}^2 - \lambda_{ii})\right\}^4\right].
\end{aligned}$$

Now we want to eliminate the λ_{ii}'s from the bound. We can do it by using the inequality:

$$\begin{aligned}
\mathrm{E}\left[\left\{\frac{1}{p}\sum_{i=1}^{p}(y_{i1}^2-\lambda_{ii})\right\}^4\right] &= \sum_{q=0}^{4}(-1)^q\binom{4}{q}\mathrm{E}\left[\left\{\frac{1}{p}\sum_{i=1}^{p}y_{i1}^2\right\}^q\right]\mathrm{E}\left[\frac{1}{p}\sum_{i=1}^{p}y_{i1}^2\right]^{4-q} \\
&\leq \sum_{q=0}^{4}\binom{4}{q}\mathrm{E}\left[\left\{\frac{1}{p}\sum_{i=1}^{p}y_{i1}^2\right\}^4\right]^{q/4}\mathrm{E}\left[\left\{\frac{1}{p}\sum_{i=1}^{p}y_{i1}^2\right\}^4\right]^{(4-q)/4} \\
&\leq 2^4\,\mathrm{E}\left[\left\{\frac{1}{p}\sum_{i=1}^{p}y_{i1}^2\right\}^4\right].
\end{aligned}$$

Therefore we have:

$$\mathrm{E}[(m-\mu)^4] \leq \frac{96}{n}\mathrm{E}\left[\left\{\frac{1}{p}\sum_{i=1}^{p}y_{i1}^2\right\}^4\right] \leq \frac{96}{n}\mathrm{E}\left[\frac{1}{p}\sum_{i=1}^{p}y_{i1}^8\right] \leq \frac{96K_2}{n} \to 0.$$

This shows that the estimator m converges to its expectation μ in quartic mean. □

A.4 Proof of Lemma 3.3

We prove this lemma by successively decomposing $d^2-\delta^2$ into terms that are easier to study.

$$d^2-\delta^2 = (\|S-mI\|^2-\|S-\mu I\|^2)+(\|S-\mu I\|^2-\mathrm{E}[\|S-\mu I\|^2]) \tag{23}$$

It is sufficient to show that both terms in parentheses on the right hand side of Equation (23) converge to zero in quadratic mean. Consider the first term. Since mI is the orthogonal projection for the inner product $\circ$ of the sample covariance matrix S onto the line spanned by the identity, we have: $\|S-\mu I\|^2-\|S-mI\|^2 = \|\mu I-mI\|^2 = (\mu-m)^2$, therefore by Lemma 3.2 it converges to zero in quadratic mean. Now consider the second term.

$$\|S-\mu I\|^2 = \mu^2-2\mu m+\|S\|^2 \tag{24}$$

Again it is sufficient to show that the three terms on the right hand side of Equation (24) converge to their expectations in quadratic mean. The first term μ^2 is equal to its expectation, so it trivially does. The second term $2\mu m$ does too by Lemma 3.2, keeping in mind that μ is bounded by Lemma 3.1. Now consider the third term $\|S\|^2$:

$$\begin{aligned}
\|S\|^2 &= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\left(\frac{1}{n}\sum_{k=1}^{n}y_{ik}y_{jk}\right)^2 \\
&= \frac{p}{n^2}\sum_{k_1=1}^{n}\sum_{k_2=1}^{n}\left(\frac{1}{p}\sum_{i=1}^{p}y_{ik_1}y_{ik_2}\right)^2 \\
&= \frac{p}{n^2}\sum_{k=1}^{n}\left(\frac{1}{p}\sum_{i=1}^{p}y_{ik}^2\right)^2+\frac{p}{n^2}\sum_{k_1=1}^{n}\sum_{\substack{k_2=1\\k_2\neq k_1}}^{n}\left(\frac{1}{p}\sum_{i=1}^{p}y_{ik_1}y_{ik_2}\right)^2
\end{aligned} \tag{25}$$

Again it is sufficient to show that both terms on the right hand side of Equation (25) converge to their expectations in quadratic mean. Consider the first term.

$$\mathrm{Var}\left[\frac{p}{n^2}\sum_{k=1}^{n}\left(\frac{1}{p}\sum_{i=1}^{p}y_{ik}^2\right)^2\right] = \frac{p^2}{n^3}\mathrm{Var}\left[\left(\frac{1}{p}\sum_{i=1}^{p}y_{i1}^2\right)^2\right]$$

$$
\begin{aligned}
&\leq \frac{p^2}{n^3}\mathrm{E}\left[\left(\frac{1}{p}\sum_{i=1}^{p} y_{i1}^2\right)^4\right] \\
&\leq \left(\frac{1}{n}\right)\left(\frac{p}{n}\right)^2\left(\frac{1}{p}\sum_{i=1}^{p}\mathrm{E}\left[y_{i1}^8\right]\right) \\
&\leq \frac{K_1^2 K_2}{n} \to 0
\end{aligned}
$$

Therefore the first term on the right hand side of Equation (25) converges to its expectation in quadratic mean. Now consider the second term.

$$
\begin{aligned}
&\mathrm{Var}\left[\frac{p}{n^2}\sum_{k_1=1}^{n}\sum_{\substack{k_2=1\\k_2\neq k_1}}^{n}\left(\frac{1}{p}\sum_{i=1}^{p} y_{ik_1}y_{ik_2}\right)^2\right] \\
&\qquad = \frac{p^2}{n^4}\sum_{k_1=1}^{n}\sum_{\substack{k_2=1\\k_2\neq k_1}}^{n}\sum_{k_3=1}^{n}\sum_{\substack{k_4=1\\k_4\neq k_3}}^{n}\mathrm{Cov}\left[\left(\frac{1}{p}\sum_{i=1}^{p} y_{ik_1}y_{ik_2}\right)^2,\left(\frac{1}{p}\sum_{i=1}^{p} y_{ik_3}y_{ik_4}\right)^2\right] \quad (26)
\end{aligned}
$$

The covariances on the right hand side of Equation (26) only depend on $(\{k_1,k_2\}\cap\{k_3,k_4\})^{\#}$, the cardinal of the intersection of the set $\{k_1,k_2\}$ with the set $\{k_3,k_4\}$. This number can be zero, one or two. We study each case separately.

$\underline{(\{k_1,k_2\}\cap\{k_3,k_4\})^{\#}=0}$

In this case $\left(\frac{1}{p}\sum_{i=1}^{p} y_{ik_1}y_{ik_2}\right)^2$ and $\left(\frac{1}{p}\sum_{i=1}^{p} y_{ik_3}y_{ik_4}\right)^2$ are independent, so their covariance is zero.

$\underline{(\{k_1,k_2\}\cap\{k_3,k_4\})^{\#}=1}$

This case occurs $4n(n-1)(n-2)$ times in the summation on the right hand side of Equation (26). Each time we have:

$$
\begin{aligned}
&\mathrm{Cov}\left[\left(\frac{1}{p}\sum_{i=1}^{p} y_{ik_1}y_{ik_2}\right)^2,\left(\frac{1}{p}\sum_{i=1}^{p} y_{ik_3}y_{ik_4}\right)^2\right] \\
&\qquad = \mathrm{Cov}\left[\left(\frac{1}{p}\sum_{i=1}^{p} y_{i1}y_{i2}\right)^2,\left(\frac{1}{p}\sum_{i=1}^{p} y_{i1}y_{i3}\right)^2\right] \\
&\qquad \leq \mathrm{E}\left[\left(\frac{1}{p}\sum_{i=1}^{p} y_{i1}y_{i2}\right)^2\left(\frac{1}{p}\sum_{i=1}^{p} y_{i1}y_{i3}\right)^2\right] \\
&\qquad \leq \mathrm{E}\left[\frac{1}{p^4}\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p} y_{i1}y_{i2}y_{j1}y_{j2}y_{k1}y_{k3}y_{l1}y_{l3}\right] \\
&\qquad \leq \frac{1}{p^4}\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}\mathrm{E}[y_{i1}y_{j1}y_{k1}y_{l1}]\mathrm{E}[y_{i2}y_{j2}]\mathrm{E}[y_{k3}y_{l3}] \\
&\qquad \leq \frac{1}{p^4}\sum_{i=1}^{p}\sum_{k=1}^{p}\mathrm{E}[y_{i1}^2 y_{k1}^2]\mathrm{E}[y_{i2}^2]\mathrm{E}[y_{k3}^2]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{1}{p^4}\sum_{i=1}^{p}\sum_{k=1}^{p}\sqrt{\mathrm{E}[y_{i1}^4]}\sqrt{\mathrm{E}[y_{k1}^4]}\mathrm{E}[y_{i2}^2]\mathrm{E}[y_{k3}^2] \\
&\leq \frac{1}{p^2}\left(\frac{1}{p}\sum_{i=1}^{p}\sqrt{\mathrm{E}[y_{i1}^4]}\mathrm{E}[y_{i1}^2]\right)^2 \\
&\leq \frac{1}{p^2}\left(\frac{1}{p}\sum_{i=1}^{p}\sqrt{\mathrm{E}[y_{i1}^4]}\sqrt{\mathrm{E}[y_{i1}^4]}\right)^2 \\
&\leq \frac{1}{p^2}\left(\frac{1}{p}\sum_{i=1}^{p}\mathrm{E}[y_{i1}^8]\right) \\
&\leq \frac{K_2}{p^2}
\end{aligned}
$$

and

$$
\begin{aligned}
&-\mathrm{Cov}\left[\left(\frac{1}{p}\sum_{i=1}^{p}y_{ik_1}y_{ik_2}\right)^2,\left(\frac{1}{p}\sum_{i=1}^{p}y_{ik_3}y_{ik_4}\right)^2\right] \\
&\qquad = -\mathrm{Cov}\left[\left(\frac{1}{p}\sum_{i=1}^{p}y_{i1}y_{i2}\right)^2,\left(\frac{1}{p}\sum_{i=1}^{p}y_{i1}y_{i3}\right)^2\right] \\
&\qquad \leq \mathrm{E}\left[\left(\frac{1}{p}\sum_{i=1}^{p}y_{i1}y_{i2}\right)^2\right]\mathrm{E}\left[\left(\frac{1}{p}\sum_{i=1}^{p}y_{i1}y_{i3}\right)^2\right] \\
&\qquad \leq \left(\frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{E}[y_{i1}y_{j1}]^2\right)^2 \\
&\qquad \leq \frac{1}{p}\left(\frac{1}{p}\sum_{i=1}^{p}\mathrm{E}[y_{i1}^2]^2\right)^2 \\
&\qquad \leq \frac{1}{p^2}\left(\frac{1}{p}\sum_{i=1}^{p}\mathrm{E}[y_{i1}^8]\right) \\
&\qquad \leq \frac{K_2}{p^2}.
\end{aligned}
$$

Therefore in this case the absolute value of the covariance on the right hand side of Equation (26) is bounded by K_2/p^2.

$\underline{(\{k_1,k_2\}\cap\{k_3,k_4\})^{\#}=2}$

This case occurs $2n(n-1)$ times in the summation on the right hand side of Equation (26). Each time we have:

$$
\begin{aligned}
&\left|\mathrm{Cov}\left[\left(\frac{1}{p}\sum_{i=1}^{p}y_{ik_1}y_{ik_2}\right)^2,\left(\frac{1}{p}\sum_{i=1}^{p}y_{ik_3}y_{ik_4}\right)^2\right]\right| \\
&\qquad = \left|\mathrm{Cov}\left[\left(\frac{1}{p}\sum_{i=1}^{p}y_{i1}y_{i2}\right)^2,\left(\frac{1}{p}\sum_{i=1}^{p}y_{i1}y_{i2}\right)^2\right]\right| \\
&\qquad \leq \frac{1}{p^4}\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}\left|\mathrm{Cov}[y_{i1}y_{i2}y_{j1}y_{j2},y_{k1}y_{k2}y_{l1}y_{l2}]\right| \quad (27)
\end{aligned}
$$

In the summation on the right hand side of Equation (27), the set of quadruples of integers between 1 and p can be decomposed into two disjoint subsets: $\{1,\ldots,p\}^4 = Q\cup R$, where

Q contains those quadruples that are made of four *distinct* integers, and R contains the remainder. Thus we can make the following decomposition:

$$\left|\mathrm{Cov}\left[\left(\frac{1}{p}\sum_{i=1}^{p} y_{ik_1}y_{ik_2}\right)^2, \left(\frac{1}{p}\sum_{i=1}^{p} y_{ik_3}y_{ik_4}\right)^2\right]\right|$$
$$\leq \frac{1}{p^4}\sum_{(i,j,k,l)\in Q} |\mathrm{Cov}[y_{i1}y_{i2}y_{j1}y_{j2}, y_{k1}y_{k2}y_{l1}y_{l2}]|$$
$$+\frac{1}{p^4}\sum_{(i,j,k,l)\in R} |\mathrm{Cov}[y_{i1}y_{i2}y_{j1}y_{j2}, y_{k1}y_{k2}y_{l1}y_{l2}]|$$

Let us express the first term of this decomposition as a function of the quantity that vanishes under Assumption 3: $v = \frac{p^2}{n^2} \times \frac{\sum_{(i,j,k,l)\in Q}(\mathrm{Cov}[y_{i1}y_{j1}, y_{k1}y_{l1}])^2}{\text{Cardinal of } Q}$. First, notice that the cardinal of Q is $p(p-1)(p-2)(p-3)$. Also, when $i \neq j$ and $k \neq l$, we have $\mathrm{E}[y_{i1}y_{j1}] = \mathrm{E}[y_{k1}y_{l1}] = 0$, therefore:

$$\begin{aligned}
|\mathrm{Cov}[y_{i1}y_{i2}y_{j1}y_{j2}, y_{k1}y_{k2}y_{l1}y_{l2}]| &= |\mathrm{E}[y_{i1}y_{i2}y_{j1}y_{j2}y_{k1}y_{k2}y_{l1}y_{l2}] \\
&\quad - \mathrm{E}[y_{i1}y_{i2}y_{j1}y_{j2}]\,\mathrm{E}[y_{k1}y_{k2}y_{l1}y_{l2}]\,| \\
&= |\mathrm{E}[y_{i1}y_{j1}y_{k1}y_{l1}]^2 - \mathrm{E}[y_{i1}y_{j1}]^2\,\mathrm{E}[y_{k1}y_{l1}]^2\,| \\
&= \mathrm{E}[y_{i1}y_{j1}y_{k1}y_{l1}]^2 \\
&= (\mathrm{Cov}[y_{i1}y_{j1}, y_{k1}y_{l1}] + \mathrm{E}[y_{i1}y_{j1}]\,\mathrm{E}[y_{k1}y_{l1}])^2 \\
&= (\mathrm{Cov}[y_{i1}y_{j1}, y_{k1}y_{l1}])^2.
\end{aligned}$$

This enables us to express the first term of the decomposition as: $\frac{n^2(p-1)(p-2)(p-3)}{p^5}v$.
Now consider the second term of the decomposition. The summation over R only extends over the quadruples (i,j,k,l) such that $i=j$ or $i=k$ or $i=l$ or $j=k$ or $j=l$ or $k=l$. Since these six conditions are symmetric, we have:

$$\frac{1}{p^4}\sum_{(i,j,k,l)\in R} |\mathrm{Cov}[y_{i1}y_{i2}y_{j1}y_{j2}, y_{k1}y_{k2}y_{l1}y_{l2}]|$$
$$\leq \frac{6}{p^4}\sum_{i=1}^{p}\sum_{k=1}^{p}\sum_{k=1}^{p} |\mathrm{Cov}[y_{i1}y_{i2}y_{i1}y_{i2}, y_{k1}y_{k2}y_{l1}y_{l2}]|$$
$$\leq \frac{6}{p^4}\sum_{i=1}^{p}\sum_{k=1}^{p}\sum_{k=1}^{p} \sqrt{\mathrm{E}[y_{i1}^2y_{i2}^2y_{i1}^2y_{i2}^2]\,\mathrm{E}[y_{k1}^2y_{k2}^2y_{l1}^2y_{l2}^2]}$$
$$\leq \frac{6}{p^4}\sum_{i=1}^{p}\sum_{k=1}^{p}\sum_{k=1}^{p} \mathrm{E}\left[y_{i1}^4\right]\mathrm{E}\left[y_{k1}^2y_{l1}^2\right]$$
$$\leq \frac{6}{p^4}\sum_{i=1}^{p}\sum_{k=1}^{p}\sum_{k=1}^{p} \mathrm{E}\left[y_{i1}^4\right]\sqrt{\mathrm{E}\left[y_{k1}^4\right]}\sqrt{\mathrm{E}\left[y_{l1}^4\right]}$$
$$\leq \frac{6}{p}\left(\frac{1}{p}\sum_{i=1}^{p}\mathrm{E}\left[y_{i1}^4\right]\right)\left(\frac{1}{p}\sum_{i=1}^{p}\sqrt{\mathrm{E}\left[y_{i1}^4\right]}\right)^2$$
$$\leq \frac{6}{p}\left(\frac{1}{p}\sum_{i=1}^{p}\mathrm{E}\left[y_{i1}^4\right]\right)^2$$
$$\leq \frac{6}{p}\left(\frac{1}{p}\sum_{i=1}^{p}\mathrm{E}\left[y_{i1}^8\right]\right)$$

$$\leq \frac{6K_2}{p}.$$

This completes the study of the decomposition, and also of the three possible cases. We can now bring all the results together to bound the summation on the right hand side of Equation (26):

$$\begin{aligned}
&\frac{p^2}{n^4}\sum_{k_1=1}^{n}\sum_{\substack{k_2=1\\k_2\neq k_1}}^{n}\sum_{k_3=1}^{n}\sum_{\substack{k_4=1\\k_4\neq k_3}}^{n}\left|\mathrm{Cov}\left[\left(\frac{1}{p}\sum_{i=1}^{p}y_{ik_1}y_{ik_2}\right)^2,\left(\frac{1}{p}\sum_{i=1}^{p}y_{ik_3}y_{ik_4}\right)^2\right]\right| \\
&\quad\leq \frac{p^2}{n^4}\left\{4n(n-1)(n-2)\frac{K_2}{p^2}+2n(n-1)\frac{n^2(p-1)(p-2)(p-3)}{p^5}v+2n(n-1)\frac{6K_2}{p}\right\} \\
&\quad\leq \frac{4K_2(1+3K_1)}{n}+2v\to 0.
\end{aligned}$$

Backing up, the second term on the right hand side of Equation (25) converges to its expectation in quadratic mean. Backing up again, the third term $\|S\|^2$ on the right hand side of Equation (24) converges to its expectation in quadratic mean. Backing up more, the second term between parentheses on the right hand side of Equation (23) converges to zero in quadratic mean. Backing up one last time, $d^2-\delta^2$ converges to zero in quadratic mean. For future reference note that, since $\|S-\mu I\|^2$ converges to its expectation δ^2 in quadratic mean and since δ^2 is bounded, $\mathrm{E}[\|S-\mu I\|^4]$ is bounded.. □

A.5 Proof of Lemma 3.4

We first prove that the unconstrained estimator $\bar{b}^2$ is consistent. As before, we do it by successively decomposing $\bar{b}^2-\beta^2$ into terms that are easier to study.

$$\begin{aligned}
\bar{b}^2-\beta^2 &= \left\{\frac{1}{n^2}\sum_{k=1}^{n}\|x_{\cdot k}x_{\cdot k}^{\mathrm{t}}-\Sigma\|^2-\mathrm{E}[\|S-\Sigma\|^2]\right\} \\
&\quad+\left\{\frac{1}{n^2}\sum_{k=1}^{n}\|x_{\cdot k}x_{\cdot k}^{\mathrm{t}}-S\|^2-\frac{1}{n^2}\sum_{k=1}^{n}\|x_{\cdot k}x_{\cdot k}^{\mathrm{t}}-\Sigma\|^2\right\} \qquad (28)
\end{aligned}$$

It is sufficient to show that both bracketed terms on the right hand side of Equation (28) converge to zero in quadratic mean. Consider the first term.

$$\begin{aligned}
\mathrm{E}[\|S-\Sigma\|^2] &= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n}x_{ik}x_{jk}-\sigma_{ij}\right)^2\right] \\
&= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Var}\left[\frac{1}{n}\sum_{k=1}^{n}x_{ik}x_{jk}\right] \\
&= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{1}{n^2}\sum_{k=1}^{n}\mathrm{Var}[x_{ik}x_{jk}] \\
&= \frac{1}{pn}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Var}[x_{i1}x_{j1}] \\
&= \frac{1}{pn}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{E}[(x_{i1}x_{j1}-\sigma_{ij})^2]
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}\left[\frac{1}{n}\|x_{\cdot 1}x_{\cdot 1}^{\mathrm{t}} - \Sigma\|^2\right] \\
&= \mathrm{E}\left[\frac{1}{n^2}\sum_{k=1}^{n}\|x_{\cdot k}x_{\cdot k}^{\mathrm{t}} - \Sigma\|^2\right]
\end{aligned}
$$

Therefore the first bracketed term on the right hand side of Equation (28) has expectation zero. For $k = 1, \ldots, n$ let $y_{\cdot k}$ denote the $p_n \times 1$ vector holding the k^{th} column of the matrix Y.

$$
\begin{aligned}
\mathrm{Var}\left[\frac{1}{n^2}\sum_{k=1}^{n}\|x_{\cdot k}x_{\cdot k}^{\mathrm{t}} - \Sigma\|^2\right] &= \frac{1}{n}\mathrm{Var}\left[\frac{1}{n}\|x_{\cdot 1}x_{\cdot 1}^{\mathrm{t}} - \Sigma\|^2\right] \\
&= \frac{1}{n}\mathrm{Var}\left[\frac{1}{n}\|y_{\cdot 1}y_{\cdot 1}^{\mathrm{t}} - \Lambda\|^2\right] \\
&= \frac{1}{p^2n^3}\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}\mathrm{Cov}[y_{i1}y_{j1} - \lambda_{ij}, y_{k1}y_{l1} - \lambda_{kl}] \\
&= \frac{1}{p^2n^3}\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}\mathrm{Cov}[y_{i1}y_{j1}, y_{k1}y_{l1}] \\
&\leq \frac{1}{p^2n^3}\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}\sqrt{\mathrm{E}[y_{i1}^2y_{j1}^2]\mathrm{E}[y_{k1}^2y_{l1}^2]} \\
&\leq \frac{1}{p^2n^3}\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}\sqrt[4]{\mathrm{E}[y_{i1}^4]\mathrm{E}[y_{j1}^4]\mathrm{E}[y_{k1}^4]\mathrm{E}[y_{l1}^4]} \\
&\leq \frac{p^2}{n^3}\left(\frac{1}{p}\sum_{i=1}^{p}\sqrt[4]{\mathrm{E}[y_{i1}^4]}\right)^4 \\
&\leq \frac{p^2}{n^3}\left(\frac{1}{p}\sum_{i=1}^{p}\mathrm{E}[y_{i1}^4]\right) \\
&\leq \frac{p^2}{n^3}\sqrt{\frac{1}{p}\sum_{i=1}^{p}\mathrm{E}[y_{i1}^8]} \\
&\leq \frac{K_1^2\sqrt{K_2}}{n}
\end{aligned}
$$

Therefore the first bracketed term on the right hand side of Equation (28) converges to zero in quadratic mean. Now consider the second term:

$$
\begin{aligned}
\frac{1}{n^2}\sum_{k=1}^{n}\left\|x_{\cdot k}x_{\cdot k}^{\mathrm{t}} - \Sigma\right\|^2 - \frac{1}{n^2}\sum_{k=1}^{n}\left\|x_{\cdot k}x_{\cdot k}^{\mathrm{t}} - S\right\|^2 &= \frac{1}{n^2}\sum_{k=1}^{n}2(S-\Sigma)\circ\left(x_{\cdot k}x_{\cdot k}^{\mathrm{t}} - \frac{S+\Sigma}{2}\right) \\
&= \frac{2}{n}(S-\Sigma)\circ\left(\frac{1}{n}\sum_{k=1}^{n}x_{\cdot k}x_{\cdot k}^{\mathrm{t}} - \frac{S+\Sigma}{2}\right) \\
&= \frac{2}{n}(S-\Sigma)\circ\left(S - \frac{S+\Sigma}{2}\right) \\
&= \frac{1}{n}\|S-\Sigma\|^2
\end{aligned}
$$

$\mathrm{E}[\|S-\Sigma\|^4]$ is bounded since $\mathrm{E}[\|S-\mu I\|^4]$ and $\|\Sigma - \mu I\|$ are bounded. Therefore the second term on the right hand side of Equation (28) converges to zero in quadratic mean. Backing up once more, $\overline{b}^2 - \beta^2$ converges to zero in quadratic mean.

Now let us turn to the constrained estimator $b^2 = \min(\bar{b}^2, d^2)$.

$$b^2 - \beta^2 = \min(\bar{b}^2, d^2) - \beta^2 \leq \bar{b}^2 - \beta^2 \leq |\bar{b}^2 - \beta^2| \leq \max(|\bar{b}^2 - \beta^2|, |d^2 - \delta^2|).$$

almost surely (a.s.). Furthermore, using $\delta^2 \geq \beta^2$, we have:

$$\begin{aligned} b^2 - \beta^2 &= \min(\bar{b}^2, d^2) - \beta^2 \\ &= \min(\bar{b}^2 - \beta^2, d^2 - \beta^2) \\ &\geq \min(\bar{b}^2 - \beta^2, d^2 - \delta^2) \\ &\geq \min(-|\bar{b}^2 - \beta^2|, -|d^2 - \delta^2|) \\ &\geq -\max(|\bar{b}^2 - \beta^2|, |d^2 - \delta^2|). \end{aligned}$$

a.s. Therefore

$$\mathrm{E}[(b^2 - \beta^2)^2] \leq \mathrm{E}[\max(|\bar{b}^2 - \beta^2|, |d^2 - \delta^2|)^2] \leq \mathrm{E}[(\bar{b}^2 - \beta^2)^2] + \mathrm{E}[(d^2 - \delta^2)^2].$$

On the right hand side, the first term converges to zero as we have shown earlier in this section, and the second term converges to zero as we have shown in Lemma 3.3. Therefore $b^2 - \beta^2$ converges to zero in quadratic mean. □

A.6 Proof of Lemma 3.5

Follows trivially from Lemmata 2.1, 3.3, and 3.4. □

A.7 Proof of Theorem 3.2

The following lemma will be useful in proving Theorems 3.2 and 3.3 and Lemma 3.6.

Lemma A.1 *If u^2 is a sequence of non-negative random variables (implicitly indexed by n, as usual) whose expectations converge to zero, and τ_1, τ_2 are two non-random scalars, and $\frac{u^2}{d^{\tau_1}\delta^{\tau_2}} \leq 2(d^2 + \delta^2)$ a.s., then:*

$$\mathrm{E}\left[\frac{u^2}{d^{\tau_1}\delta^{\tau_2}}\right] \to 0.$$

Proof of Lemma A.1 Fix $\varepsilon > 0$. Recall that the subscript n has been omitted to make the notation lighter, but is present implicitly. Let $\mathcal{N}$ denote the set of indices n such that $\delta^2 \leq \varepsilon/8$. Since $d^2 - \delta^2 \to 0$ in quadratic mean, there exists an integer n_1 such that $\forall n \geq n_1 \quad \mathrm{E}[|d^2 - \delta^2|] \leq \varepsilon/4$. For every $n \geq n_1$ inside the set $\mathcal{N}$, we have:

$$\mathrm{E}\left[\frac{u^2}{d^{\tau_1}\delta^{\tau_2}}\right] \leq 2(\mathrm{E}[d^2] + \delta^2) \leq 2(\mathrm{E}[|d^2 - \delta^2|] + 2\delta^2) \leq 2\left(\frac{\varepsilon}{4} + 2\frac{\varepsilon}{8}\right) = \varepsilon. \tag{29}$$

Now consider the complementary of the set $\mathcal{N}$. Since $\mathrm{E}[u^2] \to 0$, there exists an integer n_2 such that:

$$\forall n \geq n_2 \qquad \mathrm{E}[u^2] \leq \frac{\varepsilon^{\tau_1+\tau_2+1}}{2^{4\tau_1+3\tau_2+1}}.$$

Let $\mathbf{1}_{\{\cdot\}}$ denote the indicator function of an event, and let $\Pr(\cdot)$ denote its probability. From the proof of Lemma 3.1, δ^2 is bounded by $(1+K_1)\sqrt{K_2}$. Since $d^2-\delta^2$ converges to zero in quadratic mean, hence in probability, there exists an integer n_3 such that:

$$\forall n \geq n_3 \qquad \Pr\left(|d^2-\delta^2| \geq \frac{\varepsilon}{16}\right) \leq \frac{4\varepsilon}{16(1+K_1)\sqrt{K_2}+\varepsilon}.$$

For every $n \geq \max(n_2, n_3)$ outside the set $\mathcal{N}$, we have:

$$\begin{aligned}
\mathrm{E}\left[\frac{u^2}{d^{\tau_1}\delta^{\tau_2}}\right] &= \mathrm{E}\left[\frac{u^2}{d^{\tau_1}\delta^{\tau_2}}\mathbf{1}_{\{d^2\leq\varepsilon/16\}}\right] + \mathrm{E}\left[\frac{u^2}{d^{\tau_1}\delta^{\tau_2}}\mathbf{1}_{\{d^2>\varepsilon/16\}}\right] \\
&\leq \mathrm{E}\left[2(\delta^2+d^2)\mathbf{1}_{\{d^2\leq\varepsilon/16\}}\right] + \left(\frac{16}{\varepsilon}\right)^{\tau_1}\left(\frac{8}{\varepsilon}\right)^{\tau_2}\mathrm{E}\left[u^2\mathbf{1}_{\{d^2>\varepsilon/16\}}\right] \\
&\leq 2\left\{(1+K_1)\sqrt{K_2}+\frac{\varepsilon}{16}\right\}\Pr\left(|d^2-\delta^2| \geq \frac{\varepsilon}{16}\right) + \left(\frac{16}{\varepsilon}\right)^{\tau_1}\left(\frac{8}{\varepsilon}\right)^{\tau_2}\mathrm{E}\left[u^2\right] \\
&\leq 2\left\{(1+K_1)\sqrt{K_2}+\frac{\varepsilon}{16}\right\}\frac{4\varepsilon}{16(1+K_1)\sqrt{K_2}+\varepsilon} + \left(\frac{16}{\varepsilon}\right)^{\tau_1}\left(\frac{8}{\varepsilon}\right)^{\tau_2}\frac{\varepsilon^{\tau_1+\tau_2+1}}{2^{4\tau_1+3\tau_2+1}} \\
&\leq \varepsilon. \qquad (30)
\end{aligned}$$

Bringing together the results inside and outside the set $\mathcal{N}$ obtained in Equations (29)-(30) yields:

$$\forall n \geq \max(n_1, n_2, n_3) \qquad \mathrm{E}\left[\frac{u^2}{d^{\tau_1}\delta^{\tau_2}}\right] \leq \varepsilon.$$

This ends the proof of the lemma. □

Consider the first statement of Theorem 3.2.

$$\begin{aligned}
\|S^*-\Sigma^*\|^2 &= \left\|\frac{\beta^2}{\delta^2}(m-\mu)I + \left(\frac{a^2}{d^2}-\frac{\alpha^2}{\delta^2}\right)(S-mI)\right\|^2 \\
&= \frac{\beta^4}{\delta^4}(m-\mu)^2 + \left(\frac{a^2}{d^2}-\frac{\alpha^2}{\delta^2}\right)^2\|S-mI\|^2 \\
&\quad + 2\frac{\beta^2}{\delta^2}(m-\mu)\left(\frac{a^2}{d^2}-\frac{\alpha^2}{\delta^2}\right)(S-mI)\circ I \\
&= \frac{\beta^4}{\delta^4}(m-\mu)^2 + \left(\frac{a^2}{d^2}-\frac{\alpha^2}{\delta^2}\right)^2 d^2 \\
&\leq (m-\mu)^2 + \frac{(a^2\delta^2-\alpha^2d^2)^2}{d^2\delta^4}. \qquad (31)
\end{aligned}$$

It is sufficient to show that the expectations of both terms on the right hand side of Equation (31) converge to zero. The expectation of the first term does by Lemma 3.2. Now consider the second term. Since $\alpha^2 \leq \delta^2$ and $a^2 \leq d^2$, note that:

$$\frac{(a^2\delta^2-\alpha^2d^2)^2}{d^2\delta^4} \leq d^2 \leq 2(d^2+\delta^2) \qquad \text{a.s.}$$

Furthermore, since $a^2-\alpha^2$ and $d^2-\delta^2$ converge to zero in quadratic mean, and since α^2 and δ^2 are bounded, $a^2\delta^2-\alpha^2d^2 = (a^2-\alpha^2)\delta^2 - \alpha^2(d^2-\delta^2)$ converges to zero in quadratic mean.

Therefore the assumptions of Lemma A.1 are verified by $u^2 = (a^2\delta^2 - \alpha^2 d^2)^2$, $\tau_1 = 2$ and $\tau_2 = 4$. It implies that:

$$\mathrm{E}\left[\frac{(a^2\delta^2 - \alpha^2 d^2)^2}{d^2\delta^4}\right] \to 0.$$

The expectation of second term on the right hand side of Equation (31) converges to zero. Backing up, $\|S^* - \Sigma^*\|$ converges to zero in quadratic mean. This completes the proof of the first statement of Theorem 3.2.

Now consider the second statement.

$$\begin{aligned}\mathrm{E}\left[\left|\|S^* - \Sigma\|^2 - \|\Sigma^* - \Sigma\|^2\right|\right] &= \mathrm{E}[|(S^* - \Sigma^*) \circ (S^* + \Sigma^* - 2\Sigma)|] \\ &\leq \sqrt{\mathrm{E}[\|S^* - \Sigma^*\|^2]}\sqrt{\mathrm{E}[\|S^* + \Sigma^* - 2\Sigma\|^2]}. \qquad (32)\end{aligned}$$

As we have shown above, the first term on the right hand side of Equation (32) converges to zero. Given that $\mathrm{E}[\|\Sigma^* - \Sigma\|^2]$ is bounded, it also implies that the second term on the right hand side of Equation (32) is bounded. Therefore the product of the two terms on the right hand side of Equation (32) converges to zero. This completes the proof of the second and final statement. □

A.8 Proof of Lemma 3.6

We have:

$$\left|\frac{a^2b^2}{d^2} - \frac{\alpha^2\beta^2}{\delta^2}\right| = \frac{|a^2b^2\delta^2 - \alpha^2\beta^2 d^2|}{d^2\delta^2}.$$

Let us verify that the assumptions of Lemma A.1 hold for $u^2 = |a^2b^2\delta^2 - \alpha^2\beta^2 d^2|$, $\tau_1 = 2$ and $\tau_2 = 2$. Notice that:

$$\left|\frac{a^2b^2}{d^2} - \frac{\alpha^2\beta^2}{\delta^2}\right| \leq \frac{a^2b^2}{d^2} + \frac{\alpha^2\beta^2}{\delta^2} \leq a^2 + \alpha^2 \leq d^2 + \delta^2 \leq 2(d^2 + \delta^2)$$

a.s. Furthermore,

$$\begin{aligned}&\mathrm{E}[|a^2b^2\delta^2 - \alpha^2\beta^2 d^2|] \\ &\quad= \mathrm{E}[|(a^2b^2 - \alpha^2\beta^2)\delta^2 - \alpha^2\beta^2(d^2 - \delta^2)|] \\ &\quad= \mathrm{E}[|(a^2 - \alpha^2)(b^2 - \beta^2)\delta^2 + \alpha^2(b^2 - \beta^2)\delta^2 + (a^2 - \alpha^2)\beta^2\delta^2 - \alpha^2\beta^2(d^2 - \delta^2)|] \\ &\quad\leq \sqrt{\mathrm{E}[(a^2 - \alpha^2)^2]}\sqrt{\mathrm{E}[(b^2 - \beta^2)^2]}\delta^2 + \alpha^2\mathrm{E}[|b^2 - \beta^2|]\delta^2 + \mathrm{E}[|a^2 - \alpha^2|]\beta^2\delta^2 \\ &\qquad- \alpha^2\beta^2\mathrm{E}[|d^2 - \delta^2|].\end{aligned}$$

The right hand side converges to zero by Lemmata 3.1 3.3, 3.4, and 3.5. Therefore $\mathrm{E}[u^2] \to 0$, and the assumptions of Lemma A.1 are verified. It implies that:

$$\mathrm{E}\left[\left|\frac{a^2b^2}{d^2} - \frac{\alpha^2\beta^2}{\delta^2}\right|\right] \to 0.\square$$

A.9 Proof of Theorem 3.3

Define $\alpha_2 = \Sigma \circ S - \mu m$. Its expectation is $\mathrm{E}[\alpha_2] = \|\Sigma\|^2 - \mu^2 = \alpha^2$. We have:

$$|\alpha_2| = |\Sigma \circ S - \mu m| = |(\Sigma - \mu I) \circ (S - mI)| \leq \sqrt{\|\Sigma - \mu I\|^2}\sqrt{\|S - mI\|^2} \leq \delta d. \tag{33}$$

Let us prove that $\alpha_2 - \alpha^2$ converges to zero in quadratic mean.

$$\begin{aligned}
\mathrm{Var}[\alpha_2] &= \mathrm{Var}[\Sigma \circ S - \mu m] \\
&= \mathrm{Var}[\Sigma \circ S] + \mathrm{Var}[\mu m] - 2\mathrm{Cov}[\Sigma \circ S, \mu m] \\
&\leq 2\mathrm{Var}[\Sigma \circ S] + 2\mathrm{Var}[\mu m] \\
&\leq 2\mu^2 \mathrm{Var}[m] + 2\mathrm{Var}[\Sigma \circ S] \qquad (34)
\end{aligned}$$

The first term on the right hand side of Equation (34) converges to zero, since μ is bounded by Lemma 3.1, and since $\mathrm{Var}[m]$ converges to zero by Lemma 3.2. Consider the second term.

$$\begin{aligned}
\Sigma \circ S &= \frac{1}{p}\mathrm{tr}(\Sigma S^{\mathrm{t}}) \\
&= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma_{ij}s_{ij} \\
&= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma_{ij}\left(\frac{1}{n}\sum_{k=1}^{n}x_{ik}x_{jk}\right) \\
&= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\lambda_{ij}\left(\frac{1}{n}\sum_{k=1}^{n}y_{ik}y_{jk}\right) \\
&= \frac{1}{p}\sum_{i=1}^{p}\lambda_{ii}\left(\frac{1}{n}\sum_{k=1}^{n}y_{ik}^2\right)
\end{aligned}$$

Therefore:

$$\begin{aligned}
\mathrm{Var}[\Sigma \circ S] &= \mathrm{Var}\left[\frac{1}{p}\sum_{i=1}^{p}\lambda_{ii}\left(\frac{1}{n}\sum_{k=1}^{n}y_{ik}^2\right)\right] \\
&= \mathrm{Var}\left[\frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{p}\sum_{i=1}^{p}\lambda_{ii}y_{ik}^2\right)\right] \\
&= \frac{1}{n^2}\sum_{k_1=1}^{n}\sum_{k_2=1}^{n}\mathrm{Cov}\left[\frac{1}{p}\sum_{i=1}^{p}\lambda_{ii}y_{ik_1}^2, \frac{1}{p}\sum_{i=1}^{p}\lambda_{ii}y_{ik_2}^2\right] \\
&= \frac{1}{n^2}\sum_{k=1}^{n}\mathrm{Var}\left[\frac{1}{p}\sum_{i=1}^{p}\lambda_{ii}y_{ik}^2\right] \\
&= \frac{1}{n}\mathrm{Var}\left[\frac{1}{p}\sum_{i=1}^{p}\lambda_{ii}y_{i1}^2\right] \\
&\leq \frac{1}{n}\mathrm{E}\left[\left(\frac{1}{p}\sum_{i=1}^{p}\lambda_{ii}y_{i1}^2\right)^2\right] \\
&\leq \frac{1}{n}\mathrm{E}\left[\left(\frac{1}{p}\sum_{i=1}^{p}\lambda_{ii}^2\right)\left(\frac{1}{p}\sum_{i=1}^{p}y_{i1}^4\right)\right] \\
&\leq \frac{1}{n}\left(\frac{1}{p}\sum_{i=1}^{p}\mathrm{E}\left[y_{i1}^2\right]^2\right)\left(\frac{1}{p}\sum_{i=1}^{p}\mathrm{E}\left[y_{i1}^4\right]\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{n}\left(\frac{1}{p}\sum_{i=1}^{p}\mathrm{E}\left[y_{i1}^4\right]\right)^2 \\
&\leq \frac{1}{n}\left(\frac{1}{p}\sum_{i=1}^{p}\mathrm{E}\left[y_{i1}^8\right]\right) \\
&\leq \frac{K_2}{n}.
\end{aligned}
$$

It implies that the second term on the right hand side of Equation (34) converges to zero. Backing up, $\alpha_2 - \alpha^2$ converges to zero in quadratic mean.

Now let us find an explicit formula for the solution Σ^{**} to the optimization problem in Equation (15). This problem is very similar to the one in Theorem 2.1 but, instead of solving it with calculus as we did then, we will give an equivalent treatment based on geometry. The solution is the orthogonal projection according to the inner product $\circ$ of the true covariance matrix Σ onto the plane spanned by the identity matrix I and the sample covariance matrix S. Note that $(S - mI) \circ I = 0$, therefore $\left(I, \frac{S - mI}{\|S - mI\|}\right)$ forms an orthonormal basis for this plane. The formula for the projection has a simple expression in terms of the orthonormal basis:

$$
\begin{aligned}
\Sigma^{**} &= (\Sigma \circ I)I + \left(\Sigma \circ \frac{S - mI}{\|S - mI\|}\right)\frac{S - mI}{\|S - mI\|} \\
&= \mu I + \frac{\Sigma \circ S - \mu m}{\|S - mI\|^2}(S - mI) \\
&= \mu I + \frac{\alpha_2}{d^2}(S - mI).
\end{aligned}
$$

From now on, the proof is the same as for Theorem 3.2.

$$
\begin{aligned}
\|S^* - \Sigma^{**}\|^2 &= \left\| mI + \frac{a^2}{d^2}(S - mI) - \mu I - \frac{\alpha_2}{d^2}(S - mI) \right\|^2 \\
&= \left\| (m - \mu)I + \frac{a^2 - \alpha_2}{d^2}(S - mI) \right\|^2 \\
&= (m - \mu)^2 + \frac{(a^2 - \alpha_2)^2}{d^4}\|S - mI\|^2 + 2(m - \mu)\frac{a^2 - \alpha_2}{d^2}(S - mI) \circ I \\
&= (m - \mu)^2 + \frac{(a^2 - \alpha_2)^2}{d^2} \qquad (35)
\end{aligned}
$$

It is sufficient to show that the expectations of both terms on the right hand side of Equation (35) converge to zero. The expectation of the first term does by Lemma 3.2. Now consider the second term.

$$
\begin{aligned}
\frac{(a^2 - \alpha_2)^2}{d^2} &= \frac{a^4 + \alpha_2^2 - 2a^2\alpha_2}{d^2} \\
&\leq \frac{2a^4 + 2\alpha_2^2}{d^2} \\
&\leq 2d^2 + 2\delta^2,
\end{aligned}
$$

where we have used Equation (33). Furthermore, since $a^2 - \alpha^2$ and $\alpha_2 - \alpha^2$ both converge to zero in quadratic mean, $a^2 - \alpha_2$ also does. Therefore the assumptions of Lemma A.1 are

verified by $u^2 = (a^2 - \alpha_2)^2$, $\tau_1 = 2$ and $\tau_2 = 0$. It implies that:

$$\mathrm{E}\left[\frac{(a^2 - \alpha_2)^2}{d^2}\right] \to 0.$$

The expectation of second term on the right hand side of Equation (35) converges to zero. Backing up, $\|S^* - \Sigma^{**}\|$ converges to zero in quadratic mean. This completes the proof of the first statement of Theorem 3.3.

Now consider the second statement.

$$\begin{aligned} \mathrm{E}\left[\left|\|S^* - \Sigma\|^2 - \|\Sigma^{**} - \Sigma\|^2\right|\right] &= \mathrm{E}[|(S^* - \Sigma^{**}) \circ (S^* + \Sigma^{**} - 2\Sigma)|] \\ &\leq \sqrt{\mathrm{E}[\|S^* - \Sigma^{**}\|^2]}\sqrt{\mathrm{E}[\|S^* + \Sigma^{**} - 2\Sigma\|^2]}. \end{aligned} \quad (36)$$

As we have shown above, the first term on the right hand side of Equation (36) converges to zero. Given that $\mathrm{E}[\|S^* - \Sigma\|^2]$ is bounded, it also implies that the second term on the right hand side of Equation (36) is bounded. Therefore the product of the two terms on the right hand side of Equation (36) converges to zero. This completes the proof of the second and final statement. □

A.10 Proof of Theorem 3.4

$$\begin{aligned} \liminf(\mathrm{E}[\|\widehat{\Sigma} - \Sigma\|^2] - \mathrm{E}[\|S^* - \Sigma\|^2]) &\geq \inf(\mathrm{E}[\|\widehat{\Sigma} - \Sigma\|^2] - \mathrm{E}[\|\Sigma^{**} - \Sigma\|^2]) \\ &\quad + \lim(\mathrm{E}[\|\Sigma^{**} - \Sigma\|^2] - \mathrm{E}[\|S^* - \Sigma\|^2]). \end{aligned}$$

By construction of Σ^{**}, we have $\|\widehat{\Sigma} - \Sigma\|^2 - \|\Sigma^{**} - \Sigma\|^2 \geq 0$ a.s. , therefore the first term on the right hand side is non-negative. The second term on the right hand side is zero by Theorem 3.3. Therefore the left hand side is non-negative. This proves the first statement of Theorem 3.4. Now consider the second statement.

$$\begin{aligned} \lim(\mathrm{E}[\|\widehat{\Sigma} - \Sigma\|^2] - \mathrm{E}[\|S^* - \Sigma\|^2]) = 0 &\iff \lim(\mathrm{E}[\|\widehat{\Sigma} - \Sigma\|^2] - \mathrm{E}[\|\Sigma^{**} - \Sigma\|^2]) = 0 \\ &\iff \lim \mathrm{E}[\|\widehat{\Sigma} - \Sigma\|^2 - \|\Sigma^{**} - \Sigma\|^2] = 0 \\ &\iff \lim \mathrm{E}[\|\widehat{\Sigma} - \Sigma^{**}\|^2] = 0 \\ &\iff \lim \mathrm{E}[\|\widehat{\Sigma} - S^*\|^2] = 0 \end{aligned}$$

This completes the proof of the second and final statement. □

A.11 Proof of Theorem 3.5

Let $\lambda_{\max}(A)$ ($\lambda_{\min}(A)$) denote the largest (smallest) eigenvalue of the matrix A. The theorem is invariant to the multiplication of all the eigenvalues of Σ by a positive number. Therefore we can normalize Σ so that $\mu = 1$ without loss of generality. Then the assumption that the condition number of Σ is bounded is equivalent to the existence of two constants $\overline{\lambda}$ and $\underline{\lambda}$ independent of n such that: $0 < \underline{\lambda} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \overline{\lambda} < \infty$.

First, let us prove that the largest eigenvalue of S^* is bounded in probability. Let $Z = \Lambda^{-1/2}Y$ denote the normalized variables that are assumed to be cross-sectionally iid. We have:

$$\begin{aligned}
\lambda_{\max}(S^*) &= \lambda_{\max}\left(\frac{b^2}{d^2}mI + \frac{a^2}{d^2}S\right) \\
&= \frac{b^2}{d^2}m + \frac{a^2}{d^2}\lambda_{\max}(S) \\
&\leq \frac{b^2}{d^2}\lambda_{\max}(S) + \frac{a^2}{d^2}\lambda_{\max}(S) \\
&\leq \lambda_{\max}(S) \\
&\leq \lambda_{\max}\left(\frac{1}{n}\Lambda^{1/2}ZZ^{\mathrm{t}}\Lambda^{1/2}\right) \\
&\leq \lambda_{\max}\left(\frac{1}{n}ZZ^{\mathrm{t}}\right)\lambda_{\max}(\Lambda) \\
&\leq \lambda_{\max}\left(\frac{1}{n}ZZ^{\mathrm{t}}\right)\overline{\lambda}
\end{aligned}$$

almost surely. Assume *with* loss of generality, but temporarily, that p/n converges to some limit. Call the limit c. Assumption 1 implies that $c \leq K_1$. In this case, Yin, Bai and Krishnaiah (1988) show that:

$$\lambda_{\max}\left(\frac{1}{n}ZZ^{\mathrm{t}}\right) \to (1+\sqrt{c})^2 \qquad \text{a.s.} \tag{37}$$

It implies that:

$$\begin{aligned}
\Pr\left\{\lambda_{\max}\left(\frac{1}{n}ZZ^{\mathrm{t}}\right) \leq 2(1+\sqrt{c})^2\right\} &\to 1 \\
\Pr\left\{\lambda_{\max}(S^*) \leq 2\left(1+\sqrt{K_1}\right)^2\overline{\lambda}\right\} &\to 1.
\end{aligned} \tag{38}$$

Therefore, in the particular case where p/n converges to a limit, the largest eigenvalue of S^* is bounded in probability. Now consider the general case where p/n need not have a limit. Remember that p/n is bounded by Assumption 1. Take any subsequence along which p/n converges. Along this subsequence, the largest eigenvalue of S^* is bounded in probability. Notice that the bound in Equation (38) is independent of the particular subsequence. Since Equation (38) holds along any converging subsequence, it holds along the sequence as a whole. This proves that the largest eigenvalue of S^* is bounded in probability.

Now let us prove that the smallest eigenvalue of S^* is bounded away from zero in probability. A reasoning similar to the one above leads to: $\lambda_{\min}(S^*) \geq \lambda_{\min}(ZZ^{\mathrm{t}}/n)\underline{\lambda}$ a.s. Again assume with loss of generality, but temporarily, that p/n converges to some limit c. First consider the case $c \leq 1/2$. Bai and Yin (1993) show that:

$$\lambda_{\min}\left(\frac{1}{n}ZZ^{\mathrm{t}}\right) \to (1-\sqrt{c})^2 \qquad \text{a.s.} \tag{39}$$

It implies that:

$$\begin{aligned}
\Pr\left\{\lambda_{\min}\left(\frac{1}{n}ZZ^{\mathrm{t}}\right) \geq \frac{1}{2}(1-\sqrt{c})^2\right\} &\to 1 \\
\Pr\left\{\lambda_{\min}(S^*) \geq \frac{1}{2}\left(1-\sqrt{\frac{1}{2}}\right)^2\underline{\lambda}\right\} &\to 1.
\end{aligned} \tag{40}$$

Now turn to the other case: $c > 1/2$. In this case, we use:

$$\lambda_{\min}(S^*) = \frac{b^2}{d^2}m + \frac{a^2}{d^2}\lambda_{\min}(S) \geq \frac{b^2}{d^2}m$$

Fix any $\varepsilon > 0$. For large enough n, $p/n \geq 1/2 - \varepsilon$. Also, by Theorem 3.1, for large enough n, $\beta^2 \geq (p/n)(\mu^2 + \theta^2) - \varepsilon \geq 1/2 - 2\varepsilon$. In particular $\beta^2 \geq 1/4$ for large enough n. As a consequence, $\delta^2 \geq 1/4$ for large enough n. We can make the following decomposition:

$$\frac{b^2}{d^2}m - \frac{\beta^2}{\delta^2}\mu = \frac{\beta^2}{\delta^2}(m - \mu) + \frac{b^2 - \beta^2}{\delta^2}m + b^2 m\left(\frac{1}{d^2} - \frac{1}{\delta^2}\right) \tag{41}$$

We are going to show that all three terms on the right hand side of Equation (41) converge to zero in probability. The first term does as a consequence of Lemma 3.2 since $\beta^2/\delta^2 \leq 1$. Now consider the second term. For large enough n:

$$\mathrm{E}\left[\frac{|b^2 - \beta^2|}{\delta^2}m\right] \leq \frac{\sqrt{\mathrm{E}[(b^2 - \beta^2)^2]}\sqrt{\mathrm{E}[m^2]}}{1/4}.$$

In the numerator on the right hand side, $\mathrm{E}[(b^2 - \beta^2)^2]$ converges to zero by Lemma 3.4, and $\mathrm{E}[m^2]$ is bounded by Lemmata 3.1 and 3.2. Therefore the second term on the right hand side of Equation (41) converges to zero in first absolute moment, hence in probability. Now consider the third and last term. Since $d^2 - \delta^2$ converges to zero in probability by Lemma 3.3, and since δ^2 is bounded away from zero, $d^{-2} - \delta^{-2}$ converges to zero in probability. Furthermore, m and b^2 are bounded in probability by Lemmata 3.1, 3.2, and 3.4. Therefore the third term on the right hand side of Equation (41) converges to zero in probability. It implies that the left hand side of Equation (41) converges to zero in probability. Remember that, in the proof of Lemma 3.1, we have shown that $\delta^2 \leq (1 + K_1)\sqrt{K_2}$. For any $\varepsilon > 0$, we have:

$$\begin{aligned}
\Pr\left\{\frac{b^2}{d^2}m \geq \frac{\beta^2}{\delta^2}\mu - \varepsilon\right\} &\to 1 \\
\Pr\left\{\lambda_{\min}(S^*) \geq \frac{\beta^2}{\delta^2}\mu - \varepsilon\right\} &\to 1 \\
\Pr\left\{\lambda_{\min}(S^*) \geq \frac{\beta^2}{(1 + K_1)\sqrt{K_2}} - \varepsilon\right\} &\to 1 \\
\Pr\left\{\lambda_{\min}(S^*) \geq \frac{\frac{1}{2} - 2\varepsilon}{(1 + K_1)\sqrt{K_2}} - \varepsilon\right\} &\to 1.
\end{aligned}$$

There exists a particular value of $\varepsilon > 0$ that yields:

$$\Pr\left\{\lambda_{\min}(S^*) \geq \frac{1}{4(1 + K_1)\sqrt{K_2}}\right\} \to 1.$$

Bringing together the results obtained in the cases $c \leq 1/2$ and $c > 1/2$, we have:

$$\Pr\left\{\lambda_{\min}(S^*) \geq \min\left(\frac{1}{2}\left(1 - \sqrt{\frac{1}{2}}\right)^2 \lambda, \frac{1}{4(1 + K_1)\sqrt{K_2}}\right)\right\} \to 1. \tag{42}$$

Therefore, in the particular case where p/n converges to a limit, the smallest eigenvalue of S^* is bounded away from zero in probability. Again notice that the bound in Equation (42) does not depend on p/n. Therefore, by the same reasoning as for the largest eigenvalue, it implies that the smallest eigenvalue of S^* is bounded away from zero in probability, even in the general case where p/n need not have a limit. Bringing together the results obtained for the largest and the smallest eigenvalue, the condition number of S^* is bounded in probability. □

References

Bai, Z. D. and Yin, Y. Q. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Annals of Probability*, 21:1275–1294.

Brown, S. J. (1989). The number of factors in security returns. *Journal of Finance*, 44:1247–1262.

Crack, T. F. and Ledoit, O. (1996). Robust structure without predictability: The "compass rose" pattern of the stock market. *Journal of Finance*, 51:751–62.

Dey, D. K. and Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *Annals of Statistics*, 13:1581–1591.

Frost, P. A. and Savarino, J. E. (1986). An empirical Bayes approach to portfolio selection. *Journal of Financial and Quantitative Analysis*, 21:293–305.

Haff, L. R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Annals of Statistics*, 8:586–597.

Haff, L. R. (1982). Solutions of the Euler-Lagrange equations for certain multivariate normal estimation problems. Unpublished manuscript.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–54.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 361–379. Volume 1. J. Neyman, ed.

Jobson, J. D. and Korkie, B. (1980). Estimation for Markowitz efficient portfolios. *Journal of the American Statistical Association*, 75:544–554. Applications Section.

Kandel, S. and Stambaugh, R. F. (1995). Portfolio inefficiency and the cross-section of expected returns. *Journal of Finance*, 50:157–184.

Lin, S. P. and Perlman, M. D. (1985). A Monte-Carlo comparison of four estimators of a covariance matrix. In *Multivariate Analysis - VI*, 411–429. P. R. Krishnaiah, ed.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7:77–91.

Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the U.S.S.R. - Sbornik*, 1:457–483.

Michaud, R. O. (1989). The Markowitz optimization enigma: is 'optimized' optimal? *Financial Analysts Journal*, 31–42.

Muirhead, R. J. (1987). Developments in eigenvalue estimation. *Advances in Multivariate Statistical Analysis*, 277–288.

Muirhead, R. J. and Leung, P. L. (1987). Estimation of parameter matrices and eigenvalues in MANOVA and canonical correlation analysis. *Annals of Statistics*, 15:1651–1666.

Silverstein, J. W. (1994). Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. Unpublished.

Stein, C. (1975). Estimation of a covariance matrix. Rietz Lecture, 39th Annual Meeting IMS. Atlanta, GA.

Stein, C. (1982). Series of lectures given at the University of Washington, Seattle.

Yin, Y. Q. (1986). Limiting spectral distribution for a class of random matrices. *Journal of Multivariate Analysis*, 20:50–68.

Yin, Y. Q., Bai, Z. D., and Krishnaiah, P. R. (1988). On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probability Theory and Related Fields*, 78:509–21.