Portfolio Selection:

Improved Covariance Matrix Estimation

Olivier Ledoit*

November 1994

Abstract

This paper studies the estimation of the covariance matrix of returns on all stocks traded in the stock market, for portfolio selection. The number of observations is assumed to go to infinity, but the standard asymptotic assumption that keeps the number of variables bounded is lifted. In practice, this is appropriate when the number of traded stocks is at least of the same order of magnitude as the number of time periods, which is the usual case.

The first part characterizes intuitively and analytically the behavior of the sample covariance matrix in this case. Some of this work is potentially applicable to tests for the number of factors in the Arbitrage Pricing Theory (APT). The second part develops a simple and versatile estimator that has lower mean squared error than the sample covariance matrix. This estimator provides attractive answers to some fundamental questions in multivariate statistics. In the third and last part, Monte-Carlo simulations and historical data indicate that the new estimator improves over existing ones for portfolio selection: it yields portfolios with significantly lower risk than was previously possible. One of the empirical applications can be interpreted as a test of the Capital Asset Pricing Model (CAPM) with higher power than existing tests. It finds a significant and robust positive relationship between returns and betas, in contrast with less powerful tests in the literature.

*Sloan School of Management, Massachusetts Institute of Technology, E52-457, 50 Memorial Drive, Cambridge, MA 02142, USA; E-mail: oledoit@mit.edu; Phone: (617) 253 3637; Fax: (617) 258 6855. I wish to acknowledge the help of Professors Dimitri Bertsimas, Whitney Newey, Jack Silverstein and Bin Zhou, every member of the Finance faculty at MIT, and especially my advisor Andrew Lo. I am also indebted to my fellow Finance Ph.D. students at MIT, especially Timothy Crack. I thank the participants to seminars at MIT and the NBER for their comments. Heavy-duty computing was performed at MIT's Laboratory for Financial Engineering, with the patient and expert support of Firooz Partovi.

1 Introduction

1.1 Overview

The objective of this study is to estimate the covariance matrix of returns on all stocks traded in the stock market. This is important because the covariance matrix is a necessary input to Markowitz (1952) portfolio selection, a central method in stock market finance.

Our original approach is to assume that the number of observations T goes to infinity, as in standard asymptotics, but relax the standard asymptotic assumption that the number of variables N remains bounded by a constant: we only assume that N is bounded by a constant times T. It is a more realistic approximation of actual stock returns data, because typically the number of traded stocks N is at least of the same order of magnitude as the number of time periods T.

In the first part, we show that the sample covariance matrix is no longer consistent in this framework. Its mean squared error is of order N/T. For example, the sample covariance matrix of $N = 1,000$ stocks based on $T = 2,000$ observations is approximately as erroneous as the variance of the return on $N = 1$ stock estimated from $T = 2$ observations. Not only is the error substantial, but its nature is particularly damaging to portfolio selection: it causes the sample covariance matrix to be near-singular or singular. When the sample covariance matrix is near-singular, inverting it amplifies error and yields grossly inaccurate results for portfolio selection. This is the case if N is of the same order of magnitude as T. When the sample covariance matrix is singular, it cannot be inverted and cannot be used for portfolio selection at all. This is the case if N exceeds T.

We also review the spectral theory of large-dimensional random matrices. This theory gives the relationship between the eigenvalues of true and sample covariance matrices as a function of the ratio N/T, when T goes to infinity. It is the fact that the smallest sample covariance matrix eigenvalues are biased down towards zero that causes the singularity problem. This theory can potentially be used to test hypotheses about the eigenvalues of the covariance matrix of stock returns, such as the ones made by the Arbitrage Pricing Theory (APT).

In the second part, we improve over the sample covariance matrix. Some authors impose parsimonious structure (e.g. all pairs of stocks have the same correlation coefficient) to obtain an estimator with fewer free parameters. Better yet, Frost and Savarino (1986) combine such a "structured" estimator with the sample covariance matrix. We focus on weighted averages of a structured estimator with the sample covariance matrix and ask: what are the optimal weights? In our asymptotic framework, simple estimators of the weights minimizing mean squared error are consistent. We thus show how to improve both on any given structured estimator and on the sample covariance matrix by combining them in an asymptotically optimal way. Not only does it reduce mean squared error, but it generally escapes the singularity problem.

This method can be interpreted in Bayesian terms. The structured estimator can be called the prior, and its combination with the sample covariance matrix the posterior. Fundamental Bayesian questions have always been: Where does the prior come from? How confident are we in the prior? In finite sample, it is very hard to answer these questions satisfactorily. By contrast, in our asymptotic framework, the prior can be taken as any structured estimator, and the degree of confidence in the prior can be estimated consistently.

In the third part, we show that the improved estimator performs well in practice. In Monte-Carlo simulations, it has lower mean squared error than the sample covariance matrix, even in very small sample. Historical simulations confirm that, for a given set of constraints, our estimator yields portfolios with significantly lower risk than existing estimators.

One of our historical simulations is the first predictive Generalized Least Squares (GLS) cross-sectional regression of stock returns on betas and size. Similar regressions have been interpreted as tests of the CAPM. Thanks to our improved covariance matrix estimator, our GLS-based tests have more power than the tests in the literature, which are based on Ordinary Least Squares (OLS). By contrast with OLS tests, our GLS tests find a significant and robust positive relationship between returns and betas.

In this section, we present an overview of the paper and contrast it with the existing literature. In Section 2, we study the behavior of the sample covariance matrix when the number of variables is allowed to grow large. We develop a family of estimators that improve over the sample covariance matrix in Section 3. In Section 4, we see how these estimators perform for portfolio selection. Section 5 concludes. Appendix A contains details about the spectral theory of large-dimensional random matrices. Appendix B contains formulas for the more complicated versions of our estimator. Proofs are in Appendix C.

1.2 Comparison with Existing Literature

Jobson and Korkie (1980) show that using the sample covariance matrix for portfolio selection can cause severe problems. In some cases, it is better to use the identity matrix instead. Our main intuition is that a well-chosen linear combination of the sample covariance matrix with the identity can work even better than either. Our main contribution is to show how to choose this linear combination well.

Bawa, Brown and Klein (1979) argue that estimation risk coming from sample covariance matrix error is of the same nature as investment risk coming from stock return volatility. Their idea is of a Bayesian nature. One of their recommendations is to combine the sample covariance matrix with an "informative" prior. The more confident we are in the prior, the heavier it should weigh in the combination. They do not show how to obtain the prior and the degree of confidence in it. This is what we do.

Our paper is closest in spirit to Frost and Savarino's (1986). The difference is that they work in finite sample, while we work asymptotically. In finite sample, they have to ignore dependence between the prior and the sample covariance matrix, assume normality, and require that observations outnumber variables. Their formula is not explicit and is costly to compute for large universes of stocks. Asymptotically, we avoid all these problems. The price to pay is that peak performance only kicks in when N and T are large (larger than, say, 30), but this is almost always the case in practice.

Kandel and Stambaugh (1994) analyze cross-sectional regressions of stock returns on betas. The CAPM implies a positive slope. A problem arises because the market, with respect to which betas are measured, is only known approximately (Roll, 1977). Then the regression method matters. With Ordinary Least Squares (OLS), the regression slope can be anything, even if the CAPM holds. OLS uses the identity in place of the covariance matrix of stock return residuals. With Generalized Least Squares (GLS), however, the estimated regression slope must be close to the one implied by the CAPM, if the CAPM holds and the market proxy is close to the true market. GLS require an estimator of the covariance matrix of residuals.[1] Where to find it? Usually, the sample covariance matrix is out of the question because it is near-singular or singular. We show

[1] The term GLS sometimes means using the true covariance matrix; here, just an estimator.

that a linear combination of the identity and the sample covariance matrix can be used to run GLS regressions.

Brown (1989) finds that APT tests based on sample covariance matrix eigenvalues are extremely sensitive to the relative magnitudes of the number of time periods T and the number of stocks N. His results are obtained by Monte-Carlo simulations in a stylized case. We review an equation that gives the distribution of sample eigenvalues as a function of the distribution of true eigenvalues and the ratio N/T, when T goes to infinity. Potentially, it could be used to correct APT tests for the effect noticed by Brown.

To the best of our knowledge, the only published results on the sample covariance matrix when N goes to infinity with T characterize eigenvalues. This literature is part of the spectral theory of large-dimensional random matrices. Marčenko and Pastur (1967) first obtained its central equation, which is the one that we alluded to in the previous paragraph. The most recent and general result is by Silverstein (1994). We could only find two statistical applications in this literature: Wachter (1976) and Silverstein and Combettes (1992). Both are restricted to special cases, and study only eigenvalues. By contrast, we work in the general case, and are interested in the whole sample covariance matrix.

2 Sample Covariance Matrix

We analyze the behavior of the sample covariance matrix when the number of variables is large, the typical case for portfolio selection with stocks.

2.1 Model

Consider a very simple situation where we relax the standard asymptotic assumption that keeps the number of variables fixed.

Assumption 1 *Let* $T = 1, 2, \ldots$ *index a sequence of statistical models. For every* T, X_T *is an* $N_T \times T$ *matrix of* T *independent and identically distributed (iid) observations on a system of* N_T *random variables with mean zero and* $N_T \times N_T$ *covariance matrix* $\Sigma_T = \mathrm{E}[(1/T)X_T X_T']$, *where* $\mathrm{E}[\cdot]$ *denotes expectation and prime denotes transposition. The sample covariance matrix is*

$\tilde{\Sigma}_T = (1/T)X_T X_T'$. *Assume that there exists a constant A independent of T such that $N_T \leq A\,T$.*

All the quantities in this paper depend on T unless otherwise specified. For fluidity we omit the subscript T. Assumption 1 prevents the number of variables N from growing infinitely faster than the number of observations T.

The assumption that the random variables have mean zero is not restrictive because, in practice, we can always subtract some estimator of mean returns. How to estimate mean returns is strictly outside the scope of this paper.

Decompose the covariance matrix into eigenvectors and eigenvalues: $\Sigma = U\Lambda U'$, where U is a rotation matrix ($U'U = UU' = I$ the identity matrix) whose columns are the eigenvectors of Σ, and Λ a diagonal matrix whose diagonal elements are the eigenvalues of Σ. Define $Y = U'X$, an $N \times T$ matrix of T iid observations on a system of N uncorrelated random variables that spans the same space as the original system.

We must impose some cross-sectional restrictions in order to obtain results when we allow N to grow without bounds.

Assumption 2 *Let $(y_{11}, \ldots, y_{N1})'$ denote the first column of the matrix Y. The average eighth moment is bounded in the following sense: there exists a constant B independent of T such that* $\mathrm{E}[(1/N)\sum_{i=1}^{N} y_{i1}^8] \leq B$.

Assumption 3 $\mathrm{Cov}[y_{i1}y_{j1}, y_{k1}y_{l1}] = 0$ *when the set $\{i, j\}$ does not intersect with the set $\{k, l\}$.*

Assumptions 1-3 are implicit throughout the remainder of the paper.

2.2 Norm

The originality of this framework is that the dimension N of the covariance matrix can change as T goes to infinity, and can even go to infinity itself: the space where the covariance matrix lives is changing. This makes the definition of a norm on covariance matrices delicate, but not impossible.

Two solutions come to mind: either define a norm on an infinite-dimensional space into which every finite-dimensional space can be embedded, or define a sequence of norms directly on the finite-dimensional spaces. I choose the second option because it sticks closer to practice, where the dimension of the space, however large, is finite.

The sequence of norms (one norm corresponding to each dimension N) is built around the Frobenius norm, which is often used in linear algebra.

Definition 1 *The norm of the $N \times N$ symmetric matrix S with entries $(s_{ij})_{i,j=1,\ldots,N}$ and eigenvalues $(l_i)_{i=1,\ldots,N}$ is defined by:*

$$||S||^2 = c_N \mathrm{tr}\left(S^2\right) = c_N \sum_{i=1}^{N}\sum_{j=1}^{N} s_{ij}^2 = c_N \sum_{i=1}^{N} l_i^2, \qquad (1)$$

where tr *denotes the trace and c_N is a scalar coefficient. This norm is a quadratic form on the linear space of $N \times N$ symmetric matrices. Its associated inner product is: $S_1 \circ S_2 = c_N \mathrm{tr}(S_1 S_2)$, where S_1 and S_2 are $N \times N$ symmetric matrices.*

It is attractive for the squared norm of a matrix to accumulate the squares of individual entries. The coefficient c_N controls the asymptotic behavior of the sequence of norms. Rigorously speaking, the symbol for the norm $||\cdot||$ should be bearing the subscript N.

In order to complete the construction of the sequence of norms, we must choose what asymptotic properties we want to impose on it, and determine the sequence of coefficients c_N accordingly. Remember that an N-dimensional matrix represents a linear operator on the space of N-dimensional vectors. A desirable property is that the norm of familiar linear operators remains well-behaved as N goes to infinity.

The standard definition of the Frobenius norm uses $c_N = 1$. This may be appropriate for the standard case where the dimension N is fixed, but it would cause severe paradoxes as N goes to infinity. For example, it would make the norm of the identity matrix go to infinity with N. This is not acceptable because, as a linear operator, the identity leaves vectors unchanged, and this operation is too mild to deserve an infinite norm.

The problem with $c_N = 1$ is that the norm of a sequence of matrices could increase just because their dimension increases. All other things equal, norms and distances would be greater, the greater the dimension. In a general sense, distances would be larger between two high-dimensional matrices than between two low-dimensional ones. To present an analogy, it would be as ill-advised as measuring in the same unit the distance between two cities and the distance between two galaxies.

This paradox is resolved by defining a *relative* distance. The distance between two N-

dimensional matrices is divided by the distance between two benchmark matrices of the same dimension N. Relative distance corrects for the potentially disturbing impact of dimension. The benchmark must be chosen carefully. I take the benchmark as the distance from the null matrix to the identity. In other words, the N-dimensional identity matrix always has norm one. This convention determines c_N uniquely.

Definition 2 *The scalar coefficient not specified by Definition 1 is:* $c_N = 1/N$.

Any choice of c_N such that the norm of the identity remains bounded away from zero and infinity would induce a norm equivalent to Definition 2's. This is a very large class, and arguably it contains any norm that would make sense in this context. Equivalence means that the notions of convergence, consistency and continuity are blind to the particular norm in the class. We can thus be confident that Definitions 1-2 capture an intuitively satisfying notion of norm.

A simple example illustrates the asymptotic behavior of the norm $\|\cdot\|$ defined above. Let M_1 denote the $N \times N$ matrix with one in its top left entry and zeros everywhere else. Let M_0 denote the $N \times N$ matrix with zeros everywhere (i.e. the null matrix). M_1 and M_0 differ in a way that is independent of N: the top left entry is not the same. Yet the squared distance $\|M_1 - M_0\|^2 = 1/N$ depends on N.

This apparently surprising remark has an intuitive explanation. M_1 and M_0 disagree on the first dimension, but they agree on the $N - 1$ others. The importance of their disagreement is relative to the extent of their agreement. If $N = 1$, then M_1 and M_0 have nothing in common, and their distance is 1. If $N \to \infty$, then M_1 and M_0 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 lost among many others).

It was important to take the time to define the "right" norm because results about consistency are only as interesting as the norm that they are obtained under. If we want the appealing features of the Frobenius norm, it seems that the above choice is the only one (up to equivalence) that makes any sense as N goes to infinity.

Even though Definition 2 is crucial for theoretical results of consistency, it does not matter at all in practice. As will be seen later, the usefulness of this paper from an empirical point of view is to estimate consistently shrinkage intensities (the scalars m and r_2^2/d^2, see Section 3.2) that are *ratios* of norms or inner products of N-dimensional matrices. Therefore the scalar coefficient c_N will cancel itself out from every formula used in practice.

2.3 Consistency

Let $m = \Sigma \circ I$, where I is the identity. The scalar m measures the scale of the covariance matrix. m is the average of the diagonal elements and also the average of the eigenvalues of Σ. The scalar multiple of the identity closest to Σ is mI. mI is the orthogonal projection of Σ onto the line spanned by I. If $I \circ I = \|I\|^2$ was not equal to one, then the correct definition would be: $m = (\Sigma \circ I)/(I \circ I)$.

The mean squared error of the sample covariance matrix is of order N/T.

Theorem 1 $\mathrm{E}[\|\tilde{\Sigma} - \Sigma\|^2] - (N/T)\, m^2 \to 0$, *where convergence is meant as T goes to infinity.*

When N/T does not vanish, which is the general case under Assumption 1, the sample covariance matrix is not consistent. When N/T vanishes, which is a special case of Assumption 1, the sample covariance matrix is consistent. In particular, when N is bounded, our framework degenerates to standard asymptotics.

$\tilde{\Sigma}$ is not consistent because of its off-diagonal elements. Granted, the variance of each one of them vanishes in $1/T$, but so many of them accumulate that the error of $\tilde{\Sigma}$ as a whole does not vanish.

$T = 2,000$ time periods might sound like a lot, but it is not enough if we have as many as $N = 1,000$ stocks: it is about as bad as using two observations to estimate the variance of one random variable. 1,000 is less than half the number of stocks trading on the New York Stock Exchange (NYSE) alone. In order to estimate a $1,000 \times 1,000$ covariance matrix accurately, we need at least, say, 10,000 observations, which means 40 years of daily data, longer than the Center for Research in Security Prices (CRSP) database holds, and in any case long enough for nonstationarity to become a major concern.

Even though we have not tried to obtain a formal proof, we firmly believe that no other covariance matrix estimator is consistent under Assumptions 1-3. Yet all hope is not lost. More than its existence, it is the nature of this error that hurts portfolio selection. We will soon see that the heart of the problem lies in the smallest eigenvalues of the sample covariance matrix. First, we review the importance of covariance matrix eigenvalues for portfolio selection.

2.4 Portfolio Selection and Covariance Matrix Eigenvalues

Markowitz (1952) considers the problem of selecting the $N \times 1$ vector of weights w of a portfolio of N stocks whose returns have $N \times N$ covariance matrix Σ, under the K linear constraints defined by the $N \times K$ matrix of coefficients C and the $K \times 1$ right-hand-side vector γ. The objective is to minimize the variance of portfolio returns:

$$\begin{aligned} &\min_{w} w'\Sigma w \\ &\text{s.t.} \quad C'w = \gamma \end{aligned} \tag{2}$$

$$\rightarrow \quad w = \Sigma^{-1}C\left(C'\Sigma^{-1}C\right)^{-1}\gamma \tag{3}$$

Typical constraints impose that weights sum to one and portfolio returns have a required expectation.

Recall the decomposition $\Sigma = U\Lambda U'$. Let $u_1, \ldots, u_N$ denote the columns of U, i.e. the eigenvectors of Σ. Let $\lambda_1, \ldots, \lambda_N$ denote the diagonal terms of Λ, i.e. the eigenvalues of Σ. Let $C_* = C(C'\Sigma^{-1}C)^{-1}\gamma$. It is the linear combination of constraints where the coefficient of each constraint is its shadow price. Then Equation (3) can be rewritten as $w = \Sigma^{-1}C_* = U\Lambda^{-1}U'C_*$, or as:

$$w = \sum_{i=1}^{N} \frac{C_*'u_i}{\lambda_i} u_i. \tag{4}$$

The constrained minimum variance portfolio spreads its weight across the eigenvectors of Σ. The weight on eigenvector u_i is inversely proportional to its eigenvalue λ_i. λ_i is the variance of returns on the portfolio with weights u_i. It measures the riskiness of u_i. If an eigenvector is less risky, it receives more weight; riskier, less weight. This is the mathematical translation of the economic idea of diversification. Spreading weights across eigenvectors is like putting all the eggs in different baskets.

In practice, Σ is not known, so we can be tempted to replace it with the sample covariance matrix $\tilde{\Sigma}$. Decompose it into $\tilde{\Sigma} = \tilde{U}\tilde{\Lambda}\tilde{U}'$, where $\tilde{U}$ is the rotation matrix whose columns $\tilde{u}_1, \ldots, \tilde{u}_N$ are the eigenvectors of $\tilde{\Sigma}$, and $\tilde{\Lambda}$ the diagonal matrix whose diagonal terms $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N$ are the eigenvalues of $\tilde{\Sigma}$. Portfolio selection with $\tilde{\Sigma}$ yields weights $\tilde{w} = \sum_{i=1}^{N} (\tilde{C}_*' \tilde{u}_i / \tilde{\lambda}_i) \tilde{u}_i$, where $\tilde{C}_* = C(C'\tilde{\Sigma}^{-1}C)^{-1}\gamma$.

The true riskiness of eigenvector $\tilde{u}_i$ is $\tilde{u}_i' \Sigma \tilde{u}_i$, estimated by $\tilde{u}_i' \tilde{\Sigma} \tilde{u}_i = \tilde{\lambda}_i$. If $\tilde{\lambda}_i$ is close to zero but $\tilde{u}_i' \Sigma \tilde{u}_i$ is not, it is a catastrophe. Since weight is in $1/\tilde{\lambda}_i$, if $\tilde{\lambda}_i$ is near zero by mistake, nearly infinite weight falls on an eigenvector that is not truly riskless. It is like putting all the eggs in the same basket, and discovering that it is not safe when all the eggs get broken. A covariance matrix estimator for portfolio selection must refrain from having eigenvalues near zero, unless there is convincing evidence that it is no mistake. This is the same as saying that the covariance matrix must not be singular or even near-singular, an idea already known to Michaud (1989).

Next, we show that some eigenvalues of the sample covariance matrix are systematically too close to zero by mistake, when N is not negligible with respect to T. The sample covariance matrix is typically singular or near-singular in practical applications. This is what makes it ill-suited to portfolio selection.

2.5 Sample Covariance Matrix Eigenvalues

We are trying to show that the smallest eigenvalues of the sample covariance matrix are biased towards zero. Since they are constrained to be nonnegative, we need to show that they are biased downwards. The full picture is that the smallest eigenvalues are biased downwards and the largest ones upwards. This statement is equivalent to saying that sample eigenvalues are too dispersed.

Theorem 2 *Sample eigenvalues have approximately the same average as true ones, in the sense that* $\mathrm{E}[(1/N)\sum_{i=1}^{N} \tilde{\lambda}_i] = (1/N)\sum_{i=1}^{N} \lambda_i$ *and* $\mathrm{Var}[(1/N)\sum_{i=1}^{N} \tilde{\lambda}_i] \to 0$.

Yin (1986) proves a more general version of this result, but under stronger assumptions. Recall from above that $m = \Sigma \circ I = (1/N)\sum_{i=1}^{N} \lambda_i$.

Theorem 3 *Sample eigenvalues are more dispersed than true ones, in the sense that:*

$$\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{\lambda}_i - m\right)^2\right] = \frac{1}{N}\sum_{i=1}^{N}(\lambda_i - m)^2 + \mathrm{E}\left[\left\|\tilde{\Sigma} - \Sigma\right\|^2\right] \tag{5}$$

Yin (1986) proves a related result under stronger assumptions.[2]

$\tilde{\Sigma}$ uses all of its error to feed an increase in the dispersion of its eigenvalues. It is as if $\tilde{\Sigma}$ wanted to have the most dispersed eigenvalues, and used all that differentiates it from Σ to beat Σ at this game. Theorem 3 implies that the smallest eigenvalues of $\tilde{\Sigma}$ are biased downwards (towards zero), and the largest ones upwards. Ironically, it is due to the fact that sample covariance matrix *entries* are *un*biased, as is apparent form the proof of Theorem 3.

A property of eigenvalues helps understand the mechanism at work.

Theorem 4 *The eigenvalues are the most dispersed diagonal elements that can be obtained by rotation.*

Since $\tilde{\Sigma}$ is unbiased and U is nonstochastic, $U'\tilde{\Sigma}U$ is an unbiased estimator of $U'\Sigma U$. The diagonal elements of $U'\tilde{\Sigma}U$ are approximately as dispersed as the ones of $U'\Sigma U$. For convenience, let us speak as if they were exactly as dispersed. By contrast, $\tilde{U}'\tilde{\Sigma}\tilde{U}$ is not an unbiased estimator of $\tilde{U}'\Sigma\tilde{U}$. This is because the errors of $\tilde{U}$ and $\tilde{\Sigma}$ strongly interact. By Theorem 4, the diagonal elements of $\tilde{U}'\tilde{\Sigma}\tilde{U}$ are more dispersed than those of $U'\tilde{\Sigma}U$ and $U'\Sigma U$. This is why sample eigenvalues are more dispersed than true ones.

Evidence against the sample covariance matrix is even more damning than Theorem 3 suggests, because $\tilde{\lambda}_i = \tilde{u}_i'\tilde{\Sigma}\tilde{u}_i$ should not be compared to $\lambda_i = u_i'\Sigma u_i$, but to $\tilde{u}_i'\Sigma\tilde{u}_i$. We should compare estimated vs. true riskiness of eigenvector $\tilde{u}_i$. In portfolio selection, we entrust our money to $\tilde{u}_i$ based on $\tilde{u}_i'\tilde{\Sigma}\tilde{u}_i$, and we end up bearing the risk $\tilde{u}_i'\Sigma\tilde{u}_i$. By Theorem 4 again, the diagonal elements of $\tilde{U}'\Sigma\tilde{U}$ are even less dispersed than those of $U'\Sigma U$. Not only are sample eigenvalues more dispersed than true ones, but they should be less dispersed! Intuitively: statisticians should shy away from taking a strong stance on extremely small and large eigenvalues, because they know that they don't know everything. The sample covariance matrix is guilty of taking an unjustifiably strong stance.

How important is this effect in practice? When variables outnumber observations, it is infinitely important. Since $\tilde{\Sigma} = (1/T)XX'$ and the dimension of X is $N \times T$, the rank of $\tilde{\Sigma}$ is the minimum of N and T. When $N > T$, the rank of $\tilde{\Sigma}$ is less than its dimension N. $\tilde{\Sigma}$ is rank-deficient. This

[2] He proves that $(1/N)\sum_{i=1}^{N}(\tilde{\lambda}_i - m)^2 - \{(1/N)\sum_{i=1}^{N}(\lambda_i - m)^2 + (N/T)\, m^2\} \to 0$ in probability. His result follows from Theorems 1 and 3.

means that it is singular and that some of its eigenvalues are equal to zero. It cannot be inverted and used for portfolio selection.

By continuity, we expect the sample covariance matrix to become near-singular as the ratio N/T gets close to one. In order to see how sample covariance matrix eigenvalues change in the ratio N/T, we look more closely at a particular case. It is our experience that what follows is representative of the general case.

2.6 Particular Case: the Identity Matrix

To illustrate how dangerous the sample covariance matrix is for portfolio selection, we analyze in more detail the particular case $\Sigma = I$. Assuming that the ratio N/T converges to a finite positive limit c called the concentration, Marčenko and Pastur (1967) derive the limit of the distribution of sample eigenvalues.

A popular way to graph eigenvalues is to sort them in descending order, and plot the eigenvalues as a function of their rank. We follow this convention, with one adjustment due to the fact that the number of eigenvalues goes to infinity. We plot the eigenvalues as a function of their relative rank, defined as the rank divided by the total number of eigenvalues. As N goes to infinity, the relative rank remains between zero (largest eigenvalues) and one (smallest).

By assumption, $\Sigma = I$, therefore true eigenvalues are all equal to one. Their graph is a horizontal line at one. Figure 1 plots sample eigenvalues for various concentrations, as given by Marčenko and Pastur's asymptotic approximation. If concentration was zero, sample eigenvalues would also plot as a horizontal line at one. However, for positive concentrations, even small ones, the smallest eigenvalues are substantially biased towards zero. Bias becomes more severe as concentration increases to one. When $c > 1$, the smallest eigenvalues are *equal* to zero.

Figure 1 speaks against using the sample covariance matrix for portfolio selection unless N is negligible with respect to T, which is rarely the case in practice. From the above discussion, it is because the sample covariance matrix uses the accumulation of errors off the diagonal to bias the smallest eigenvalues downwards and the largest ones upwards. This is a widespread phenomenon. For example, it is well-known that the smallest estimated betas are biased downwards and the largest ones upwards. It can even be said that this phenomenon plays an important role in the popularity of alternatives to the maximum likelihood such as Bayesian statistics and decision theory. It is particularly pronounced here because the excess dispersion of sample eigenvalues is in N/T, instead of e.g. $1/T$ for betas. Also, it is particularly damaging, because the downwards bias of the smallest eigenvalues, when it draws them close to zero, has infinitely destructive consequences on portfolio selection.

The bias of the eigenvalues of the sample covariance matrix is intimately related to the unbiasedness of its entries. To put it bluntly, either the eigenvalues or the entries must be biased: we cannot have it both ways. Equation (4) makes it clear that portfolio selection calls for minimally biased eigenvalues, even if the price to pay is to bias the entries. This is the topic of Section 3.

2.7 Potential Applications to Tests for the Number of Factors in the APT

Some of the plots in Figure 1 bear a striking resemblance to plots of the eigenvalues of the sample covariance matrix of stock returns in tests for the number of factors in the APT. There, the emphasis is not on the smallest eigenvalues, but on the largest ones: are they large enough to support the APT? As can be seen from Figure 1, the largest sample eigenvalues are severely biased upwards, therefore inference must be drawn cautiously. This is the point made by Brown (1989), based on Monte-Carlo simulations. The review by Connor and Korajczyk (1992) makes it clear that this is a pervasive problem in the literature.

Marčenko and Pastur (1967) solve much more than the special case $\Sigma = I$. They derive a general equation that yields the distribution of sample eigenvalues as a function of the distribution of true eigenvalues and the concentration. An original approach to APT tests would be to use this equation in reverse to back up true eigenvalues from sample eigenvalues. This is an appealing direction for future research, but there is one obstacle. It is an ill-posed problem.

Infinitesimal errors on the estimation of sample eigenvalues are amplified into large errors on true eigenvalues as we go through the equation in reverse. For example, Black and Scholes (1973) obtain a partial differential equation that determines the value $V(S,t)$ of a European option as a function of the stock price S and time t. They know $V(\cdot, t_2)$ at expiration date t_2, and want $V(\cdot, t_1)$ today at $t_1 < t_2$. This is a well-posed problem. Reverse the direction of time and it becomes an ill-posed problem. It would not be possible to deduce $V(\cdot, t_2)$ from $V(\cdot, t_1)$ for $t_2 > t_1$. More precisely, a lot of very different solutions $V(\cdot, t_2)$ correspond to almost exactly the same initial conditions $V(\cdot, t_1)$. Fortunately for option pricing, time flows in the right direction.

The distribution of sample eigenvalues is a smoothed-out version of the distribution of true eigenvalues. It is a general fact that "un-smoothing" is an ill-posed problem. Figuratively, this is because the resolution of the picture is diminished by the action of smoothing. In our case, it is the error of sample eigenvectors that smoothes out true eigenvalues into sample eigenvalues. For option pricing, it is the uncertainty about the terminal value of the stock price that makes today's option value $V(\cdot, t_1)$ smoother than the terminal payoff $V(\cdot, t_2)$.

Ill-posedness makes it hard to obtain reliable estimators of true eigenvalues. Getting confidence intervals is probably even harder. Not surprisingly, the degree of ill-posedness increases in the ratio N/T. We interpret it as: we cannot get something for nothing. We firmly believe that ill-posedness is not an artifact of the Marčenko and Pastur equation, but a deep feature of the problem itself.

However, the degree of ill-posedness is not uniform. The problem is better posed around isolated eigenvalues. In practice, we expect the largest eigenvalues to be quite isolated. This may be what makes it possible to recover them. Some more details are in Appendix A. For a different and innovative approach, see Adamek (1994).

3 Improved Covariance Matrix Estimation

We derive an estimator that improves over the sample covariance matrix when the number of variables N is not negligible with respect to the number of observations T. Generalizations are described.

3.1 Linear Shrinkage of Sample Eigenvalues

As we saw in Section 2, the problem with the sample covariance matrix is that its eigenvalues can be too dispersed. The line of attack is suggested by established methods in multivariate statistics. Muirhead (1987) reviews decision-theoretic alternatives to the sample covariance matrix and concludes that they "have a tendency to move the sample eigenvalues together in an intuitively appealing way." Shrinking sample eigenvalues together is attractive for portfolio selection because it reduces singularity by pulling the smallest eigenvalues away from zero. We follow this approach.

To simplify matters, we focus on linear shrinkage. That is, we consider improved eigenvalues estimators of the form $\widehat{\lambda}_i = \alpha + \beta\widetilde{\lambda}_i$, $i = 1, \ldots, N$, where α and β are scalars.[3] This is equivalent to replacing $\widetilde{\Lambda}$ with $\widehat{\Lambda} = \alpha I + \beta\widetilde{\Lambda}$. Following the decision-theoretic literature, we keep the same eigenvectors as the sample covariance matrix. The improved estimator is: $\widehat{\Sigma} = \widetilde{U}\widehat{\Lambda}\widetilde{U}' = \widetilde{U}(\alpha I + \beta\widetilde{\Lambda})\widetilde{U}' = \alpha I + \beta\widetilde{\Sigma}$.

The central question is to find the coefficients α and β. If we were only trying to avoid singularity, the choice of α and β would be ad-hoc. Instead, we ought to be minimizing some criterion. A natural candidate is the mean squared error:

$$\begin{aligned} &\min_{\alpha,\beta} \mathrm{E}\left[\left\|\widehat{\Sigma} - \Sigma\right\|^2\right] \\ &\text{s.t.} \quad \widehat{\Sigma} = \alpha I + \beta\widetilde{\Sigma}. \end{aligned} \tag{6}$$

Is it compatible with the need to avoid singularity? $\mathrm{E}[\|\widehat{\Sigma} - \Sigma\|^2] = \mathrm{E}[\|\widetilde{U}'\widehat{\Sigma}\widetilde{U} - \widetilde{U}'\Sigma\widetilde{U}\|^2] = \mathrm{E}[(1/N)\sum_{i=1}^{N}(\widetilde{u}_i'\widehat{\Sigma}\widetilde{u}_i - \widetilde{u}_i'\Sigma\widetilde{u}_i)^2]$ + constant, where the constant does not depend on α and β. Therefore choosing α and β to minimize mean squared error is the same as choosing them to minimize the distance between the estimated riskiness $\widetilde{u}_i'\widehat{\Sigma}\widetilde{u}_i$ of eigenvector $\widetilde{u}_i$ and its true riskiness $\widetilde{u}_i'\Sigma\widetilde{u}_i$, on average across $i = 1, \ldots, N$. For portfolio selection, this is a very good criterion, since so much rides on estimating the riskiness of each eigenvector well. The mean squared error criterion is in alignment with the objectives of portfolio selection. Even more alignment could conceivably be achieved, for example by letting the criterion depend on the matrix of portfolio selection constraints C (cf. Equation (2)), but this is left to future research.[4]

$\widetilde{u}_1'\Sigma\widetilde{u}_1, \ldots, \widetilde{u}_N'\Sigma\widetilde{u}_N$ are even less dispersed than true eigenvalues, so we anticipate that our estimator's eigenvalues will be less dispersed than true ones. This should keep the smallest eigenvalues of $\widehat{\Sigma}$ safely away from zero.

[3] One advantage of linear shrinkage is that it preserves the ordering of the eigenvalues (if $\beta \geq 0$), an intuitively appealing property whose theoretical importance is proven by Sheena and Takemura (1992).

[4] I thank Fischer Black for this suggestion.

3.2 Optimal Linear Shrinkage

If we could observe the true covariance matrix Σ, we could easily solve Equation (6).

Theorem 5 *Let* $m = \Sigma \circ I$. *Let* $r_1^2 = \|\Sigma - mI\|^2$, $r_2^2 = \mathrm{E}[\|\tilde{\Sigma} - \Sigma\|^2]$ *and* $d^2 = \mathrm{E}[\|\tilde{\Sigma} - mI\|^2]$. *The solution* $\hat{\Sigma}$ *to Equation (6) is:*

$$\hat{\Sigma} = \frac{r_2^2}{d^2}\, mI + \frac{r_1^2}{d^2}\tilde{\Sigma}. \tag{7}$$

Its mean squared error is $\mathrm{E}[\|\hat{\Sigma} - \Sigma\|^2] = r_1^2 r_2^2 / d^2 < \min(r_1^2, r_2^2)$.

By Theorem 3, $r_1^2 + r_2^2 = d^2$, so $\hat{\Sigma}$ is a weighted average of mI and $\tilde{\Sigma}$. The weight placed on mI increases with the error of $\tilde{\Sigma}$ and decreases with the error of mI. For the weight on $\tilde{\Sigma}$, it is the opposite. The dispersion of the eigenvalues of $\hat{\Sigma}$ is $\mathrm{E}[\|\hat{\Sigma} - mI\|^2] = r_1^4/d^2 < r_1^2$: the eigenvalues of $\hat{\Sigma}$ are even less dispersed than Σ's. This effect becomes more pronounced as the error of $\tilde{\Sigma}$ increases, i.e. as the ratio N/T increases. $\hat{\Sigma}$ is the projection of Σ onto the line between mI and $\tilde{\Sigma}$. Figure 2 shows this geometrical interpretation.

Unfortunately, $\hat{\Sigma}$ is not an estimator because it depends on the unobservable matrix Σ. As we saw, in general it is impossible to estimate Σ consistently. However, we do not need all the entries of Σ: the four parameters m, r_1^2, r_2^2 and d^2 suffice. The key insight of this paper is that, as T goes to infinity, even if N goes to infinity too, it is possible to estimate these four parameters consistently.

First, Theorem 2 reveals that m can be estimated simply by $\widehat{m} = (1/N)\sum_{i=1}^{N} \tilde{\lambda}_i$: the average of sample eigenvalues is a consistent estimator of the average of true eigenvalues. Second, a natural estimator of $d^2 = \mathrm{E}[\|\tilde{\Sigma} - mI\|^2]$ is $\hat{d}^2 = \|\tilde{\Sigma} - \widehat{m}I\|^2$.

Theorem 6 $\hat{d} - d \xrightarrow{\mathrm{P}} 0$, *where* $\xrightarrow{\mathrm{P}}$ *denotes convergence in probability as* T *goes to infinity.*

Third, let the $N \times 1$ vector $x_{\cdot t}$ denote the t^{th} column of the observations matrix X for $t = 1, \ldots, T$. $\tilde{\Sigma} = (1/T)XX'$ can be rewritten as $\tilde{\Sigma} = (1/T)\sum_{t=1}^{T} x_{\cdot t}x'_{\cdot t}$. $\tilde{\Sigma}$ is the average of the matrices $x_{\cdot t}x'_{\cdot t}$ $(t = 1, \ldots, T)$. Since the matrices $x_{\cdot t}x'_{\cdot t}$ are iid across $t = 1, \ldots, T$, we can estimate the error $d^2 = \mathrm{E}[\|\tilde{\Sigma} - \Sigma\|]$ of their average by seeing how far each one of them deviates from the average.

Theorem 7 *Define* $\hat{r}_2^2 = (1/T^2)\sum_{t=1}^{T} \|x_{\cdot t}x'_{\cdot t} - \tilde{\Sigma}\|^2$. *Then* $\hat{r}_2^2 - r_2^2 \xrightarrow{\mathrm{P}} 0$.

Finally, Theorem 3 can be rewritten as $r_1^2 = d^2 - r_2^2$.

Theorem 8 *Define* $\widehat{r}_1^2 = \widehat{d}^2 - \widehat{r}_2^2$. *Then* $\widehat{r}_1^2 - r_1^2 \xrightarrow{\mathrm{P}} 0$.

If, for a given realization, $\widehat{d}^2 < \widehat{r}_2^2$, then we recommend correcting $\widehat{d}^2$ and/or $\widehat{r}_2^2$ so that they are equal. It can be shown that this does not affect the validity of the theorems.

Please note that Theorems 6, 7 and 8 are non-trivial since, in spite of the division by N in the definition of the norm $\|\cdot\|$, the scalars d, r_1, and r_2 do not converge to zero (except in special cases), as is apparent from the proofs.

Plugging consistent estimators in place of the unobservable parameters in Equation (7) yields a consistent estimator of $\widehat{\Sigma}$ with the same asymptotic properties. This is the main result of the paper.

Theorem 9 *The improved estimator*

$$\hat{\hat{\Sigma}} = \frac{\widehat{r}_2^2}{\widehat{d}^2}\,\widehat{m}I + \frac{\widehat{r}_1^2}{\widehat{d}^2}\tilde{\Sigma} \tag{8}$$

estimates the solution $\widehat{\Sigma}$ *to Equation (6) consistently, i.e.* $\|\hat{\hat{\Sigma}} - \widehat{\Sigma}\|^2 \xrightarrow{\mathrm{P}} 0$. *Both* $\hat{\hat{\Sigma}}$ *and* $\widehat{\Sigma}$ *have the same asymptotic mean squared error, i.e.* $\mathrm{E}[\|\hat{\hat{\Sigma}} - \Sigma\|^2] - \mathrm{E}[\|\widehat{\Sigma} - \Sigma\|^2] \to 0$, *and* $\widehat{r}_1^2\widehat{r}_2^2/\widehat{d}^2$ *estimates it consistently, i.e.* $(\widehat{r}_1^2\widehat{r}_2^2/\widehat{d}^2) - (r_1^2 r_2^2/d^2) \xrightarrow{\mathrm{P}} 0$.

$\hat{\hat{\Sigma}}$ is an improved estimator of the covariance matrix. It is a consistent estimator of the linear combination of the sample covariance matrix with the identity matrix that minimizes mean squared error. It is easy to verify that $\hat{\hat{\Sigma}}$ is invariant by rotation, i.e. premultiplying the observations X by a rotation matrix V ($V'V = VV' = I$) changes $\hat{\hat{\Sigma}}$ into $V\hat{\hat{\Sigma}}V'$.

By Theorem 1, the weight on $\widehat{m}I$ increases in N/T. If N remains bounded, asymptotically all the weight is on the sample covariance matrix $\tilde{\Sigma}$.

The advantage of our framework over finite sample statistics is that we do not have to take into account the error of estimators of the unobservable parameters m, r_1^2, r_2^2 and d^2. The advantage over standard asymptotics is that we encompass realistic situations where the sample covariance matrix is not optimal.

3.3 Generalization

$\hat{\hat{\Sigma}}$ is a weighted average of $\widehat{m}I$ and $\tilde{\Sigma}$. $\widehat{m}I$ can be thought of as an estimator of the covariance matrix. It has asymptotically minimum mean squared error among a certain class of estimators:

scalar multiples of the identity matrix. This class imposes a lot of structure on the covariance matrix: no covariances, and all variances are the same. There is only one free parameter, as opposed to $N(N+1)/2$ for $\tilde{\Sigma}$. This parsimonious structure makes $\widehat{m}I$ heavily biased, but at least it prevents it from being singular, a problem that hurts the unstructured, unbiased estimator $\tilde{\Sigma}$.

Other structures can be imposed on the covariance matrix. Frost and Savarino (1986) impose that all stock returns have the same variance and all pairs of stock returns have the same covariance. They have two free parameters. We can also impose that the covariance matrix is diagonal (N parameters), or that all correlation coefficients are equal ($N + 1$ parameters).

We call such estimators: "structured." Other structured estimators of interest in Finance are the index models. For example, Sharpe's (1963) single index model assumes that the idiosyncratic risks of different stocks are uncorrelated. The idiosyncratic risk is the fraction of the risk that is not systematic risk. Systematic risk is the fraction of the risk that can be explained as covariance with an index, usually a broad-based market index. In general, if there are K indices, then we need to estimate the covariance matrix of the indices ($K(K+1)/2$ parameters), the covariance of each stock with each index (KN parameters) and each stock's idiosyncratic risk (N parameters), for a total of $(K+1)(N+K/2)$ free parameters. When $K \ll N$, this is still much fewer parameters than the sample covariance matrix.

Structured estimators are popular for portfolio selection. They are carefully designed to avoid the singularity problem of the sample covariance matrix. Their main selling point is that they do not place infinite weights on risky eigenvectors by mistake.

However, the way that they obtain this desirable feature is ad-hoc. They impose arbitrary structure that they know is wrong, then disregard any evidence that goes against it. They throw away all sample information that does not fit in their arbitrarily specified structure. It would be better to recycle the information that they ignore, in an optimal way. We recommend taking a well-chosen weighted average of a structured estimator and the sample covariance matrix.

Let $\bar{\Sigma}$ denote any given structured estimator of interest to the statistician. Consider the problem:

$$\begin{aligned} &\min_{\omega} \mathrm{E}\left[\left\|\hat{\Sigma} - \Sigma\right\|^2\right] \\ \text{s.t.}\quad &\hat{\Sigma} = \omega\bar{\Sigma} + (1-\omega)\tilde{\Sigma}. \end{aligned} \tag{9}$$

$\widehat{\Sigma}$ is a weighted average of two estimators, one generally singular ($\widetilde{\Sigma}$), and the other one generally not ($\overline{\Sigma}$). Which one does it inherit its properties from? An elementary result from matrix algebra answers.

Proposition 1 *The smallest eigenvalue of* $\widehat{\Sigma} = \omega\overline{\Sigma} + (1-\omega)\widetilde{\Sigma}$ *is at least as large as* ω *times the smallest eigenvalue of* $\overline{\Sigma}$*.*

$\overline{\Sigma}$ is constructed so that its smallest eigenvalues do not come near zero. Therefore $\widehat{\Sigma}$ is generally not singular, unless ω is very small. If ω was very small, then it would mean that the sample covariance matrix can hardly be improved on. From what we have seen so far, this would be rather surprising when N is of the same order of magnitude as T.

Again, let us pretend for a moment that we can observe Σ. As above, let $r_1^2 = \mathrm{E}[\|\overline{\Sigma} - \Sigma\|^2]$, $r_2^2 = \mathrm{E}[\|\widetilde{\Sigma} - \Sigma\|^2]$ and $d^2 = \mathrm{E}[\|\widetilde{\Sigma} - \overline{\Sigma}\|^2]$. In addition, let $\varphi = \mathrm{E}[(\overline{\Sigma} - \Sigma) \circ (\widetilde{\Sigma} - \Sigma)]$ measure the "covariance" between the errors of both estimators.

Theorem 10 *Then the solution to Equation (9) is given by:*

$$\widehat{\Sigma} = \frac{r_2^2 - \varphi}{d^2}\,\overline{\Sigma} + \left(1 - \frac{r_2^2 - \varphi}{d^2}\right)\widetilde{\Sigma}. \tag{10}$$

The geometric interpretation is the same as in Figure 2, except that $\overline{\Sigma}$ replaces mI and that the triangle $(\overline{\Sigma}, \Sigma. \widetilde{\Sigma})$ does not necessarily have a right angle at Σ anymore. In the particular case $\varphi = 0$, the weight on $\overline{\Sigma}$ reduces to r_2^2/d^2, as in Equation (7). This simplification takes place (asymptotically) for $\overline{\Sigma} = \widehat{m}I$, but not necessarily for other structured estimators.

Again the problem is to estimate the unobservable parameters r_1^2, r_2^2, d^2 and φ consistently. We do not provide formal proofs of consistency, since they would have to be rewritten for every structured estimator $\overline{\Sigma}$. We just indicate how the general logic of the argument for $\overline{\Sigma} = \widehat{m}I$ can be extended to other structured estimators. In Section 4, we provide empirical support for these extensions.

We can take the same estimator $\widehat{r}_2^2$ as before. The estimator of d^2 becomes $\widehat{d^2} = \|\widetilde{\Sigma} - \overline{\Sigma}\|^2$. The additional complication is that we need an estimator $\widehat{\varphi}$ of φ. Since $d^2 = r_1^2 + r_2^2 - 2\varphi$, $\widehat{\varphi}$ would let us estimate r_1^2 by $\widehat{r}_1^2 = \widehat{d^2} - \widehat{r}_2^2 + 2\widehat{\varphi}$.

Let $\overline{\Sigma} = [\overline{\sigma}_{ij}]_{i,j=1,\ldots,N}$ and $\widetilde{\Sigma} = [\widetilde{\sigma}_{ij}]_{i,j=1,\ldots,N}$. Since $\varphi = (1/N)\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{Cov}[\overline{\sigma}_{ij}. \widetilde{\sigma}_{ij}]$, all we need is estimators of $\varphi_{ij} = \text{Cov}[\overline{\sigma}_{ij}, \widetilde{\sigma}_{ij}]$ for $i, j = 1, \ldots, N$. They are usually suggested by the nature of $\overline{\Sigma}$. The idea is that, if we can estimate $\overline{\sigma}_{ij}$, then we can estimate the error on $\overline{\sigma}_{ij}$, and its covariance with the error on $\widetilde{\sigma}_{ij}$. Please keep in mind that φ_{ij} vanishes in $1/T$, even though φ itself may be of order N/T. Therefore, in the more complicated cases, the delta method can be used to estimate φ_{ij} consistently. Given the estimators $\widehat{\varphi}_{ij}$ for $i, j = 1, \ldots, N$, we form $\widehat{\varphi} = (1/N) \sum_{i=1}^{N} \sum_{j=1}^{N} \widehat{\varphi}_{ij}$.

Appendix B gives the formula of $\widehat{\varphi}_{ij}$ $(i, j = 1, \ldots, N)$ for various structured estimators.

3.4 Comparison with Previous Work in Multivariate Statistics

This approach has an obvious Bayesian interpretation. Bayesian statistics combine sample information with other sources of information. The other sources are summarized in a "prior" distribution of the unknown parameter. In our case, the prior distribution puts all its mass on a sphere centered on $\overline{\Sigma}$ with radius $\widehat{r}_1$. Then sample information reveals that the true parameter also lies on another sphere, with center $\widetilde{\Sigma}$ and radius $\widehat{r}_2$. Combining prior and sample yields a posterior distribution. In our case, the true covariance matrix must lie on the intersection of the two spheres. This intersection is a circle. At the center of this circle stands the improved estimator $\widehat{\widehat{\Sigma}}$. This interpretation is shown in Figure 3.

Fundamental Bayesian questions are: Where does the prior come from? How confident are we in the prior? In finite sample, it is very hard to answer these questions satisfactorily. If the statistician chooses the prior without looking at data at all, it might be very inaccurate. Empirical Bayesians do look at data, but then they pretend that they did not, and ignore dependence between prior and sample. In some cases, dependence can safely be neglected, but how do we know that?

By contrast, in our asymptotic framework, we can build the prior around any structured estimator already used in practice. Furthermore, the degree of confidence in the prior can be estimated consistently. In particular, we estimate the parameter φ that captures dependence between prior and sample. We find out for any given prior whether φ can be neglected, and if it cannot be, we account for it in Equation (10).

In the established nomenclature, our work is not pure Bayesian because we estimate the prior from the sample. It is not empirical Bayesian either because it takes into account the dependence between the estimated prior and the sample. It is decision theory.

For the covariance matrix, previous literature on decision theory (and on pure and empirical Bayesian statistics too) has been only in finite sample. The reason is that, under standard asymptotics, the sample covariance matrix is consistent, so there is no need to seek alternatives. Decision theory in finite sample is not very tractable. Also, it relies on the Wishart distribution, which has two limitations: random variables must be normally distributed, and if variables outnumber observations then the Wishart density does not exist (because $\tilde{\Sigma}$ is rank-deficient). For portfolio selection, both limitations are serious.

One of our contributions is to realize that these are not limitations of decision theory itself, but of finite sample. In stock market finance, we are fortunate enough to have large numbers of observations, which make asymptotic approximations realistic, and large numbers of variables, which open the door to improvements over the sample covariance matrix. This is the ideal situation to free decision theory from finite sample drawbacks. All that is needed is to relax the standard asymptotic assumption that keeps the number of variables bounded. $\hat{\hat{\Sigma}}$ is the first estimator of the covariance matrix based on *asymptotic* decision theory.

Stein (1975) suggests that invariance by rotation is an important property for covariance matrix estimators. Intuitively, it means that the statistician lets the data speak without putting a spin on what they say. This excludes all of the structured estimators cited above except $\widehat{m}I$. The existing literature does not contain any estimator invariant by rotation and theoretically motivated when $N > T$. Perhaps more importantly, it contains no estimator that is invariant by rotation and is known not to be singular or near-singular when $N > T$. This has lead some to believe that the inverse of the covariance matrix could not be estimated at all when $N > T$.

Now it can be. The estimator $\hat{\hat{\Sigma}}$ of Section 3.2 is invariant by rotation. It has a sound theoretical motivation when $N > T$. As a matter of fact, it does not even matter whether $N > T$, which is satisfying because we should expect some continuity between $N = 999, T = 1000$ and $N = 1000$, $T = 999$. The eigenvalues of $\hat{\hat{\Sigma}}$ are asymptotically even less dispersed than Σ's, which prevents $\hat{\hat{\Sigma}}$ from being near-singular or singular. The dispersion of the eigenvalues of $\hat{\hat{\Sigma}}$ actually *decreases* in the ratio N/T. Therefore $\hat{\hat{\Sigma}}^{-1}$ is the first estimator of inverse of the covariance matrix that is invariant by rotation and can be used when variables outnumber observations.

4 Application to Portfolio Selection

The goal of this section is to find out how the asymptotic results of Section 3 carry through to large but finite sample. We first compare $\hat{\hat{\Sigma}}$ to other estimators in terms of mean squared error in Monte-Carlo simulations. Then we apply $\hat{\hat{\Sigma}}$ to historical stock returns data.

4.1 Monte-Carlo Simulations

Our purpose is to compare the mean squared errors of various estimators across a range of situations. We focus on estimators that are invariant by rotation, therefore we use Equation (9) for $\hat{\hat{\Sigma}}$.

The benchmark is the mean squared error of the covariance matrix. We report the Percentage Relative Improvement in Average Loss of $\hat{\hat{\Sigma}}$, defined as: $\text{PRIAL}(\hat{\hat{\Sigma}}) = (\text{E}[\|\tilde{\Sigma} - \Sigma\|^2] - \text{E}[\|\hat{\hat{\Sigma}} - \Sigma\|^2])/\text{E}[\|\tilde{\Sigma} - \Sigma\|^2] \times 100$. If the PRIAL is positive (negative), then $\hat{\hat{\Sigma}}$ performs better (worse) than $\tilde{\Sigma}$. The PRIAL of the sample covariance matrix is zero by definition. The PRIAL cannot exceed 100%. We compare the PRIAL of $\hat{\hat{\Sigma}}$ to the PRIAL of other estimators from finite sample decision theory.

Haff (1980) introduces an estimator with an empirical Bayesian interpretation. Like $\hat{\hat{\Sigma}}$, 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 N and T. If the criterion is the mean squared error, Haff's approach suggests:

$$\hat{\Sigma}_{\text{EB}} = \frac{NT - 2T - 2}{NT^2}\,\widehat{m}_{\text{EB}}\,I + \frac{T}{T+1}\tilde{\Sigma} \tag{11}$$

with $\widehat{m}_{\text{EB}} = [\det(\tilde{\Sigma})]^{1/N}$. When $N > T$ we take $\widehat{m}_{\text{EB}} = \widehat{m}$ 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 $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_N$ by:

$$T\widetilde{\lambda}_i \Bigg/ \left(T - N + 1 + 2\widetilde{\lambda}_i \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{1}{\widetilde{\lambda}_i - \widetilde{\lambda}_j} \right) \qquad i = 1, \ldots, N. \tag{12}$$

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.[5] Haff (1982) independently obtains a closely related estimator. In any given simulation, we call $\widehat{\Sigma}_{\mathrm{SH}}$ the better performing estimator of the two. The other one is not reported. The initials SH stand for Stein and Haff.[6]

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:

$$\frac{T}{T + N + 1 - 2i} \widetilde{\lambda}_i, \tag{13}$$

where sample eigenvalues $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_N$ are sorted in descending order. We call this estimator $\widehat{\Sigma}_{\mathrm{MX}}$. The initials MX stand for minimax.

We simulate normally distributed random variables. The true covariance matrix Σ can be taken diagonal without loss of generality. We draw its eigenvalues according to a log-normal distribution. We set their average equal to one without loss of generality. We let their dispersion r_1^2 vary around the central value 1/2. We let the ratio N/T vary around the central value 1/2. Finally, we let the product NT vary around the central value 800. We study the influences of r_1^2, N/T and NT separately. When one parameter moves, the other two remain fixed at their central values.

The asymptotic PRIAL of $\widehat{\widehat{\Sigma}}$ implied by Theorems 1 and 9 is $(N/T)/[(N/T) + r_1^2] \times 100$. The PRIAL increases in N/T and decreases in r_1^2. This is intuitive because N/T is the error on $\widetilde{\Sigma}$ and

[5] Intuitively, isotonic regression restores the ordering by assigning the same value to a subsequence of corrected eigenvalues that would violate it.

[6] When $N > T$ some of the terms $\widetilde{\lambda}_i - \widetilde{\lambda}_j$ in formula (12) result in a division by zero. We just ignore them. Nonetheless, when N is too large compared to T, the isotonic regression does not converge. In this case $\widehat{\Sigma}_{\mathrm{SH}}$ does not exist.

r_1^2 is the error on $\widehat{m}I$.

When all three parameters are fixed at their central values, we get the results in Table 1. "Risk" means the average mean squared error over 1,000 simulations. For the central values of the parameters, the asymptotic PRIAL of $\hat{\hat{\Sigma}}$ is 50%. Table 1 shows that asymptotic behavior is practically attained for $N = 20$ and $T = 40$. $\hat{\hat{\Sigma}}$ improves substantially over $\tilde{\Sigma}$ and $\hat{\Sigma}_{\text{EB}}$ and moderately over $\hat{\Sigma}_{\text{SH}}$ and $\hat{\Sigma}_{\text{MX}}$. This may be due to the fact that $\hat{\Sigma}_{\text{SH}}$ and $\hat{\Sigma}_{\text{MX}}$ were originally derived under another loss function than the mean squared error.

Estimator	$\tilde{\Sigma}$	$\hat{\hat{\Sigma}}$	$\hat{\Sigma}_{\text{EB}}$	$\hat{\Sigma}_{\text{SH}}$	$\hat{\Sigma}_{\text{MX}}$
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	0.00%	49.31%	4.69%	42.74%	40.02%

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

When we increase N/T from zero to infinity, the asymptotic PRIAL of $\hat{\hat{\Sigma}}$ increases from 0% to 100% with an "S" shape. Figure 4 confirms this.[7] $\hat{\hat{\Sigma}}$ always has lower mean squared error than $\tilde{\Sigma}$ and $\hat{\Sigma}_{\text{EB}}$. It usually has slightly lower mean squared error than $\hat{\Sigma}_{\text{SH}}$ and $\hat{\Sigma}_{\text{MX}}$. $\hat{\Sigma}_{\text{SH}}$ is not defined for high values of N/T. $\hat{\Sigma}_{\text{MX}}$ performs slightly better than $\hat{\hat{\Sigma}}$ for the highest values of N/T. This may be due to the fact that $\hat{\hat{\Sigma}}$ does not attain its asymptotic performance for values of T below 10.

When we increase r_1^2 from zero to infinity, the asymptotic PRIAL of $\hat{\hat{\Sigma}}$ decreases from 100% to 0% with a reverse "S" shape. Figure 5 confirms this. $\hat{\hat{\Sigma}}$ has lower mean squared error than $\tilde{\Sigma}$ always, and than $\hat{\Sigma}_{\text{EB}}$ almost always. $\hat{\hat{\Sigma}}$ always has lower mean squared error than $\hat{\Sigma}_{\text{SH}}$ and $\hat{\Sigma}_{\text{MX}}$. When r_1^2 gets too large, $\hat{\Sigma}_{\text{SH}}$ and $\hat{\Sigma}_{\text{MX}}$ perform worse than the sample covariance matrix. The reason is that $\widehat{m}I$ is very erroneous, and they shrink sample eigenvalues together too much. It is very reassuring that, in a case where its leading competitors perform much worse than $\tilde{\Sigma}$, $\hat{\hat{\Sigma}}$ performs at least as well as $\tilde{\Sigma}$.

When we increase NT from zero to infinity, we should see the PRIAL of $\hat{\hat{\Sigma}}$ converge to its asymptotic value of 1/2. Figure 6 confirms this. $\hat{\hat{\Sigma}}$ always has lower mean squared error than $\tilde{\Sigma}$ and $\hat{\Sigma}_{\text{EB}}$. It has moderately lower mean squared error than $\hat{\Sigma}_{\text{SH}}$ and $\hat{\Sigma}_{\text{MX}}$, except when T is below

[7] Corresponding tables of results are available from the author upon request. Standard errors on our estimators of the mean squared error have the same order of magnitude as in Table 1.

20. When T is below 20, $\hat{\hat{\Sigma}}$ performs slightly worse than $\hat{\Sigma}_{\rm SH}$ and moderately worse than $\hat{\Sigma}_{\rm MX}$, but still substantially better than $\tilde{\Sigma}$.

When the number of variables N is large, $\hat{\hat{\Sigma}}$ and $\tilde{\Sigma}$ take much less time to compute than $\hat{\Sigma}_{\rm EB}$, $\hat{\Sigma}_{\rm SH}$ and $\hat{\Sigma}_{\rm MX}$ because they do not need eigenvalues and determinants. Indeed the number and nature of operations needed to compute $\hat{\hat{\Sigma}}$ are of the same order as for $\tilde{\Sigma}$. It can be an enormous advantage when the covariance matrix is very large. The only seemingly slow step is the estimation of r_2^2, but it can be accelerated by writing:

$$\widehat{r_2^2} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\frac{1}{T}\left(X^{\wedge 2}\right)\left(X^{\wedge 2}\right)'\right]_{ij} - \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\left(\frac{1}{T}XX'\right)^{\wedge 2}\right]_{ij}$$

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

Simulations not reported here study departures from normality. These departures have little impact on the above results. In relative terms, $\tilde{\Sigma}$ and $\hat{\Sigma}_{\rm EB}$ appear to suffer the most; then $\hat{\Sigma}_{\rm SH}$ and $\hat{\Sigma}_{\rm MX}$; and $\hat{\hat{\Sigma}}$ appears to suffer the least.

We draw the following conclusions from these simulations. The asymptotic theory developed in Sections 2-3 approximates finite sample behavior well, as soon as T and N become of the order of 20 to 40. $\hat{\hat{\Sigma}}$ has lower mean squared error than the sample covariance matrix across the wide range of simulations studied. $\hat{\hat{\Sigma}}$ usually improves over existing finite sample decision theory estimators, in terms of mean squared error.[8] It sometimes performs substantially better than them. It never performs substantially worse than them.

This set of simulations indicates that the estimator $\hat{\hat{\Sigma}}$ from Section 3.2 can be used as an all-purpose estimator of the covariance matrix.

[8] We acknowledge that $\hat{\Sigma}_{\rm SH}$ and $\hat{\Sigma}_{\rm MX}$ were designed with another criterion than the mean squared error in mind. Our conclusions say nothing about performance under any other criterion. Nevertheless, the mean squared error is an important criterion. Also, there is some similarity between criteria, as suggested by the fact that $\hat{\Sigma}_{\rm SH}$ and $\hat{\Sigma}_{\rm MX}$ do perform well in terms of mean squared error.

4.2 Historical Data

This section takes the covariance matrix estimator $\hat{\hat{\Sigma}}$ to the data. The objective is to estimate how well it would have performed over the past, had it been used for portfolio selection.

Monthly stock returns in excess of the riskfree rate and capitalizations from July 1926 to June 1993 are drawn from the Center for Research in Security Prices (CRSP) database. Let y denote any year between 1936 and 1992. Stock returns from July of year $y - 10$ to June of year y are used to estimate the covariance matrix of stock returns. Stocks with missing observations are excluded. We consider only common stocks traded on the New York Stock Exchange (NYSE) or the American Stock Exchange (AMEX).[9] We require stocks to have a valid market capitalization in June of year y.

From these data, we extract two factors that past research has associated with stock returns. The first one is the beta with respect to a CRSP value-weighted index including dividends.[10] The second factor is the logarithm in base 10 of the market capitalization in dollars of a given stock, minus the average logarithm of market capitalization across all stocks in the dataset in year y. We call this factor: "size", for brevity. The average is subtracted because a stock with the same 50 million dollars capitalization would have been relatively large in 1936, and relatively small in 1992. Thus, a stock with "size" one (respectively minus one) is ten times larger (respectively smaller) than the average stock in the market.

We compare different covariance matrices, either of the structured type ($\bar{\Sigma}$) or the asymptotic shrinkage type ($\hat{\hat{\Sigma}}$). We do not include the other estimators because they are too costly to compute or not defined at all when N is much larger than T, which is the case here.[11] $\bar{\Sigma}$ can be either $\widehat{m}I$ as in Section 3.2, or any of the four structured estimators in Appendix B. Each of these five structured estimators gives a shrinkage estimator. Therefore there are ten estimators in total.

We impose different sets of portfolio constraints. We always make weights sum up to one. In addition, we impose zero, one or two constraints chosen among the following two: the weighted average of betas has a required value; the weighted average of sizes has a required value. Therefore

[9] AMEX stocks do not appear in the CRSP database before July 1963. We do not include them before $y = 1973$.

[10] Before July 1963, the NYSE index; afterwards, the NYSE and AMEX index.

[11] The number of stocks N grows from 340 in 1936 to 1105 in 1992. The number of time periods T is 120 (ten years of monthly data).

there are four possible sets of constraints.

Based on these data, we buy at the end of June of year y forty different kinds of minimum variance portfolios corresponding to the ten covariance matrices and the four sets of constraints. We hold them until the end of June of year $y + 1$, at which time they are rebalanced in a similar fashion, incorporating fresh data. This scheme yields a time series of monthly returns for each of the forty kinds of portfolios from July 1936 to June 1993. Since each rebalancing is based only on information that is available at the time, we are simulating realistic investment strategies. Tests based on strategies such as these ones, i.e. that do not require hindsight, are called predictive. They are easier to interpret than non-predictive tests. In addition, since we measure true buy-and-hold returns and rebalance portfolios only once a year, transactions costs are quite limited. We ignore them.

The most urgent questions concern shrinkage weight $(r_2^2 - \varphi)/d^2$: Is it between zero and one? Is it relatively stable over time? Does it make intuitive sense? Qualitatively, the answers to these three questions appears to be yes in Figure 7. Weights are between 0.07 and 0.93 for every structured estimator and every year. Each structured estimator's weights remain within the same range of width 0.3 (approximately) throughout the 67 years. The ordering between weights remains the same over time, and makes intuitive sense. Diagonal structured estimators are given the least weight, probably because the true covariance matrix is far from being diagonal. The structured estimators that have the most free parameters are given the highest weights, probably because they are the least biased. Qualitative evidence from Figure 7 is very reassuring about the estimators of shrinkage weights, which are among our main contributions.

The most important question about the empirical properties of our method is: Does shrinkage help minimize variance? Table 2 provides evidence that it does. The table shows the ex-post standard deviation of the ex-ante unconstrained minimum variance portfolio. For all five structured estimators, shrinkage yields portfolios with significantly lower variance. In some cases, variance diminishes a lot.

These results might be criticized as relying only on the unconstrained minimum variance portfolio. Therefore, for each structured estimator, we consider three portfolios: zero beta and size

-1; unit beta and size -1; zero beta and unit size.[12] If an investor believed that returns are driven by beta and/or size, she would select some combination of these three portfolios. Then we give the benefit of hindsight to structured estimators, but only to them. That is, we choose the combination of these three portfolios with the lowest variance *based on ex-post variances and covariances*. We compare it to the ex-ante minimum variance portfolio from the corresponding shrinkage estimator. This is unfair because hindsight is such a strong advantage. It biases our results towards not finding that shrinkage helps reduce variance.

Results are in Table 3. Again, all five shrinkage estimators (without hindsight) yield portfolios with lower variance than their corresponding structured estimators do (even with hindsight). In this sense, it can be said that our method yields portfolios with lower variance than could possibly be attained before. Table 3 demonstrates empirically that our estimator $\hat{\hat{\Sigma}}$ achieves its goal: it helps portfolio selection minimize variance.

Portfolios with lower variance than was previously possible open a new investment opportunity. From an economic perspective, it is interesting to know whether this new opportunity is attractive: Does it let investors improve their risk-return tradeoff? The risk-return tradeoff can be summarized by the Sharpe ratio: mean divided by standard deviation of portfolio returns.[13]

Figure 8 plots the ex-post means and standard deviations of the ex-ante minimum variance portfolios constrained to have a specified beta between zero and one, and size zero. On each graph, portfolios obtained from a structured estimator are plotted as a dashed line, together with portfolios from the corresponding shrinkage estimator as a solid line. As seen above, the solid line ventures further into low-risk territory than the dashed line. However, the risk-return tradeoff does not seem to improve much. The dotted line, whose slope is the maximum Sharpe ratio of all the portfolios on the figure, is practically tangent to both the solid line of shrinkage estimator portfolios and the dashed line of structured estimator portfolios.

This is especially true when $\bar{\Sigma}$ is given by the single index model, which is the structured estimator closest to what actual investors would use. For the other $\bar{\Sigma}$s, our interpretation is that combining a structured estimator with the sample covariance matrix goes a long way towards fixing its intrinsic flaws, if any exist.

[12] Remember that size one (minus one) means ten times larger (smaller) market capitalization than market average.

[13] Returns are in excess of the riskfree rate.

Overall, the message is that low risk portfolios are penalized by low returns. They do not offer more attractive investment opportunities. While this may sound a little disappointing to a practitioner, it is on the contrary very satisfying for an economist. In equilibrium, there should be no easy and permanent way to attain an abnormally favorable risk-return tradeoff. It is rather remarkable that agents priced fairly the low-risk portfolios identified in this paper... even long before they were identified! This can be interpreted as strong support for equilibrium theory of risk-return tradeoff.

Since a particular version of this theory, the Capital Asset Pricing Model (CAPM), has recently been challenged on empirical grounds, it is natural to extract from shrinkage covariance matrix estimators quantitative evidence on this subject beyond Figure 8.

4.3 Testing an Implication of the CAPM

The CAPM implies, among other things, a positive relationship between returns and betas. A familiar method to test this is to run a cross-sectional regression of returns on betas: the CAPM predicts a positive slope. As Fama (1970) clearly explains, this is equivalent to forming minimum variance portfolios with betas of one and zero respectively, and then testing whether they have different mean returns. This brings back CAPM tests to portfolio selection, where shrinkage covariance matrix estimators can be used.

Most existing tests run Ordinary Least Squares (OLS) regressions. This corresponds to using the structured estimator $\overline{\Sigma} = \widehat{m}I$ for portfolio selection. No doubt it can be replaced by an improved estimator of the covariance matrix. This corresponds to running Generalized Least Squares (GLS) regressions. Amihud, Christensen and Mendelson (1994) are among the few who run GLS. The problem is that they allow themselves to "peek into the future" to estimate the covariance matrix. Their test is not predictive. Its interpretation is not straightforward, because real-life investors cannot peek into the future. Furthermore, the ex-post returns that they report are not truly ex-post because they come from a period that has already been used to estimate the covariance matrix. This feature can bias standard errors towards zero, t-statistics away from zero, and tests of the CAPM towards finding a significantly positive slope. We avoid these problems by running a predictive test.

Another difficulty is estimating betas. Since beta estimates contain error, the largest ones are biased upwards, and the smallest ones downwards, by now a familiar phenomenon. Some authors aggregate stocks into portfolios, on the assumption that betas can be estimated more accurately for portfolios. Typically, portfolios are formed by ranking stocks on the basis of their betas estimated over a given period, then portfolio betas are estimated over a later period. This ensures that betas vary across portfolios, but prevents portfolio beta estimates from being biased. What this procedure actually does is shrink beta estimates together.

Since shrinkage is the general answer to such problems, why not apply the technique of Section 3? As it turns out, there is a direct correspondence between shrinking sample eigenvalues when T and N both go to infinity, and shrinking beta estimates (or sample means) when T is fixed and N goes to infinity. Thus, the asymptotic linear shrinkage developed in Sections 3.1-3.2 can be applied to betas too. However, linear shrinkage has no impact on t-statistics of regression slopes: it only changes the intercept. In other words, if the bias of betas is nearly linear, then there is little reason to fix it. For this reason, we do not elaborate on this point here, and work with unadjusted betas. This more naive approach is less arbitrary than forming portfolios, and — if anything — makes it harder to find a significant relationship between returns and betas.

Previous OLS regressions of returns on betas found a positive slope, but with some serious limitations. First, it is not always statistically significant. Second, Tinic and West (1984) show that the return-beta relationship weakens substantially if the month of January is excluded from the period. Also, Lakonishok and Shapiro (1986) find that it disappears if size is included in the regression. Finally, Fama and French (1992) report that it flattens out over the period 1963-1992.

Using the same database as these authors, we reproduce their OLS results in Table 4. The t-statistic for significance of the slope of returns on betas is 1.03 over the full period 1936-1992. It goes down to -0.33 if January is excluded, to 0.17 if size is included, and to 0.60 over 1963-1992. Actual results may differ somewhat from previously published ones, but the conclusions are identical.

Now, we change only one step: instead of using the structured estimator $\bar{\Sigma} = \widehat{m} I$ for portfolio selection, we use a shrinkage estimator. This corresponds to upgrading from OLS to GLS. In Table 5, we report the results obtained with the shrinkage estimator corresponding to the single index model, since this is the best-known structured estimator among the ones in Appendix B.

The t-statistic for significance of the slope of returns on betas is now 1.91 over the full period 1936-1992. It is statistically significant at the 5% level against a one-sided alternative. It only goes down to 1.62 if January is excluded, to 1.44 if size is included, and to 1.16 over 1963-1992. This relationship is much more robust than under OLS.[14]

The change comes from two sources: standard deviations go down, because GLS is more efficient than OLS, and slope estimates go up. Kandel and Stambaugh (1994) explain theoretically why this should be anticipated. They show that OLS slope estimates can be more sensitive than GLS to misspecification of the market proxy used to estimate betas.

In conclusion, the first predictive GLS cross-sectional regression of stock returns on betas, conducted thanks to the asymptotic linear shrinkage estimator of the covariance matrix developed in Section 3, finds a more significant and robust positive relationship between returns and betas than similar OLS regressions do. The relationship is not as strong as theory suggests, but this is hardly surprising given the error of beta estimates. Predictive GLS regressions support the existence of a positive linear relationship between returns and betas.

5 Conclusion

Directions for future research include using the spectral theory of large-dimensional random matrices to test for the number of factors in the APT; translating asymptotic shrinkage techniques to beta estimation; searching for the best frequency at which to sample stock returns for covariance matrix estimation; accounting for some type of Autoregressive Conditional Heteroskedasticity (ARCH) effects; bringing improved covariance matrix estimators to other areas of empirical stock market finance such as event studies.

In this paper, we demonstrate the importance of a seldom-used framework for covariance matrix estimation: letting the number of variables and the number of observations go to infinity together. This framework is particularly well-suited for stock returns data, because the number of stocks

[14] The interested reader can find results for other asymptotic shrinkage estimators in Table 6. All slope estimates are positive. The results that we choose to comment are neither the weakest nor the strongest, and are close to another structured estimator's results. We believe that they are the most credible.

traded in the stock market is at least of the same order of magnitude as the number of time periods. The covariance matrix of stock returns is important because it is a necessary input into portfolio selection, a central method in stock market finance.

We show that, in this framework, the sample covariance matrix is not well-behaved, especially through its eigenvalues. This work has potential implications for tests of the number of factors in the APT based on sample covariance matrix eigenvalues. We also show that it is easy to improve over the sample covariance matrix by shrinking its eigenvalues together in an asymptotically optimal way. In particular, this yields the first rotation-invariant estimator of the inverse of the covariance matrix to retain some theoretical motivation when variables outnumber observations. Generalizations provide attractive asymptotic extensions to familiar finite sample Bayesian and decision theory methods.

Monte-Carlo simulations reveal that peak asymptotic performance is attained as soon as the number of observations and the number of variables become of order 20 to 40. The asymptotic shrinkage estimator has lower mean squared error than the sample covariance matrix in all situations simulated. It compares favorably overall in terms of mean squared error with existing finite sample estimators. The asymptotic shrinkage estimator has the potential to replace the sample covariance matrix as an all-purpose estimator.

More importantly for Finance, this asymptotic shrinkage technology helps portfolio selection minimize variance, as tests on historical data show. It opens new investment opportunities: equity portfolios with lower risk than was previously possible. These opportunities, however, are only slightly more attractive than existing ones because lower risk is penalized by lower return. In a related investigation of the risk-return tradeoff, the improved covariance matrix estimator is used to perform the first predictive GLS cross-sectional regression of returns on betas. This test concludes that the positive relationship between returns and betas predicted by the CAPM is statistically significant and robust, in stark contrast with tests based on less efficient OLS regressions.

References

Adamek, P. (1994). Approximate factor structure: a test for number of factors. Technical report, MIT Sloan School of Management.

Amihud, Y., Christensen, B. J., and Mendelson, H. (1992). Further evidence on the risk-return relationship. Technical report, New York University.

Bawa, V. S., Brown, S. J., and Klein, R. W. (1979). *Estimation Risk and Optimal Portfolio Choice*. Bell Laboratory Series. North Holland, New York. Studies in Bayesian Econometrics.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Polititcal Economy*, 81:637–654.

Brown, S. J. (1989). The number of factors in security returns. *Journal of Finance*, 44(5):1247–1262.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage and mean variance analysis on large asset markets. *Econometrica*, 51:1281–1304.

Connor, G. and Korajczyk, R. A. (1992). The arbitrage pricing theory and multifactor models of asset returns. In *Finance Handbook*. R. Jarrow, V. Maksimovic and W. Ziemba, eds.

Dey, D. K. and Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *Annals of Statistics*, 13(4):1581–1591.

Fama, E. F. (1970). *Foundations of Finance*. Basic Books, New York.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *Journal of Finance*.

Frost, P. A. and Savarino, J. E. (1986). An empirical Bayes approach to portfolio selection. *Journal of Financial and Quantitative Analysis*, 21(3):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.

Jobson, J. D. and Korkie, B. (1980). Estimation for Markowitz efficient portfolios. *Journal of the American Statistical Association*, 75(371):544–554. Applications Section.

Kandel, S. and Stambaugh, R. F. (1994). Portfolio inefficiency and the cross-section of expected returns. Technical report, Wharton School.

Lakonishok, J. and Shapiro, A. C. (1986). Systematic risk, total risk and size as determinants of stock market returns. *Journal of Banking and Finance*, 10:115–132.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1):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(4):457–483.

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

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

Roll, R. (1977). A critique of the asset pricing theory's test; part I: on past and potential testability of the theory. *Journal of Financial Economics*, 4:129–176.

Sharpe, W. F. (1963). A simplified model for portfolio analysis. *Management Science*.

Sheena, Y. and Takemura, A. (1992). Inadmissibility of non-order-preserving orthogonally invariant estimators of the covariance matrix in the case of Stein's loss. *Journal of Multivariate Analysis*, 41:117–131.

Silverstein, J. W. (1994). Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*. Submitted.

Silverstein, J. W. and Choi, S.-I. (1994). Analysis of the limiting spectral distribution of large dimensional random matrices. *SIAM Journal on Mathematical Analysis*. Submitted.

Silverstein, J. W. and Combettes, P. L. (1992). Signal detection via spectral theory of large dimensional random matrices. *IEEE Transactions on Signal Processing*, 40:2100–2105.

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.

Tinic, S. M. and West, R. R. (1984). Risk and return: January vs. the rest of the yer. *Journal of Financial Economics*, 13:561–574.

Wachter, K. W. (1976). Probability plotting points for principal components. In *Proceedings of the Ninth Interface Symposium on Computer Science and Statistics*, pages 299–308. Hoaglin and Welsch, eds.

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

Appendices

A Spectral Theory of Large Random Matrices

This appendix gives details about the spectral theory of large-dimensional random matrices. To our knowledge, it is the first time that this theory has been mentioned in the finance literature. It bears directly on tests for the number of factors in the Arbitrage Pricing Theory (APT) based on the largest eigenvalues of the sample covariance matrix. Since this is somewhat outside the scope of the paper, we do not provide proofs.

A.1 Mathematical Tools

A *cumulative distribution function (c.d.f.)* is a nondecreasing right-continuous function defined on the real line whose limit is zero at $-\infty$ and one at $+\infty$.

Definition 3 *Let S be a symmetric matrix. Its* spectral c.d.f. *is the function defined by $F^S(x) =$ proportion of eigenvalues of $S \leq x$. If the matrix S is random, so is the value of its spectral c.d.f. $F^S(x)$.*

The spectral c.d.f. is in one-to-one correspondence with the system of eigenvalues. It is a convenient way to summarize the behavior of eigenvalues without invoking the joint density. The joint density would become very complicated as the number of eigenvalues grows.

Definition 4 *The linear operator L transforms the c.d.f. F with support $[0, +\infty)$ into the nondecreasing function: $LF(x) = \int_{-\infty}^{x} t \; dF(t)$.*

The inversion formula is: $F(x) = L^{-1}[LF](x) = LF(1) + LF(x)/x + \int_1^x LF(t)\; dt/t^2$ for $x > 0$, $F(0) = L^{-1}[LF](0) = \lim_{x \searrow 0} F(x)$, and $F(x) = L^{-1}[LF](x) = 0$ for $x < 0$. This linear operator is only introduced to simplify equations. Its presence can often be ignored when thinking of the problem intuitively.

Definition 5 *If F is a nondecreasing function verifying $\int_{-\infty}^{+\infty} dF(t)/(1 + |t|) < \infty$, then its Stieltjes transform s_F is defined by:*

$$s_F(z) = \int_{-\infty}^{+\infty} \frac{dF(t)}{t - z} \tag{14}$$

for z on the strict upper half $\mathbf{C}^+$ of the complex plane. Where possible, extend s_F by continuity to real x: $s_F(x) = \lim_{z \in \mathbf{C}^+ \to x} s_F(z)$.

The inversion formula is $F(t) = \lim_{\varepsilon \searrow 0}(1/\pi)\mathrm{Im}[\int_{-\infty}^{t} s_F(x + i\varepsilon)dx]$ at all points of continuity of F, where Im denotes the imaginary part of a complex number. If F is regular enough at x, e.g. twice differentiable in a neighborhood of x, then $s_F(x)$ exists and is equal to $\lim_{\varepsilon \searrow 0} \int_{|t-x| \geq \varepsilon} dF(t)/(t - x) + i\pi F'(x)$, where prime denotes the derivative (no confusion with the transposition is possible). The real and imaginary parts of s_F satisfy the Laplace equation over $\mathbf{C}^+$:

$$\frac{\partial^2 \mathrm{Re}\,[s_F(x+iy)]}{\partial x^2} + \frac{\partial^2 \mathrm{Re}\,[s_F(x+iy)]}{\partial y^2} = 0 \qquad (15)$$

$$\frac{\partial^2 \mathrm{Im}\,[s_F(x+iy)]}{\partial x^2} + \frac{\partial^2 \mathrm{Im}\,[s_F(x+iy)]}{\partial y^2} = 0' \qquad (16)$$

where Re denotes the real part. For fixed $y > 0$, the function $x \mapsto (1/\pi)\mathrm{Im}[s_F(x + iy)]$ is the convolution of the density $F'(x)$ with the Cauchy kernel $x \mapsto (y/\pi)/(x^2 + y^2)$.

Definition 6 *The c.d.f.'s $(F_n)_{n \geq 1}$ converge* in distribution *to F if $F_n(x) \to F(x)$ at all points of continuity of F.*

With these mathematical tools, we can expose the results of the spectral theory of large-dimensional random matrices that are relevant to some tests of the APT.

A.2 Asymptotic Results

Recall that $Y = U'X$ is an $N \times T$ matrix of T iid observations on a system of N uncorrelated random variables that spans the same space as the original system. Let $(y_{11}, \ldots, y_{N1})'$ denote the first column of the matrix Y. $y_{11}, \ldots, y_{N1}$ are uncorrelated with variances $\lambda_1, \ldots, \lambda_N$ respectively. We need to strengthen Assumption 3.

Assumption 4 $y_{11}/\sqrt{\lambda_1}, \ldots, y_{N1}/\sqrt{\lambda_N}$ *are iid.*

We maintain Assumption 4 throughout the remainder of this appendix. The following theorem was first proven by Marčenko and Pastur (1967). It was later generalized by a number of authors. The latest and most general version is by Silverstein (1994).

Proposition 2 *Assume that the ratio N/T converges to a finite positive limit c called the concentration. Assume that the spectral c.d.f. F^{Σ} of the true covariance matrix Σ converges in distribution to a c.d.f. H. Then the spectral c.d.f. $F^{\tilde{\Sigma}}$ of the sample covariance matrix $\tilde{\Sigma}$ converges almost surely in distribution to a nonrandom c.d.f. G.*

The fact that the sample spectral c.d.f. $F^{\tilde{\Sigma}}$ is asymptotically nonrandom is quite remarkable. Even though $\tilde{\Sigma}$ randomly moves around its expectation Σ, its eigenvalues remain the same (in some sense). The error on sample eigenvalues is all bias and no variance. Bias comes from the fact that G is different from H.

Basic qualitative properties are established by Silverstein and Choi (1994).

Proposition 3 *G is uniquely determined by H and c. H is uniquely determined by G and c. G converges in distribution to H as c goes to zero. G has a continuous derivative, except possibly at zero. The masses $G\{0\}$ and $H\{0\}$ that G and H respectively place at zero are related by: $G\{0\} = \max(H\{0\}, 1 - 1/c)$.*

The particular shape of the distribution of the random variables X does not matter, except through the covariance matrix Σ. Under standard asymptotics, c is zero: sample and true eigenvalues coincide. Even though the distribution of true eigenvalues need not be smooth (e.g. for $\Sigma = I$ it is discontinuous at one), the distribution of sample eigenvalues must be, except possibly at zero. Intuitively, the error of sample covariance matrix eigenvectors smoothes out sample eigenvalues. If H places some mass at zero, then G places at least the same mass at zero. Intuitively, true eigenvalues at zero do not get smoothed out because the observed variance of their corresponding eigenvectors is exactly zero in every sample. If $c > 1$, then $\tilde{\Sigma}$ is rank-deficient, therefore it can have more eigenvalues equal to zero than Σ.

The equation linking H to G is due to Marčenko and Pastur (1967):

$$\forall z \in \mathbf{C}^{+} \qquad s_{LH}\left(\frac{z}{1 - c\, s_{LG}(z)}\right) = s_{LG}(z). \tag{17}$$

It is our contribution to introduce the linear operator L. It simplifies the equation. Equation (17) clearly displays how nonzero concentrations drive G and H apart. An additional advantage is that s_{LG} and s_{LH} are better behaved near zero than the Stieltjes transforms s_G and s_H used previously.

Yin (1986) derives another equation with H and G.

Proposition 4 *Assume that all the moments $h_1, h_2, \ldots$ of H exist and satisfy Carleman's condition $\sum_{k=1}^{\infty} h_{2k}^{-1/2k} = +\infty$. Then all the moments $g_1, g_2, \ldots$ of G exist and satisfy Carleman's condition. They are given by:*

$$\forall k = 1, 2, \ldots \qquad g_k = \sum_{w=1}^{k} c^{k-w} \sum \frac{k!}{n_1! n_2! \cdots n_w!} h_1^{n_1} h_2^{n_2} \cdots h_k^{n_k}, \tag{18}$$

where the inner sum extends over all w-tuples of nonnegative integers $(n_1, n_2, \ldots, n_w)$ such that $\sum_{i=1}^{w} n_i = k - w + 1$ and $\sum_{i=1}^{w} i n_i = k$.

Carleman's condition ensures that a distribution is uniquely determined by its moments. It is verified by most familiar distributions whose moments exist. For the first moment, Equation (18) yields $g_1 = h_1$, a result that we have already seen in Theorem 2. For the second moment, $g_2 = h_2 + c h_1^2$, a result that we have already seen in Footnote 2. The second and higher moments of the sample spectral c.d.f. are larger than those of the true spectral c.d.f. The difference increases in the concentration. This means that sample eigenvalues are more dispersed than true ones. Excess dispersion increases in the concentration.

A.3 From True to Sample Eigenvalues

For $\Sigma = I$, all eigenvalues are equal to one. The true spectral c.d.f. is $H(x) = \mathbf{I}_{[0,+\infty)}(x)$, where $\mathbf{I}$ denotes the indicator function of a set. Marčenko and Pastur solve Equation (17) explicitly in this important particular case. Define $a_c = (1 - \sqrt{c})^2$ and $b_c = (1 + \sqrt{c})^2$. Let $\psi_c(t) = \sqrt{(t - a_c)(b_c - t)}/(2\pi c t)$ for $a_c \leq x \leq b_c$ and $\psi_c(t) = 0$ otherwise. Then $G(x) = \int_{-\infty}^{x} \psi_c(t)\, dt$ if $0 < c \leq 1$, and $G(x) = (1 - 1/c)\mathbf{I}_{[0,+\infty)}(x) + \int_{-\infty}^{x} \psi_c(t)\, dt$ if $c > 1$. This is the formula that yields Figure 1.

In the general case, remember that the sample spectral c.d.f. G has a continuous derivative G', except possibly at zero. Silverstein and Choi (1994) show that for every $x \neq 0$ for which $G'(x) > 0$, $\pi c G'(x)$ is the imaginary part of the unique $z \in \mathbf{C}^+$ satisfying:

$$x = -\frac{1}{z} + c \int_{-\infty}^{+\infty} \frac{t}{1 + tz} dH(t). \tag{19}$$

When H is discrete and its support has a finite number of points n_H, z is the root of a polynomial of degree at most $n_H + 1$. For $n_H \leq 3$, the polynomial equation can be solved in closed form, which yields an explicit formula for $G'(x)$. A Fortran routine by Wachter (1976) implements it for $n_H = 2$. Otherwise, it is straightforward to solve Equation (19) numerically. In particular, it is a well-posed problem.

A.4 From Sample to True Eigenvalues

The APT makes assumptions about the eigenvalues of the true covariance matrix Σ of the returns on all stocks traded in the stock market (Chamberlain and Rothschild, 1983). Some authors have tried to test these assumptions by using the eigenvalues of the sample covariance matrix $\tilde{\Sigma}$. As Brown (1989) points out and our analysis confirms, sample eigenvalues do not estimate true eigenvalues well when N is of the same order of magnitude as T, which is the usual case. In particular, the largest sample eigenvalues are upward biased estimators of the largest true eigenvalues. How can we use the spectral theory of large-dimensional matrices for such tests?

Theorem 3 states that the true spectral c.d.f. H is uniquely determined by the sample spectral c.d.f. G and the concentration c. It is easy to obtain a smooth nonparametric estimator $\widehat{G}$ of G. Can we plug it, along with $c = N/T$, into Equation (17) in order to back up an estimator $\widehat{H}$ of H?

$\widehat{G}$ can be used to estimate the complex function s_{LG} by $s_{L\widehat{G}}$ over $\mathbf{C}^+$. Equation (17) then yields an estimator $s_{L\widehat{H}}$ of the complex function s_{LH}, but not over all of $\mathbf{C}^+$: only over the domain $\widehat{D} = \{z/[1 - c\, s_{L\widehat{G}}(z)], z \in \mathbf{C}^+\}$. This domain is included in $\mathbf{C}^+$, but excludes a portion of $\mathbf{C}^+$ near the real axis. A typical domain $\widehat{D}$ is shown in Figure 9.

From the Stieltjes transform $s_{L\widehat{H}}$, we need to back up an estimate of the distribution of true eigenvalues H. Roughly, the Stieltjes inversion formula is: $\lim_{\varepsilon \searrow 0} \mathrm{Im}[s_{L\widehat{H}}(x + i\varepsilon)] = \pi x H'(x)$, where $H'(x)$ is the density of true eigenvalues. Therefore we can estimate $H'(x)$ if we know $s_{L\widehat{H}}(x + i\varepsilon)$ for small $\varepsilon > 0$. What we need is to extend our estimator $s_{L\widehat{H}}$ from the domain $\widehat{D}$ towards the real axis.

The imaginary part of $s_{L\widehat{H}}$ satisfies the Laplace equation (15)-(16) over $\mathbf{C}^+$, and in particular between $\widehat{D}$ and the real line. Our goal is to solve this partial differential equation over $\mathbf{C}^+ - \widehat{D}$. The boundary of $\mathbf{C}^+ - \widehat{D}$ is divided into two pieces: the frontier with $\widehat{D}$, where we know the value of $s_{L\widehat{H}}$, and the real axis, where we want to know it. Since we do not have any information about the function on a piece of the boundary of the domain, this p.d.e. has a "free boundary."

Solving the Laplace equation with a free boundary is an ill-posed problem.

Even infinitesimal errors on the value of $s_{L\widehat{H}}$ over the domain $\widehat{D}$ are amplified into large errors near the real axis. To put it in another way, there are some very different values of $s_{L\widehat{H}}$ near the real axis that imply almost the same values of $s_{L\widehat{H}}$ on the domain $\widehat{D}$. Available data do not provide much guidance in choosing between them. If $s_{L\widehat{H}}$ oscillated wildly over the real line, the Laplace equation would smooth it out so that we would not notice it over $\widehat{D}$.

In practice, for high values of c, sample eigenvalues look a lot like in Figure 1, regardless of how true eigenvalues are distributed. It is possible to back up the average and the dispersion of true eigenvalues, but not much more than that when N is of the same order of magnitude as T.

The degree of ill-posedness is determined by how far from the real line the domain $\widehat{D}$ is. It increases in the concentration c. If c is negligible, then the domain $\widehat{D}$ is so close to the real line that ill-posedness is negligible. In practice, c is not negligible, which is why we want to improve over sample eigenvalues in the first place.

There is, however, one reason to hope that this approach can potentially be yield APT tests: the degree of ill-posedness is not uniform. It is roughly proportional to the density of sample eigenvalues. In Figure 9, there are a lot of small eigenvalues and a few large ones. This is realistic for the stock market. We can see that the domain $\widehat{D}$ gets closer to the real axis around large eigenvalues (large values of x). It may make it easier to estimate the density of true eigenvalues $h(x)$ when x is large. Silverstein and Combettes (1992) make a similar argument in the context of signal detection. It suggests that the problem of estimating large, isolated eigenvalues may not be ill-posed, even if the concentration is not negligible. This suggestion will be explored in future research.

In the end, it may even turn out that large, isolated true eigenvalues are actually well estimated by large, isolated sample eigenvalues. This kind of reassurance, however, cannot come from standard asymptotics. Therefore it is essential to recognize in APT tests that the number of variables N is not negligible with respect to the number of observations T. The spectral theory of large-dimensional random matrices offers one possible way to do this. Another way is proposed by Adamek (1994). He obtains very interesting results by assuming that the number of variables N goes to infinity while the number of observations T remains fixed.

B Additional Formulas

This appendix discusses the optimal combination of a structured estimator $\overline{\Sigma} = [\overline{\sigma}_{ij}]_{i,j=1,\ldots,N}$ and the sample covariance matrix $\tilde{\Sigma} = [\tilde{\sigma}_{ij}]_{i,j=1,\ldots,N}$. Section 3.3 shows the importance of $\varphi_{ij} = \mathrm{Cov}[\overline{\sigma}_{ij}, \tilde{\sigma}_{ij}]$ for $i, j = 1, \ldots, N$, and $\varphi = (1/N)\sum_{i=1}^{N}\sum_{j=1}^{N}\varphi_{ij}$. This section shows how to estimate these parameters for various choices of the structured estimators $\overline{\Sigma}$.

B.1 All Variances, Respectively Covariances, Are Equal

Frost and Savarino (1986) propose a structured estimator of the covariance matrix with two free parameters: one on the diagonal, the other one off the diagonal. They obtain $\overline{\Sigma} = \widehat{m}I + \widehat{q}(\mathbf{1}\mathbf{1}' - I)$, where $\widehat{q} = \frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{i-1}\tilde{\sigma}_{ij}$ is the average of the off-diagonal elements of the sample covariance matrix, and $\mathbf{1}$ is a conformable column vector of ones. On the diagonal, φ_{ii} is at most of order $1/T$ for each $i = 1, \ldots, N$. Off the diagonal, $\mathrm{Var}[\tilde{\sigma}_{ij}]$ is of order $1/T$ and $\mathrm{Var}[\overline{\sigma}_{ij}]$ is at most of order $1/(NT)$, therefore φ_{ij} is at most of order $1/(\sqrt{N}T)$ for $i, j = 1, \ldots, N$, $i \neq j$. This makes φ at most of order $\sqrt{N}/T$: it vanishes asymptotically. In conclusion, for this choice of prior, we recommend $\widehat{\varphi} = 0$.

B.2 Diagonal Matrix

If we impose that $\overline{\Sigma}$ is diagonal, then $\varphi_{ij} = 0$ for $i, j = 1, \ldots, N$, $i \neq j$. Since φ_{ii} is of order $1/T$ for $i = 1, \ldots, N$, this makes φ at most of order $1/T$. For this choice of prior too, we recommend $\widehat{\varphi} = 0$.

B.3 All Correlation Coefficients Are Equal

We can impose that all pairs of stock returns have the same correlation coefficient. On the diagonal, $\overline{\sigma}_{ii} = \tilde{\sigma}_{ii}$, therefore $\varphi_{ii} = \mathrm{Cov}[\overline{\sigma}_{ii}, \tilde{\sigma}_{ii}] = \mathrm{Var}[\tilde{\sigma}_{ii}]$, which can be estimated by $\widehat{\varphi}_{ii} = (1/T^2)\sum_{t=1}^{T}(x_{it}^2 - \tilde{\sigma}_{ii})^2$ for $i = 1, \ldots, N$, as in Theorem 7.

Let $\widehat{\rho} = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{i-1} \widetilde{\sigma}_{ij} / \sqrt{\widetilde{\sigma}_{ii} \widetilde{\sigma}_{jj}}$ denote the average of the correlation coefficients in the sample covariance matrix. Off the diagonal, $\overline{\sigma}_{ij} = \widehat{\rho} \sqrt{\widetilde{\sigma}_{ii} \widetilde{\sigma}_{jj}}$. $\mathrm{Cov}[\widehat{\rho}, \widetilde{\sigma}_{ij}]$ is of order at most $1/(NT)$, therefore it can be neglected. $\mathrm{Cov}[\widetilde{\sigma}_{ii}, \widetilde{\sigma}_{ij}]$ can be estimated by $\widehat{v}_{ii,ij} = (1/T^2) \sum_{t=1}^{T} (x_{it}^2 - \widetilde{\sigma}_{ii})(x_{it} x_{jt} - \widetilde{\sigma}_{ij})$. Using the delta method, $\varphi_{ij} = \mathrm{Cov}[\overline{\sigma}_{ij}, \widetilde{\sigma}_{ij}]$ can be estimated by $\widehat{\varphi}_{ij} = \widehat{\rho} \left(\widehat{v}_{ii,ij} \sqrt{\widetilde{\sigma}_{jj} / \widetilde{\sigma}_{ii}} + \widehat{v}_{jj,ij} \sqrt{\widetilde{\sigma}_{ii} / \widetilde{\sigma}_{jj}} \right) / 2$, for $i, j = 1, \ldots, N$, $i \neq j$. These formulas yield the estimator $\widehat{\varphi} = (1/N) \sum_{i=1}^{N} \sum_{j=1}^{N} \widehat{\varphi}_{ij}$ that we recommend in this case.

B.4 Single Index Model

The matrix of observations is $X = [x_{it}]_{\substack{i=1,\ldots,N \\ t=1,\ldots,T}}$. On the diagonal, $\overline{\sigma}_{ii} = \widetilde{\sigma}_{ii}$, therefore $\varphi_{ii} = \mathrm{Cov}[\overline{\sigma}_{ii}, \widetilde{\sigma}_{ii}] = \mathrm{Var}[\widetilde{\sigma}_{ii}]$, which can be estimated by $\widehat{\varphi}_{ii} = (1/T^2) \sum_{t=1}^{T} (x_{it}^2 - \widetilde{\sigma}_{ii})^2$ for $i = 1, \ldots, N$, as in Theorem 7.

Let $[x_{Mt}]_{t=1,\ldots,T}$ denote returns on the market index. Let $\widetilde{\sigma}_{MM} = (1/T) \sum_{t=1}^{T} x_{Mt}^2$, and for $i = 1, \ldots, N$, let $\widetilde{\sigma}_{iM} = (1/T) \sum_{t=1}^{T} x_{it} x_{Mt}$. Off the diagonal, $\overline{\sigma}_{ij} = \widetilde{\sigma}_{iM} \widetilde{\sigma}_{jM} / \widetilde{\sigma}_{MM}$. $\mathrm{Cov}[\widetilde{\sigma}_{iM}, \widetilde{\sigma}_{ij}]$ can be estimated by $\widehat{v}_{iM,ij} = (1/T^2) \sum_{t=1}^{T} (x_{it} x_{Mt} - \widetilde{\sigma}_{iM})(x_{it} x_{jt} - \widetilde{\sigma}_{ij})$. Similarly, $\mathrm{Cov}[\widetilde{\sigma}_{MM}, \widetilde{\sigma}_{ij}]$ can be estimated by $\widehat{v}_{MM,ij} = (1/T^2) \sum_{t=1}^{T} (x_{Mt}^2 - \widetilde{\sigma}_{MM})(x_{it} x_{jt} - \widetilde{\sigma}_{ij})$. Using the delta method, $\varphi_{ij} = \mathrm{Cov}[\overline{\sigma}_{ij}, \widetilde{\sigma}_{ij}]$ can be estimated by $\widehat{\varphi}_{ij} = \widehat{v}_{iM,ij} \widetilde{\sigma}_{jM} / \widetilde{\sigma}_{MM} + \widehat{v}_{jM,ij} \widetilde{\sigma}_{iM} / \widetilde{\sigma}_{MM} - \widehat{v}_{MM,ij} \widetilde{\sigma}_{iM} \widetilde{\sigma}_{jM} / \widetilde{\sigma}_{MM}^2$, for $i, j = 1, \ldots, N$, $i \neq j$. These formulas yield the estimator $\widehat{\varphi} = (1/N) \sum_{i=1}^{N} \sum_{j=1}^{N} \widehat{\varphi}_{ij}$ that we recommend when $\overline{\Sigma}$ is given by the single index model.

The extension to multiple index models is tedious but straightforward.

C Proofs

We prove the theorems contained in the main body of the text. The propositions in Appendix A and the formulas in Appendix B are not proven.

C.1 Theorem 1

Recall that the matrix Y is defined as $Y = U'X$, where U is a rotation matrix containing the eigenvectors of the covariance matrix Σ. Let $[\lambda_{ij}]_{i,j=1,\ldots,N}$ denote the entries of $\Lambda = U' \Sigma U$. The rotation matrix U is such that $\lambda_{ij} = 0$ when $i \neq j$ and $\lambda_{11}, \ldots, \lambda_{NN}$ are the eigenvalues of the covariance matrix Σ. Please be aware that the eigenvalues of Σ are also denoted $\lambda_1, \ldots, \lambda_N$ elsewhere in the text. Let $r_2^2 = \mathrm{E}[\|\tilde{\Sigma} - \Sigma\|^2]$.

$$\begin{aligned}
r_2^2 &= \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} y_{it}y_{jt} - \lambda_{ij}\right)^2\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T} y_{it}y_{jt} - \lambda_{ij}\right)^2\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{Var}\left[\frac{1}{T}\sum_{t=1}^{T} y_{it}y_{jt}\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{T}\mathrm{Var}\left[y_{i1}y_{j1}\right] \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\mathrm{E}\left[y_{i1}^2 y_{j1}^2\right] - \mathrm{E}\left[y_{i1}y_{j1}\right]^2\right) \\
&= \frac{1}{NT}\sum_{i=1}^{N}\left(\mathrm{E}\left[y_{i1}^4\right] - 2\mathrm{E}\left[y_{i1}^2\right]^2\right) + \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{E}\left[y_{i1}^2\right]\mathrm{E}\left[y_{j1}^2\right] \\
&= \frac{1}{NT}\sum_{i=1}^{N}\left(\mathrm{E}\left[y_{i1}^4\right] - 2\mathrm{E}\left[y_{i1}^2\right]^2\right) + \frac{N}{T}m^2.
\end{aligned}$$

Therefore

$$\begin{aligned}
\left|\frac{N}{T}m^2 - r_2^2\right| &\leq \frac{1}{NT}\sum_{i=1}^{N}\mathrm{E}\left[y_{i1}^4\right] + \frac{2}{NT}\sum_{i=1}^{N}\mathrm{E}\left[y_{i1}^2\right]^2 \\
&\leq \frac{3}{T}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\left[y_{i1}^8\right]} \\
&\leq \frac{3\sqrt{B}}{T} \to 0
\end{aligned}$$

Therefore $(N/t)m^2 - r_2^2 \to 0$. □

C.2 Theorem 2

Let $\tilde{\Sigma} = [\tilde{\sigma}_{ij}]_{i,j=1,\ldots,N}$ and $\Sigma = [\sigma_{ij}]_{i,j=1,\ldots,N}$. Then $\mathrm{E}[(1/N)\sum_{i=1}^{N}\tilde{\lambda}_i] = \mathrm{E}[(1/N)\sum_{i=1}^{N}\tilde{\sigma}_{ii}] = (1/N)\sum_{i=1}^{N}\sigma_{ii} = (1/N)\sum_{i=1}^{N}\lambda_i$. This proves the first statement of Theorem 2.

Now let us prove the second statement. Recall that the matrix Y is defined as $Y = U'X$, where

U is a rotation matrix containing the eigenvectors of the covariance matrix Σ. Let $[y_{it}]_{\substack{i=1,\ldots,N \\ t=1,\ldots,T}}$ denote the entries of the matrix Y. Let $[\lambda_{ij}]_{i,j=1,\ldots,N}$ denote the entries of $\Lambda = U'\Sigma U$. The rotation matrix U is such that $\lambda_{ij} = 0$ when $i \neq j$ and $\lambda_{11}, \ldots, \lambda_{NN}$ are the eigenvalues of the covariance matrix Σ.

$$\begin{aligned}
\mathrm{E}\left[(\widehat{m} - m)^4\right] &= \mathrm{E}\left[\left\{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T}\left(y_{it}^2 - \lambda_{ii}\right)\right\}^4\right] \\
&= \mathrm{E}\left[\left\{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{N}\sum_{i=1}^{N}\left(y_{it}^2 - \lambda_{ii}\right)\right\}^4\right] \\
&= \frac{1}{T^4}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\sum_{t_3=1}^{T}\sum_{t_4=1}^{T}\mathrm{E}\left[\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_{it_1}^2 - \lambda_{ii}\right)\right\}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_{it_2}^2 - \lambda_{ii}\right)\right\}\right. \\
&\qquad\qquad \left.\times\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_{it_3}^2 - \lambda_{ii}\right)\right\}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_{it_4}^2 - \lambda_{ii}\right)\right\}\right] \qquad (20)
\end{aligned}$$

In the summation on the right hand side of Equation (20), the expectation is nonzero only if $t_1 = t_2$ or $t_1 = t_3$ or $t_1 = t_4$ or $t_2 = t_3$ or $t_2 = t_4$ or $t_3 = t_4$. Since these six conditions are symmetric we have:

$$\begin{aligned}
&\mathrm{E}\left[(\widehat{m} - m)^4\right] \\
&\quad\leq \frac{6}{T^4}\sum_{t_1=1}^{T}\sum_{t_3=1}^{T}\sum_{t_4=1}^{T}\left|\mathrm{E}\left[\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_{it_1}^2 - \lambda_{ii}\right)\right\}^2\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_{it_3}^2 - \lambda_{ii}\right)\right\}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_{it_4}^2 - \lambda_{ii}\right)\right\}\right]\right| \\
&\quad\leq \frac{6}{T^4}\sum_{t_1=1}^{T}\sum_{t_3=1}^{T}\sum_{t_4=1}^{T}\sqrt{\mathrm{E}\left[\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_{it_1}^2 - \lambda_{ii}\right)\right\}^4\right]} \\
&\qquad\qquad\times\sqrt{\mathrm{E}\left[\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_{it_3}^2 - \lambda_{ii}\right)\right\}^2\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_{it_4}^2 - \lambda_{ii}\right)\right\}^2\right]} \\
&\quad\leq \frac{6}{T^4}\sum_{t_1=1}^{T}\sum_{t_3=1}^{T}\sum_{t_4=1}^{T}\sqrt{\mathrm{E}\left[\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_{it_1}^2 - \lambda_{ii}\right)\right\}^4\right]} \\
&\qquad\qquad\times\sqrt[4]{\mathrm{E}\left[\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_{it_3}^2 - \lambda_{ii}\right)\right\}^4\right]}\sqrt[4]{\mathrm{E}\left[\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_{it_4}^2 - \lambda_{ii}\right)\right\}^4\right]} \\
&\quad\leq \frac{6}{T}\sqrt{\frac{1}{T}\sum_{t_1=1}^{T}16\mathrm{E}\left[\left\{\frac{1}{N}\sum_{i=1}^{N}y_{it_1}^2\right\}^4\right]}\sqrt[4]{\frac{1}{T}\sum_{t_3=1}^{T}16\mathrm{E}\left[\left\{\frac{1}{N}\sum_{i=1}^{N}y_{it_3}^2\right\}^4\right]}
\end{aligned}$$

$$\times \sqrt[4]{\frac{1}{T}\sum_{t_4=1}^{T} 16\mathrm{E}\left[\left\{\frac{1}{N}\sum_{i=1}^{N} y_{it_4}^2\right\}^4\right]}$$

$$\leq \frac{384B}{T} \to 0$$

where B is defined by Assumption 2. Therefore $\widehat{m} - m$ converges to zero in quartic mean, hence in quadratic mean and in probability. For future reference note that $m = (1/N)\sum_{i=1}^{N} \mathrm{E}[y_{i1}^2] \leq \{(1/N)\sum_{i=1}^{N} \mathrm{E}[y_{i1}^8]\}^{1/4} \leq B^{1/4}$, therefore m is bounded. □

C.3 Theorem 3

We have $\mathrm{E}[(1/N)\sum_{i=1}^{N}(\tilde{\lambda}_i - m)^2] = \mathrm{E}[\|\tilde{\Sigma} - mI\|^2]$ and $\mathrm{E}[(1/N)\sum_{i=1}^{N}(\lambda_i - m)^2] = \mathrm{E}[\|\Sigma - mI\|^2]$. Note that $\Sigma - mI$ and $\tilde{\Sigma} - \Sigma$ are orthogonal in the sense that $\mathrm{E}[(\Sigma - mI) \circ (\tilde{\Sigma} - \Sigma)] = (\Sigma - mI) \circ \mathrm{E}[\tilde{\Sigma} - \Sigma] = (\Sigma - mI) \circ (\Sigma - \Sigma) = 0$. Therefore the triangle $(mI, \Sigma, \tilde{\Sigma})$ has a right angle at Σ. Then Theorem 3 follows from Pythagorus' Theorem. □

C.4 Theorem 4

Let S denote an $N \times N$ symmetric matrix and V an $N \times N$ rotation matrix: $VV' = V'V = I$. First, note that $(1/N)\mathrm{tr}(V'SV) = (1/N)\mathrm{tr}(S)$. The average of the diagonal elements is invariant by rotation. Call it m. Let v_i denote the i^{th} column of V. The dispersion of the diagonal elements of $V'SV$ is $(1/N)\sum_{i=1}^{N}(v_i'Sv_i - m)^2$. Note that $(1/N)\sum_{i=1}^{N}(v_i'Sv_i - m)^2 + (1/N)\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N}(v_i'Sv_j)^2 = (1/N)\mathrm{tr}[(V'SV - mI)^2] = (1/N)\mathrm{tr}[(S - mI)^2]$ is invariant by rotation. Therefore the rotation V maximizes the dispersion of the diagonal elements of $V'SV$ if and only if it minimizes $(1/N)\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N}(v_i'Sv_j)^2$. This is achieved by setting $v_i'Sv_j$ to zero for all $i, j = 1, \ldots, N$, $i \neq j$. In this case, $V'SV$ is a diagonal matrix, call it D. $V'SV = D$ is equivalent to $S = VDV'$. Since V is a rotation and D is diagonal, the column of V must contain the eigenvectors of S and the diagonal of D its eigenvalues. Therefore the dispersion of the diagonal elements of $V'SV$ is maximized when these diagonal elements are equal to the eigenvalues of S. This completes the proof of Theorem 4. □

C.5 Theorem 5

First, we prove that the solution to Equation (6) is of the form $\hat{\Sigma} = \omega mI + (1 - \omega)\tilde{\Sigma}$ for some weight ω. Since $\hat{\Sigma}$ is the orthogonal projection of Σ onto the plane spanned by I and $\tilde{\Sigma}$, $(\hat{\Sigma} - \Sigma)\perp I$ where $\perp$ denotes orthogonality. Since $\tilde{\Sigma}$ is an unbiased estimator of Σ and I is nonstochastic, $\mathrm{E}[\tilde{\Sigma} \circ I] = \Sigma \circ I$ and $(\tilde{\Sigma} - \Sigma)\perp I$. Since mI is the orthogonal projection of Σ onto the line spanned by I, $(\Sigma - mI)\perp I$. Combining the last result with the first two yields $(\hat{\Sigma} - mI)\perp I$ and $(\tilde{\Sigma} - mI)\perp I$, therefore both $\hat{\Sigma} - mI$ and $\tilde{\Sigma} - mI$ belong to the orthogonal of I in the plane spanned by I and $\tilde{\Sigma}$, which is a subspace of dimension one. $\hat{\Sigma} - mI$ and $\tilde{\Sigma} - mI$ must be parallel, which means that $\hat{\Sigma}$ is on the line going from mI to $\tilde{\Sigma}$.

Now, we find the weight ω. The proof relies on elementary geometric relations in the triangle $(mI, \Sigma, \tilde{\Sigma})$ with right angle at Σ. $\hat{\Sigma}$ is the orthogonal projection of Σ onto the line going from mI to $\tilde{\Sigma}$. Let $d_1^2 = \mathrm{E}[||\hat{\Sigma} - mI||^2]$ and $d_2^2 = \mathrm{E}[||\hat{\Sigma} - \tilde{\Sigma}||]$. The cosine of the angle at mI can be expressed in two different ways: d_1/r_1 and r_1/d, therefore the two ratios must be equal and $d_1 = r_1^2/d$. Similarly the cosine of the angle at $\tilde{\Sigma}$ can be expressed in two different ways: d_2/r_2 and r_2/d, therefore the two ratios must be equal and $d_2 = r_2^2/d$. Note that $d_1 + d_2 = d$ as expected. These values for d_1 and d_2 yield $\hat{\Sigma} = (d_2/d)mI + (d_1/d)\tilde{\Sigma} = (r_2^2/d^2)mI + (r_1^2/d^2)\tilde{\Sigma}$.

Finally, we compute the mean squared error of $\hat{\Sigma}$. Let $r_0^2 = \mathrm{E}[||\hat{\Sigma} - \Sigma||^2]$. The angle of $(\tilde{\Sigma}, mI, \Sigma)$ at mI and the angle of $(\hat{\Sigma}, \Sigma, \tilde{\Sigma})$ at Σ are equal. Equating their cosines yields $r_1/d = r_0/r_2$, therefore $r_0 = r_1 r_2/d$ and $||\hat{\Sigma} - \Sigma||^2 = r_1^2 r_2^2/d^2$. Note that Theorem 7 can also be proved by calculus alone.

□

C.6 Theorem 6

First, it is convenient to prove the following lemma.

Lemma 1 $E[||\tilde{\Sigma}||^2]$ *is bounded.*

Let $[\lambda_{ij}]_{i,j=1,\dots,N}$ denote the entries of $\Lambda = U'\Sigma U$. The rotation matrix U is such that $\lambda_{ij} = 0$ when $i \neq j$ and $\lambda_{11}, \dots, \lambda_{NN}$ are the eigenvalues of the covariance matrix Σ.

$$\begin{aligned}
\mathrm{E}[||\tilde{\Sigma} - \Sigma||^2] &= \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} y_{it}y_{jt} - \lambda_{ij}\right)^2\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T} y_{it}y_{jt} - \lambda_{ij}\right)^2\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{Var}\left[\frac{1}{T}\sum_{t=1}^{T} y_{it}y_{jt}\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{T}\mathrm{Var}\left[y_{i1}y_{j1}\right] \\
&\leq \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{E}\left[y_{i1}^2 y_{j1}^2\right] \\
&\leq \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sqrt{\mathrm{E}\left[y_{i1}^4\right]\mathrm{E}\left[y_{j1}^4\right]} \\
&\leq \frac{N}{T}\left(\frac{1}{N}\sum_{i=1}^{N}\sqrt{\mathrm{E}\left[y_{i1}^4\right]}\right)^2 \\
&\leq \frac{N}{T}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\left[y_{i1}^8\right]} \\
&\leq A\sqrt{B}
\end{aligned}$$

where A and B are defined by Assumptions 1-2. Therefore $\mathrm{E}[||\tilde{\Sigma} - \Sigma||^2]$ is bounded. $||\Sigma||^2 = (1/N)\sum_{i=1}^{N}\mathrm{E}[y_{i1}^2]^2 \leq \{(1/N)\sum_{i=1}^{N}\mathrm{E}[y_{i1}^8]\}^{1/2} \leq \sqrt{B}$ implies that $\mathrm{E}[||\tilde{\Sigma}||^2]$ is bounded. For future reference note that it implies that d^2, r_1^2 and r_2^2 are bounded too. □

Now we turn to the proof of Theorem 6. We successively decompose $\widehat{d}^2 - d^2$ into terms that are easier to study.

$$\widehat{d}^2 - d^2 = \left\{ \left\|\tilde{\Sigma} - \widehat{m}I\right\|^2 - \left\|\tilde{\Sigma} - mI\right\|^2 \right\} + \left\{ \left\|\tilde{\Sigma} - mI\right\|^2 - \mathrm{E}\left[\left\|\tilde{\Sigma} - mI\right\|^2 \right] \right\} \tag{21}$$

It is sufficient to show that both bracketed terms on the right hand side of Equation (21) converge to zero in quadratic mean. Consider the first term: $\|\tilde{\Sigma} - \widehat{m}I\|^2 - \|\tilde{\Sigma} - mI\|^2 = (\widehat{m} - m)^2$, therefore by the proof of Theorem 2 it converges to zero in quadratic mean. Now consider the second term.

$$\left\|\tilde{\Sigma} - mI\right\|^2 = m^2 - 2m\widehat{m} + \left\|\tilde{\Sigma}\right\|^2 . \tag{22}$$

Again it is sufficient to show that the three terms on the right hand side of Equation (22) converge to their expectations in quadratic mean. The first term m^2 trivially does. The second term $2m\widehat{m}$ does too by the proof of Theorem 2, keeping in mind that m is bounded. Now consider the third term $\|\tilde{\Sigma}\|^2$.

$$\begin{aligned} \left\|\tilde{\Sigma}\right\|^2 &= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\frac{1}{T} \sum_{t=1}^{T} y_{it} y_{jt} \right)^2 \\ &= \frac{N}{T^2} \sum_{t=1}^{T} \sum_{\tau=1}^{T} \left(\frac{1}{N} \sum_{i=1}^{N} y_{it} y_{i\tau} \right)^2 \\ &= \frac{N}{T^2} \sum_{t=1}^{T} \left(\frac{1}{N} \sum_{i=1}^{N} y_{it}^2 \right)^2 + \frac{N}{T^2} \sum_{t=1}^{T} \sum_{\substack{\tau=1 \\ \tau \neq t}}^{T} \left(\frac{1}{N} \sum_{i=1}^{N} y_{it} y_{i\tau} \right)^2 \end{aligned} \tag{23}$$

Again it is sufficient to show that both terms on the right hand side of Equation (23) converge to their expectations in quadratic mean. Consider the first term.

$$\begin{aligned} \mathrm{Var}\left[\frac{N}{T^2} \sum_{t=1}^{T} \left(\frac{1}{N} \sum_{i=1}^{N} y_{it}^2 \right)^2 \right] &= \frac{N^2}{T^3} \mathrm{Var}\left[\left(\frac{1}{N} \sum_{i=1}^{N} y_{i1}^2 \right)^2 \right] \\ &\leq \frac{N^2}{T^3} \mathrm{E}\left[\left(\frac{1}{N} \sum_{i=1}^{N} y_{i1}^2 \right)^4 \right] \\ &\leq \left(\frac{1}{T} \right) \left(\frac{N}{T} \right)^2 \left(\frac{1}{N} \sum_{i=1}^{N} \mathrm{E}\left[y_{i1}^8 \right] \right) \end{aligned}$$

$$\leq \frac{A^2B}{T} \to 0$$

Therefore the first term on the right hand side of Equation (23) converges to its expectation in quadratic mean. Now consider the second term.

$$\begin{aligned}
\operatorname{Var}&\left[\frac{N}{T^2}\sum_{t=1}^{T}\sum_{\substack{\tau=1\\ \tau\neq t}}^{T}\left(\frac{1}{N}\sum_{i=1}^{N} y_{it}y_{i\tau}\right)^2\right] \\
&= \frac{N^2}{T^4}\sum_{t_1=1}^{T}\sum_{\substack{\tau_1=1\\ \tau_1\neq t_1}}^{T}\sum_{t_2=1}^{T}\sum_{\substack{\tau_2=1\\ \tau_2\neq t_2}}^{T} \operatorname{Cov}\left[\left(\frac{1}{N}\sum_{i=1}^{N} y_{it_1}y_{i\tau_1}\right)^2, \left(\frac{1}{N}\sum_{i=1}^{N} y_{it_2}y_{i\tau_2}\right)^2\right] \quad (24)
\end{aligned}$$

The covariances on the right hand side of Equation (24) only depend on $(\{t_1, \tau_1\} \cap \{t_2, \tau_2\})^{\#}$ the number of elements in the intersection of the set $\{t_1, \tau_1\}$ with the set $\{t_2, \tau_2\}$. This number can be zero, one or two. We study each case separately.

$(\{t_1, \tau_1\} \cap \{t_2, \tau_2\})^{\#} = 0$

In this case $((1/N)\sum_{i=1}^{N} y_{it_1}y_{i\tau_1})^2$ and $((1/N)\sum_{i=1}^{N} y_{it_2}y_{i\tau_2})^2$ are independent, so their covariance is zero.

$(\{t_1, \tau_1\} \cap \{t_2, \tau_2\})^{\#} = 1$

This case occurs $4t(t-1)(t-2)$ times in the summation on the right hand side of Equation (24). Each time we have:

$$\begin{aligned}
\operatorname{Cov}&\left[\left(\frac{1}{N}\sum_{i=1}^{N} y_{it_1}y_{i\tau_1}\right)^2, \left(\frac{1}{N}\sum_{i=1}^{N} y_{it_2}y_{i\tau_2}\right)^2\right] \\
&= \operatorname{Cov}\left[\left(\frac{1}{N}\sum_{i=1}^{N} y_{i1}y_{i2}\right)^2, \left(\frac{1}{N}\sum_{i=1}^{N} y_{i1}y_{i3}\right)^2\right] \\
&\leq \operatorname{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} y_{i1}y_{i2}\right)^2 \left(\frac{1}{N}\sum_{i=1}^{N} y_{i1}y_{i3}\right)^2\right] \\
&\leq \operatorname{E}\left[\frac{1}{N^4}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N} y_{i1}y_{i2}y_{j1}y_{j2}y_{k1}y_{k3}y_{l1}y_{l3}\right] \\
&\leq \frac{1}{N^4}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N} \operatorname{E}\left[y_{i1}y_{j1}y_{k1}y_{l1}\right]\operatorname{E}\left[y_{i2}y_{j2}\right]\operatorname{E}\left[y_{k3}y_{l3}\right]
\end{aligned}$$

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

and

$$
\begin{aligned}
-\mathrm{Cov}&\left[\left(\frac{1}{N} \sum_{i=1}^{N} y_{it_1} y_{i\tau_1} \right)^2, \left(\frac{1}{N} \sum_{i=1}^{N} y_{it_2} y_{i\tau_2} \right)^2 \right] \\
&= -\mathrm{Cov}\left[\left(\frac{1}{N} \sum_{i=1}^{N} y_{i1} y_{i2} \right)^2, \left(\frac{1}{N} \sum_{i=1}^{N} y_{i1} y_{i3} \right)^2 \right] \\
&\leq \mathrm{E}\left[\left(\frac{1}{N} \sum_{i=1}^{N} y_{i1} y_{i2} \right)^2 \right] \mathrm{E}\left[\left(\frac{1}{N} \sum_{i=1}^{N} y_{i1} y_{i3} \right)^2 \right] \\
&\leq \left(\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{E}\left[y_{i1} y_{j1}\right]^2 \right)^2 \\
&\leq \frac{1}{N^2} \left(\frac{1}{N} \sum_{i=1}^{N} \mathrm{E}\left[y_{i1}^8\right] \right) \\
&\leq \frac{B}{N^2}.
\end{aligned}
$$

Therefore in this case the absolute value of the covariance on the right hand side of Equation (24) is bounded by B/N^2.

$\underline{(\{t_1, \tau_1\} \cap \{t_2. \tau_2\})^{\#} = 2}$

This case occurs $2t(t-1)$ times in the summation on the right hand side of Equation (24). Each time we have:

$$
\left| \mathrm{Cov}\left[\left(\frac{1}{N} \sum_{i=1}^{N} y_{it_1} y_{i\tau_1} \right)^2, \left(\frac{1}{N} \sum_{i=1}^{N} y_{it_2} y_{i\tau_2} \right)^2 \right] \right|
$$

$$\begin{aligned} &= \left|\text{Cov}\left[\left(\frac{1}{N}\sum_{i=1}^{N} y_{i1}y_{i2}\right)^2, \left(\frac{1}{N}\sum_{i=1}^{N} y_{i1}y_{i2}\right)^2\right]\right| \\ &\leq \frac{1}{N^4}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N} \left|\text{Cov}\left[y_{i1}y_{i2}y_{j1}y_{j2}, y_{k1}y_{k2}y_{l1}y_{l2}\right]\right| \qquad (25) \end{aligned}$$

Now consider the summation on the right hand side of Equation (25). When i, j, k, l are all pairwise distinct, Assumption 3 ensures that $\text{E}[y_{i1}y_{j1}y_{k1}y_{l1}] = \text{E}[y_{i1}y_{j1}]\text{E}[y_{k1}y_{l1}]$, which in turn implies:

$$\begin{aligned} \text{Cov}\left[y_{i1}y_{i2}y_{j1}y_{j2}, y_{k1}y_{k2}y_{l1}y_{l2}\right] &= \text{E}\left[y_{i1}y_{i2}y_{j1}y_{j2}y_{k1}y_{k2}y_{l1}y_{l2}\right] \\ &\quad - \text{E}\left[y_{i1}y_{i2}y_{j1}y_{j2}\right]\text{E}\left[y_{k1}y_{k2}y_{l1}y_{l2}\right] \\ &= \text{E}\left[y_{i1}y_{j1}y_{k1}y_{l1}\right]^2 - \text{E}\left[y_{i1}y_{j1}\right]^2\text{E}\left[y_{k1}y_{l1}\right] \\ &= 0. \end{aligned}$$

Therefore the summation on the right hand side of Equation (25) only extends over the set $S = \{(i, j, k, l) : i, j, k, l = 1, \ldots, N; \{i, j, k, l\}^{\#} \leq 3\}$, with the convention that $\{2, 2, 3, 4\}^{\#} = 3$.

$$\begin{aligned} &\left|\text{Cov}\left[\left(\frac{1}{N}\sum_{i=1}^{N} y_{it_1}y_{i\tau_1}\right)^2 \cdot \left(\frac{1}{N}\sum_{i=1}^{N} y_{it_2}y_{i\tau_2}\right)^2\right]\right| \\ &\qquad \leq \frac{1}{N^4}\sum_{(i,j,k,l)\in S} \left|\text{Cov}\left[y_{i1}y_{i2}y_{j1}y_{j2}, y_{k1}y_{k2}y_{l1}y_{l2}\right]\right| \\ &\qquad \leq \frac{1}{N^4}\sum_{(i,j,k,l)\in S} \sqrt{\text{E}\left[y_{i1}^2y_{i2}^2y_{j1}^2y_{j2}^2\right]\text{E}\left[y_{k1}^2y_{k2}^2y_{l1}^2y_{l2}^2\right]} \\ &\qquad \leq \frac{1}{N^4}\sum_{(i,j,k,l)\in S} \text{E}\left[y_{i1}^2y_{j1}^2\right]\text{E}\left[y_{k1}^2y_{l1}^2\right] \\ &\qquad \leq \frac{1}{N^4}\sum_{(i,j,k,l)\in S} \sqrt{\text{E}\left[y_{i1}^4\right]\text{E}\left[y_{j1}^4\right]\text{E}\left[y_{k1}^4\right]\text{E}\left[y_{l1}^4\right]} \end{aligned}$$

The summation only extends over the quadruples (i, j, k, l) where $i = j$ or $i = k$ or $i = l$ or $j = k$ or $j = l$ or $k = l$. Since these six conditions are symmetric we have:

$$\left|\text{Cov}\left[\left(\frac{1}{N}\sum_{i=1}^{N} y_{it_1}y_{i\tau_1}\right)^2 \cdot \left(\frac{1}{N}\sum_{i=1}^{N} y_{it_2}y_{i\tau_2}\right)^2\right]\right|$$

$$\begin{aligned}
&\leq \frac{6}{N^4}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{k=1}^{N}\sqrt{\mathrm{E}\left[y_{i1}^4\right]\mathrm{E}\left[y_{i1}^4\right]\mathrm{E}\left[y_{k1}^4\right]\mathrm{E}\left[y_{l1}^4\right]}\\
&\leq \frac{6}{N}\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\left[y_{i1}^4\right]\right)\left(\frac{1}{N}\sum_{i=1}^{N}\sqrt{\mathrm{E}\left[y_{i1}^4\right]}\right)^2\\
&\leq \frac{6}{N}\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\left[y_{i1}^4\right]\right)^2\\
&\leq \frac{6}{N}\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\left[y_{i1}^8\right]\right)\\
&\leq \frac{6B}{N}.
\end{aligned}$$

Having studied the three possible cases, we can now bound the summation on the right hand side of Equation (24):

$$\begin{aligned}
&\frac{N^2}{T^4}\sum_{t_1=1}^{T}\sum_{\substack{\tau_1=1\\ \tau_1\neq t_1}}^{T}\sum_{t_2=1}^{T}\sum_{\substack{\tau_2=1\\ \tau_2\neq t_2}}^{T}\left|\mathrm{Cov}\left[\left(\frac{1}{N}\sum_{i=1}^{N}y_{it_1}y_{i\tau_1}\right)^2,\left(\frac{1}{N}\sum_{i=1}^{N}y_{it_2}y_{i\tau_2}\right)^2\right]\right|\\
&\qquad\leq \frac{N^2}{T^4}\left\{4t(t-1)(t-2)\frac{B}{N^2}+2t(t-1)\frac{6B}{N}\right\}\\
&\qquad\leq \frac{4B(1+3A)}{T}\rightarrow 0.
\end{aligned}$$

Backing up, the second term on the right hand side of Equation (23) converges to its expectation in quadratic mean. Backing up again, the third term $\|\tilde{\Sigma}\|^2$ on the right hand side of Equation (22) converges to its expectation in quadratic mean. Backing up more, the second bracketed term on the right hand side of Equation (21) converges to zero in quadratic mean. Backing up one last time, $\widehat{d}^2 - d^2$ converges to zero in quadratic mean, hence in probability. For future reference note that, since $\|\tilde{\Sigma} - mI\|^2$ converges to its expectation d^2 in quadratic mean and since d^2 is bounded, $\mathrm{E}[\|\tilde{\Sigma} - mI\|^4$ is bounded. □

C.7 Theorem 7

Again we prove this theorem by successively decomposing $\hat{r}_2^2 - r_2^2$ into terms that are easier to study.

$$\begin{aligned} \hat{r}_2^2 - r_2^2 &= \left\{ \frac{1}{T^2} \sum_{t=1}^{T} \left\| x_{\cdot t} x_{\cdot t}{}^{\mathsf{T}} - \Sigma \right\|^2 - \mathrm{E}\left[\left\| \tilde{\Sigma} - \Sigma \right\|^2 \right] \right\} \\ &\quad + \left\{ \frac{1}{T^2} \sum_{t=1}^{T} \left\| x_{\cdot t} x_{\cdot t}{}^{\mathsf{T}} - \tilde{\Sigma} \right\|^2 - \frac{1}{T^2} \sum_{t=1}^{T} \left\| x_{\cdot t} x_{\cdot t}{}^{\mathsf{T}} - \Sigma \right\|^2 \right\} \end{aligned} \tag{26}$$

It is sufficient to show that both bracketed terms on the right hand side of Equation (26) converge to zero in quadratic mean. Consider the first term.

$$\begin{aligned} \mathrm{E}\left[\left\| \tilde{\Sigma} - \Sigma \right\|^2 \right] &= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{E}\left[\left(\frac{1}{T} \sum_{t=1}^{T} x_{it} x_{jt} - \sigma_{ij} \right)^2 \right] \\ &= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{Var}\left[\frac{1}{T} \sum_{t=1}^{T} x_{it} x_{jt} - \sigma_{ij} \right] \\ &= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{1}{T^2} \sum_{t=1}^{T} \mathrm{Var}\left[x_{it} x_{jt} - \sigma_{ij} \right] \\ &= \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{Var}\left[x_{i1} x_{j1} - \sigma_{ij} \right] \\ &= \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{E}\left[\left(x_{i1} x_{j1} - \sigma_{ij} \right)^2 \right] \\ &= \mathrm{E}\left[\frac{1}{T} \left\| x_1 x_1{}^{\mathsf{T}} - \Sigma \right\|^2 \right] \\ &= \mathrm{E}\left[\frac{1}{T^2} \sum_{t=1}^{T} \left\| x_{\cdot t} x_{\cdot t}{}^{\mathsf{T}} - \Sigma \right\|^2 \right] \end{aligned}$$

Therefore the first bracketed term on the right hand side of Equation (26) has expectation zero. For $t = 1, \ldots, T$ let $y_{\cdot t}$ denote the $n \times 1$ vector holding the t^{th} column of the matrix Y.

$$\begin{aligned} \mathrm{Var}\left[\frac{1}{T^2} \sum_{t=1}^{T} \left\| x_{\cdot t} x_{\cdot t}{}^{\mathsf{T}} - \Sigma \right\|^2 \right] &= \frac{1}{T} \mathrm{Var}\left[\frac{1}{T} \left\| x_{\cdot 1} x_{\cdot 1}{}^{\mathsf{T}} - \Sigma \right\|^2 \right] \\ &= \frac{1}{T} \mathrm{Var}\left[\frac{1}{T} \left\| y_{\cdot 1} y_{\cdot 1}{}^{\mathsf{T}} - \Lambda \right\|^2 \right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{N^2T^3}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N}\operatorname{Cov}\left[y_{i1}y_{j1}-\lambda_{ij},\, y_{k1}y_{l1}-\lambda_{kl}\right] \\
&= \frac{1}{N^2T^3}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N}\operatorname{Cov}\left[y_{i1}y_{j1},\, y_{k1}y_{l1}\right] \\
&\leq \frac{1}{N^2T^3}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N}\sqrt{\mathrm{E}\left[y_{i1}^2y_{j1}^2\right]\mathrm{E}\left[y_{k1}^2y_{l1}^2\right]} \\
&\leq \frac{1}{N^2T^3}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N}\sqrt[4]{\mathrm{E}\left[y_{i1}^4\right]\mathrm{E}\left[y_{j1}^4\right]\mathrm{E}\left[y_{k1}^4\right]\mathrm{E}\left[y_{l1}^4\right]} \\
&\leq \frac{N^2}{T^3}\left(\frac{1}{N}\sum_{i=1}^{N}\sqrt[4]{\mathrm{E}\left[y_{i1}^4\right]}\right)^4 \\
&\leq \frac{N^2}{T^3}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\left[y_{i1}^8\right]} \\
&\leq \frac{A^2\sqrt{B}}{T}
\end{aligned}
$$

Therefore the first bracketed term on the right hand side of Equation (26) converges to zero in quadratic mean. For future reference note that, since $\mathrm{E}[\|\tilde{\Sigma}-\Sigma\|^2]$ is bounded, it implies that $\mathrm{E}[\{(1/T^2)\sum_{t=1}^{T}\|x_{\cdot t}x_{\cdot t}^{\top}\|^2\}^2]$ is bounded. Now consider the second term.

$$
\begin{aligned}
&\mathrm{E}\left[\left\{\frac{1}{T^2}\sum_{t=1}^{T}\left\|x_{\cdot t}x_{\cdot t}^{\top}-\tilde{\Sigma}\right\|^2-\frac{1}{T^2}\sum_{t=1}^{T}\left\|x_{\cdot t}x_{\cdot t}^{\top}-\Sigma\right\|^2\right\}^2\right] \\
&\qquad\leq \mathrm{E}\left[\frac{1}{T^3}\sum_{t=1}^{T}\left\{\left\|x_{\cdot t}x_{\cdot t}^{\top}-\tilde{\Sigma}\right\|^2-\left\|x_{\cdot t}x_{\cdot t}^{\top}-\Sigma\right\|^2\right\}^2\right] \\
&\qquad\leq \mathrm{E}\left[\frac{1}{T^3}\sum_{t=1}^{T}\left\{2\left(\tilde{\Sigma}-\Sigma\right)\circ\left(x_{\cdot t}x_{\cdot t}^{\top}-\frac{\tilde{\Sigma}+\Sigma}{2}\right)\right\}^2\right] \\
&\qquad\leq \mathrm{E}\left[\frac{4}{T^3}\sum_{t=1}^{T}\left\|\tilde{\Sigma}-\Sigma\right\|^2\left\|x_{\cdot t}x_{\cdot t}^{\top}-\frac{\tilde{\Sigma}+\Sigma}{2}\right\|^2\right] \\
&\qquad\leq \mathrm{E}\left[\frac{4}{T}\left\|\tilde{\Sigma}-\Sigma\right\|^2\left(\frac{1}{T^2}\sum_{t=1}^{T}\left\|x_{\cdot t}x_{\cdot t}^{\top}-\frac{\tilde{\Sigma}+\Sigma}{2}\right\|^2\right)\right] \\
&\qquad\leq \frac{4}{T}\sqrt{\mathrm{E}\left[\left\|\tilde{\Sigma}-\Sigma\right\|^4\right]}\sqrt{\mathrm{E}\left[\left(\frac{1}{T^2}\sum_{t=1}^{T}\left\|x_{\cdot t}x_{\cdot t}^{\top}-\frac{\tilde{\Sigma}+\Sigma}{2}\right\|^2\right)^2\right]} \qquad (27)
\end{aligned}
$$

It is sufficient to show that the last two terms on the right hand side of Equation (27) are bounded. It is true for $\mathrm{E}[\|\tilde{\Sigma} - \Sigma\|^4]$ since $\mathrm{E}[\|\tilde{\Sigma} - mI\|^4]$ and $\|\Sigma - mI\|$ are bounded. Now consider the last term.

$$\frac{1}{T^2}\sum_{t=1}^{T}\left\|x_{\cdot t}x_{\cdot t}^{\mathsf{T}} - \frac{\tilde{\Sigma}+\Sigma}{2}\right\|^2 \leq \frac{2}{T^2}\sum_{t=1}^{T}\left\|x_{\cdot t}x_{\cdot t}^{\mathsf{T}} - \Sigma\right\|^2 + \frac{1}{2T}\left\|\tilde{\Sigma} - \Sigma\right\|^2$$

Since $\mathrm{E}[\{(1/T^2)\sum_{t=1}^{T}\|x_{\cdot t}x_{\cdot t}^{\mathsf{T}}\|^2\}^2]$ and $\mathrm{E}[\|\tilde{\Sigma} - \Sigma\|^4]$ are bounded, so is the last term on the right hand side of Equation (27). Backing up, the second term on the right hand side of Equation (26) converges to zero in quadratic mean. Backing up once more, $\widehat{r}_2^2 - r_2^2$ converges to zero in quadratic mean, hence in probability. □

C.8 Theorem 8

Follows trivially from the previous two theorems. □

C.9 Theorem 9

C.9.1 $\|\widehat{\widehat{\Sigma}} - \widehat{\Sigma}\|^2 \xrightarrow{P} 0$

As usual the subscript $_t$, which should index all quantities unless otherwise specified, has been omitted to make notation lighter.

$$\begin{aligned}\left\|\widehat{\widehat{\Sigma}} - \widehat{\Sigma}\right\| &= \left\|\frac{\widehat{r}_2^2}{\widehat{d}^2}(\widehat{m} - m)I + \left(\frac{\widehat{r}_1^2}{\widehat{d}^2} - \frac{r_1^2}{d^2}\right)\left(\tilde{\Sigma} - mI\right)\right\| \\ &\leq |\widehat{m} - m| + \left|\frac{\left(\widehat{r}_1^2 - r_1^2\right)d^2 - r_1^2\left(\widehat{d}^2 - d^2\right)}{\widehat{d}^2 d^2}\right|\left\|\tilde{\Sigma} - mI\right\| \end{aligned} \qquad (28)$$

It is sufficient to prove that both terms on the right hand side of Equation (28) converge to zero in probability. The first term does by Theorem 2. Now consider the second term. Note that both its factors $|(\widehat{r}_1^2 - r_1^2)d^2 - r_1^2(\widehat{d}^2 - d^2)|/(\widehat{d}^2 d^2)$ and $\|\tilde{\Sigma} - mI\|$ are bounded in probability, therefore it is sufficient to prove that either one of them converges to zero in probability. Since d^2 and r_1^2 are bounded by Lemma 1, we have: $(\widehat{r}_1^2 - r_1^2)d^2 - r_1^2(\widehat{d}^2 - d^2) \xrightarrow{P} 0$. Let S_1 denote the set of indices t such that

$$\left|\frac{\left(\widehat{r}_1^2 - r_1^2\right)d^2 - r_1^2\left(\widehat{d}^2 - d^2\right)}{\widehat{d}^2 d^2}\right| \leq \sqrt{\left|\left(\widehat{r}_1^2 - r_1^2\right)d^2 - r_1^2\left(\widehat{d}^2 - d^2\right)\right|}.$$

If the set S_1 is infinite then $|(\widehat{r}_1^2 - r_1^2)d^2 - r_1^2(\widehat{d^2} - d^2)|/(\widehat{d^2}d^2) \xrightarrow{P} 0$ as t tends to infinity inside the set S_1, and so does the second term on the right hand side of Equation (28). If the complementary to the set S_1 is infinite then $\widehat{d^2}d^2 \leq |(\widehat{r}_1^2 - r_1^2)d^2 - r_1^2(\widehat{d^2} - d^2)|^{1/2} \xrightarrow{P} 0$ as t tends to infinity outside the set S_1. By Theorem 6 it implies that $d \to 0$, therefore $\|\widetilde{\Sigma} - mI\| \xrightarrow{P} 0$ as t tends to infinity outside the set S_1, and so does the second term on the right hand side of Equation (28). Bringing together the results obtained for t inside and outside the set S_1 yields that the second term on the right hand side of Equation (28) converges to zero in probability. Backing up, $\|\widehat{\widehat{\Sigma}} - \widehat{\Sigma}\| \xrightarrow{P} 0$ and so does $\|\widehat{\widehat{\Sigma}} - \widehat{\Sigma}\|^2$. □

C.9.2 $\mathrm{E}[\|\widehat{\widehat{\Sigma}} - \Sigma\|^2] - \mathrm{E}[\|\widehat{\Sigma} - \Sigma\|^2] \to 0$

$$\begin{aligned}
\mathrm{E}\left[\left|\left\|\widehat{\widehat{\Sigma}} - \Sigma\right\|^2 - \left\|\widehat{\Sigma} - \Sigma\right\|^2\right|\right] &= \mathrm{E}\left[\left|\left(\widehat{\widehat{\Sigma}} - \widehat{\Sigma}\right) \circ \left(\widehat{\widehat{\Sigma}} + \widehat{\Sigma} - 2\Sigma\right)\right|\right] \\
&\leq \sqrt{\mathrm{E}\left[\left\|\widehat{\widehat{\Sigma}} - \widehat{\Sigma}\right\|^2\right]}\sqrt{\mathrm{E}\left[\left\|\widehat{\widehat{\Sigma}} + \widehat{\Sigma} - 2\Sigma\right\|^2\right]} \qquad (29)
\end{aligned}$$

It is sufficient to prove that the first term on the right hand side of Equation (29) converges to zero and that the second term is bounded. Consider the first term.

$$\begin{aligned}
\mathrm{E}\left[\left\|\widehat{\widehat{\Sigma}} - \widehat{\Sigma}\right\|^2\right] &= \mathrm{E}\left[\left\|\frac{r_2^2}{d^2}(\widehat{m} - m)I + \left(\frac{\widehat{r}_1^2}{\widehat{d^2}} - \frac{r_1^2}{d^2}\right)\left(\widetilde{\Sigma} - \widehat{m}I\right)\right\|^2\right] \\
&= \mathrm{E}\left[\frac{r_2^4}{d^4}(\widehat{m} - m)^2\right] + \mathrm{E}\left[\left(\frac{\widehat{r}_1^2}{\widehat{d^2}} - \frac{r_1^2}{d^2}\right)^2 \left\|\widetilde{\Sigma} - \widehat{m}I\right\|^2\right] \\
&\leq \mathrm{E}\left[(\widehat{m} - m)^2\right] + \mathrm{E}\left[\frac{\left(\widehat{r}_1^2 d^2 - r_1^2\widehat{d^2}\right)^2}{\widehat{d^2}d^4}\right] \qquad (30)
\end{aligned}$$

It is sufficient to show that both terms on the right hand side of Equation (30) converges to zero. The first term does by the proof of Theorem 2. Now consider the second term. Since $\widehat{r}_1^2 \leq \widehat{d^2}$ and $\widehat{r}_1^2 \leq \widehat{d^2}$, note for future reference that $(\widehat{r}_1^2 d^2 - r_1^2\widehat{d^2})^2/(\widehat{d^2}d^4) \leq 4\widehat{d^2}$. Fix $\varepsilon > 0$. Let S_3 denote the set of indices t such that $d^2 \leq \varepsilon/8$. Since $\widehat{d^2} - d^2 \to 0$ in quadratic mean,

$\exists T_1 \quad \forall T \quad T \geq T_1 \Rightarrow \mathrm{E}[|\widehat{d^2} - d^2|] \leq \varepsilon/8$. We have:

$$
\begin{aligned}
\forall T \quad T \in S_3, t \geq T_1 \Rightarrow \mathrm{E}\left[\frac{\left(\hat{r}_1^2 d^2 - r_1^2 \widehat{d^2}\right)^2}{\widehat{d^2} d^4}\right] &\leq 4\mathrm{E}\left[\widehat{d^2}\right] \\
&\leq 4\mathrm{E}\left[\left|\widehat{d^2} - d^2\right|\right] + 4d^2 \\
&\leq 4\frac{\varepsilon}{8} + 4\frac{\varepsilon}{8} \\
&\leq \varepsilon. \qquad (31)
\end{aligned}
$$

Since $\hat{r}_1^2 - r_1^2$ and $\widehat{d^2} - d^2$ converge to zero in quadratic mean and since r_1^2 and d^2 are bounded, $\hat{r}_1^2 d^2 - r_1^2 \widehat{d^2} = (\hat{r}_1^2 - r_1^2)d^2 - r_1^2(\widehat{d^2} - d^2) \to 0$ in quadratic mean, therefore $\exists T_2 \quad \forall T \quad T \geq T_2 \Rightarrow \mathrm{E}[(\hat{r}_1^2 d^2 - r_1^2 \widehat{d^2})^2] \leq \varepsilon^4/1024$. Denote $\Pr(\cdot)$ the probability of an event. We have: $\forall T \quad T \notin S_3. T \geq T_2 \Rightarrow$

$$
\begin{aligned}
\mathrm{E}\left[\frac{\left(\hat{r}_1^2 d^2 - r_1^2 \widehat{d^2}\right)^2}{\widehat{d^2} d^4}\right] &= \mathrm{E}\left[\left.\frac{\left(\hat{r}_1^2 d^2 - r_1^2 \widehat{d^2}\right)^2}{\widehat{d^2} d^4}\right| \widehat{d^2} \leq \frac{\varepsilon}{8}\right] \Pr\left(\widehat{d^2} \leq \frac{\varepsilon}{8}\right) \\
&\quad + \mathrm{E}\left[\left.\frac{\left(\hat{r}_1^2 d^2 - r_1^2 \widehat{d^2}\right)^2}{\widehat{d^2} d^4}\right| \widehat{d^2} > \frac{\varepsilon}{8}\right] \Pr\left(\widehat{d^2} > \frac{\varepsilon}{8}\right) \\
&\leq \mathrm{E}\left[4\widehat{d^2} \middle| \widehat{d^2} \leq \frac{\varepsilon}{8}\right] \Pr\left(\widehat{d^2} \leq \frac{\varepsilon}{8}\right) \\
&\quad + \frac{8}{\varepsilon d^4}\mathrm{E}\left[\left(\hat{r}_1^2 d^2 - r_1^2 \widehat{d^2}\right)^2 \middle| \widehat{d^2} > \frac{\varepsilon}{8}\right] \Pr\left(\widehat{d^2} > \frac{\varepsilon}{8}\right) \\
&\leq 4\frac{\varepsilon}{8} + \frac{512}{\varepsilon^3}\mathrm{E}\left[\left(\hat{r}_1^2 d^2 - r_1^2 \widehat{d^2}\right)^2\right] \\
&\leq \frac{\varepsilon}{2} + \frac{512}{\varepsilon^3}\frac{\varepsilon^4}{1024} \\
&\leq \varepsilon. \qquad (32)
\end{aligned}
$$

Bringing together the results from Equations (31)-(32) yields:

$$
\forall T \quad T \geq \max(T_1 . T_2) \Rightarrow \mathrm{E}\left[\frac{\left(\hat{r}_1^2 d^2 - r_1^2 \widehat{d^2}\right)^2}{\widehat{d^2} d^4}\right] \leq \varepsilon,
$$

therefore the second term on the right hand side of Equation (30) converges to zero. Backing up, the first term on the right hand side of Equation (29) converges to zero. Since $\mathrm{E}[\|\hat{\Sigma}-\Sigma\|^2]$ is bounded, it implies that the second term on the right hand side of Equation (29) is bounded too. Backing up once more yields $\mathrm{E}[\|\hat{\hat{\Sigma}}-\Sigma\|^2]-\mathrm{E}[\|\hat{\Sigma}-\Sigma\|^2]\rightarrow 0$. □

C.9.3 $(\hat{r}_1^2\hat{r}_2^2/\hat{d}^2)-(r_1^2r_2^2/d^2)\xrightarrow{\mathrm{P}}0$

$$\frac{\hat{r}_1^2\hat{r}_2^2}{\hat{d}^2}-\frac{r_1^2r_2^2}{d^2}=\frac{\left(\hat{r}_1^2\hat{r}_2^2-r_1^2r_2^2\right)d^2-r_1^2r_2^2\left(\hat{d}^2-d^2\right)}{\hat{d}^2d^2}$$

By Theorems 6-8 and Lemma 1 the numerator on the right hand side converges to zero in probability. Let S_2 denote the set of indices t such that

$$\left|\frac{\left(\hat{r}_1^2\hat{r}_2^2-r_1^2r_2^2\right)d^2-r_1^2r_2^2\left(\hat{d}^2-d^2\right)}{\hat{d}^2d^2}\right|\leq\sqrt{\left|\left(\hat{r}_1^2\hat{r}_2^2-r_1^2r_2^2\right)d^2-r_1^2r_2^2\left(\hat{d}^2-d^2\right)\right|}.$$

If the set S_2 is infinite then $(\hat{r}_1^2\hat{r}_2^2/\hat{d}^2)-(r_1^2r_2^2/d^2)\xrightarrow{\mathrm{P}}0$ as t tends to infinity inside the set S_2. If the complementary to the set S_2 is infinite then $\hat{d}^2d^2\leq|(\hat{r}_1^2\hat{r}_2^2-r_1^2r_2^2)d^2-r_1^2r_2^2(\hat{d}^2-d^2)|^{1/2}\xrightarrow{P}0$ as t tends to infinity outside the set S_2. By Theorem 6 it implies that $d^2\rightarrow 0$, therefore $(r_1^2r_2^2/d^2)\xrightarrow{\mathrm{P}}0$ as t tends to infinity outside the set S_2, and so does $(\hat{r}_1^2\hat{r}_2^2/\hat{d}^2)$. Bringing together the results obtained for t inside and outside the set S_2 yields $(\hat{r}_1^2\hat{r}_2^2/\hat{d}^2)-(r_1^2r_2^2/d^2)\xrightarrow{\mathrm{P}}0$. □

C.10 Theorem 10

This is similar to the proof of Theorem 5. $\hat{\Sigma}$ is the orthogonal projection of Σ on the line between $\bar{\Sigma}$ and $\tilde{\Sigma}$. Let $d_1^2=\mathrm{E}[\|\hat{\Sigma}-mI\||]$, $d_2^2=\mathrm{E}[\|\hat{\Sigma}-\tilde{\Sigma}\||]$ and $r_0^2=\mathrm{E}[\|\hat{\Sigma}-\Sigma\|^2]$. The orthogonality condition $(\bar{\Sigma}-\hat{\Sigma})\perp(\Sigma-\hat{\Sigma})$ implies $d_1^2+r_0^2=r_1^2$. Also, the orthogonality condition $(\tilde{\Sigma}-\hat{\Sigma})\perp(\Sigma-\hat{\Sigma})$ implies $d_2^2+r_0^2=r_2^2$. Subtracting one equation from the other yields $d_1^2-d_2^2=r_1^2-r_2^2$. Since $\bar{\Sigma}$, $\hat{\Sigma}$ and $\tilde{\Sigma}$ are aligned, we have $d_1+d_2=d$, which implies $d_1^2-d_2^2=d_1^2+(d-d_1)^2=2d_1d-d^2$. Therefore $2d_1d-d^2=r_1^2-r_2^2$, i.e. $d_1=(r_1^2+d^2-r_2^2)/2d$. By symmetry, $d_2=(r_2^2+d^2-r_1^2)/2d$. Note that $d_1+d_2=d$ as expected. These values for d_1 and d_2 yield $\hat{\Sigma}=(d_2/d)\bar{\Sigma}+(d_1/d)\tilde{\Sigma}=[(r_2^2+d^2-r_1^2)/(2d^2)]\bar{\Sigma}+[(r_1^2+d^2-r_2^2)/(2d^2)]\tilde{\Sigma}$.

	Structured	Shrinkage	T-Statistic
$\widehat{m}I$	20.3	10.9	7.01
B.1	20.3	10.6	7.20
B.2	16.0	9.6	8.33
B.3	13.8	9.6	6.37
B.4	11.5	9.3	4.94

Table 2: Comparison of the Ex-Post Standard Deviations of Ex-Ante Minimum Variance Portfolios. Standard deviations are quoted in percents on an annual basis. The portfolios are obtained using a structured estimator of the covariance matrix, or its associated shrinkage estimator. The t-statistic tests the null hypothesis that a given structured estimator and its associated shrinkage estimator yield ex-ante minimum variance portfolios with the same ex-post variance of returns. This hypothesis is rejected in all five cases. Shrinkage helps portfolio selection minimize variance.

	Structured With Hindsight	Shrinkage Without Hindsight	T-Statistic
$\widehat{m}I$	11.8	10.9	1.88
B.1	11.8	10.6	3.80
B.2	11.0	9.6	4.38
B.3	12.4	9.6	5.53
B.4	11.0	9.3	5.65

Table 3: Comparison of the Ex-Post Standard Deviations of Minimum Variance Portfolios. Standard deviations are quoted in percents on an annual basis. The portfolios are obtained using a structured estimator of the covariance matrix, or its associated shrinkage estimator. For structured estimators, the minimum variance portfolio is chosen ex-post among linear combinations of three portfolios that span the ex-ante mean-variance efficient set, assuming that returns are driven by beta and size only. For shrinkage estimators, the minimum variance portfolio is chosen ex-ante, without the benefit of hindsight. This makes it harder to help portfolio selection minimize variance. The t-statistic tests the null hypothesis that a given structured estimator and its associated shrinkage estimator yield minimum variance portfolios with the same ex-post variance of returns. All reject the null. The t-statistic of 1.88 is significant at the 5% level against the one-sided alternative that shrinkage helps minimize variance.

	Plain Regression	Excluding January	Including Size	1963-1992
Slope	2.33	-0.77	0.33	1.88
Standard Error	(2.27)	(2.31)	(1.93)	(3.15)
T-Statistic	1.03	-0.33	0.17	0.60

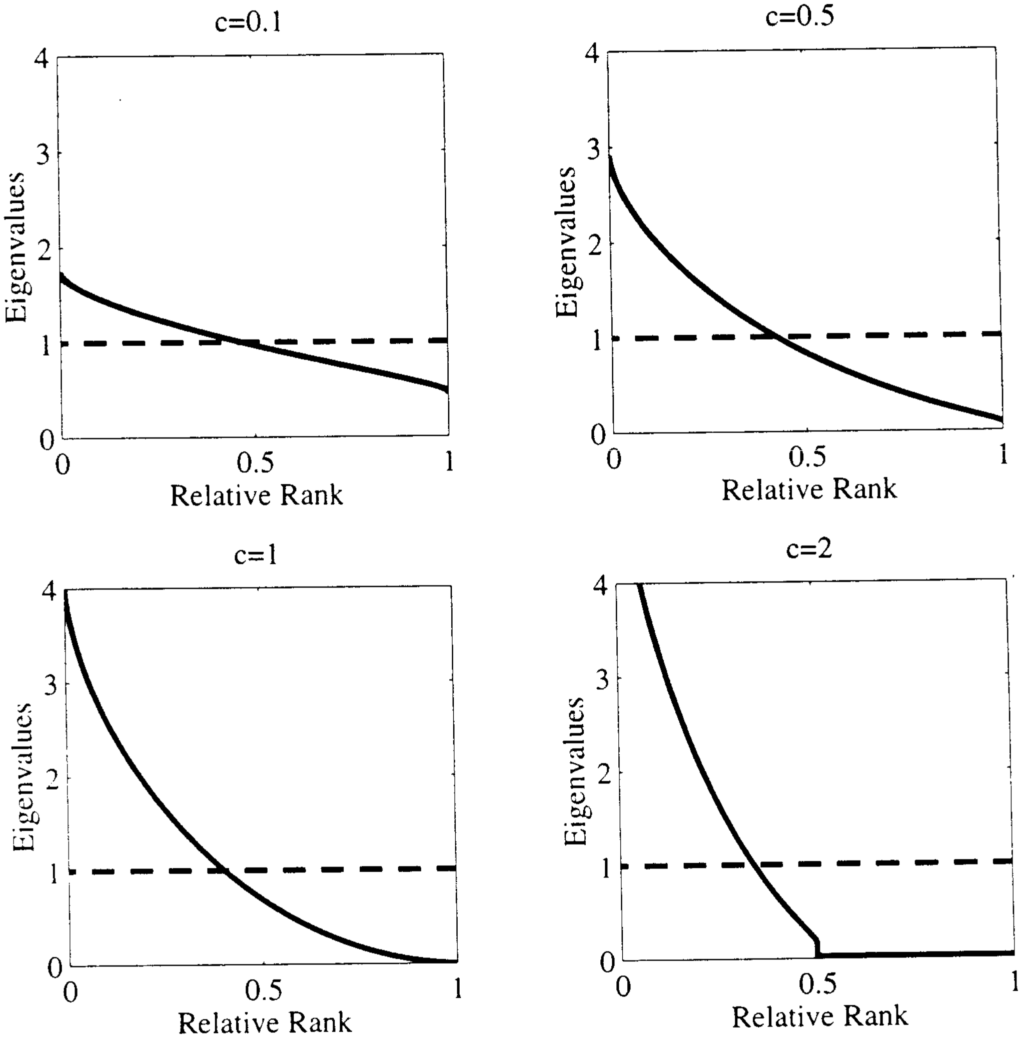
Table 4: Predictive OLS Cross-Sectional Regression of Returns on Betas over 1936-1992. Data come from the Center for Research in Security Prices (CRSP) database. Slope estimates are quoted in percents on an annual basis. Returns are in excess of the riskfree rate. The universe for a given year includes all common stocks traded on the NYSE and (after 1963) AMEX, with all valid monthly returns over the past 10 years and valid market capitalization. Returns are buy-and-hold, with annual rebalancing.

	Plain Regression	Excluding January	Including Size	1963-1992
Slope	3.51	3.08	2.57	3.08
Standard Error	(1.84)	(1.90)	(1.78)	(2.66)
T-Statistic	1.91	1.62	1.44	1.16

Table 5: Predictive GLS Cross-Sectional Regression of Returns on Betas over 1936-1992. Data come from the Center for Research in Security Prices (CRSP) database. Slope estimates are quoted in percents on an annual basis. Returns are in excess of the riskfree rate. The universe for a given year includes all common stocks traded on the NYSE and (after 1963) AMEX, with all valid monthly returns over the past 10 years and valid market capitalization. Returns are buy-and-hold, with annual rebalancing. The covariance matrix estimate required for GLS is obtained from the asymptotic shrinkage estimator associated with the structured estimator from Appendix B.4 (single index model).

	Plain Regression	Excluding January	Including Size	1963-1992
Section 3.2	2.58 (1.82) 1.42	2.23 (1.88) 1.19	1.14 (1.73) 0.66	2.35 (2.56) 0.92
Appendix B.1	2.53 (1.81) 1.40	2.06 (1.86) 1.11	1.14 (1.71) 0.66	2.22 (2.55) 0.87
Appendix B.2	3.61 (1.82) 1.98	3.44 (1.88) 1.83	2.65 (1.80) 1.47	3.01 (2.63) 1.14
Appendix B.3	3.39 (1.77) 1.92	4.56 (1.80) 2.53	3.42 (1.71) 2.00	3.85 (2.45) 1.57

Table 6: Predictive GLS Cross-Sectional Regression of Returns on Betas over 1936-1992. Data come from the Center for Research in Security Prices (CRSP) database. In each cell, the first number is the slope estimates are quoted in percents on an annual basis; the second number (in parenthesis) is the standard error on this number; and the third number is the t-statistic obtained by dividing the above two numbers. Returns are in excess of the riskfree rate. The universe for a given year includes all common stocks traded on the NYSE and (after 1963) AMEX, with all valid monthly returns over the past 10 years and valid market capitalization. Returns are buy-and-hold, with annual rebalancing. The covariance matrix estimates required for GLS is obtained from the asymptotic shrinkage estimator associated with the structured estimator from Section 3.2, and Appendices B.1, B.2 and B.3 respectively.



c=0.1

c=0.5

c=1

c=2

Figure 1: Sample vs. True Eigenvalues. The solid line represents the distribution of the eigenvalues of the sample covariance matrix. Eigenvalues are sorted in descending order, then plotted against their relative rank, defined as the ratio of the rank to the total number of eigenvalues N. When N changes, the relative rank remains between zero (largest eigenvalues) and one (smallest). We assume that the true covariance matrix is the identity, i.e. true eigenvalues are equal to one. The distribution of true eigenvalues is plotted as the dashed horizontal line. Distributions are obtained in the limit as the number of observations T and the number of variables N both go to infinity, with their ratio N/T converging to a finite positive limit c called the concentration. The four plots correspond to different concentrations. Marčenko and Pastur (1967) give an explicit formula for the asymptotic distribution. The smallest eigenvalues of the sample covariance matrix are severely biased downwards and the largest ones upwards. Bias increases in the concentration.

Geometric Interpretation of Theorem 5

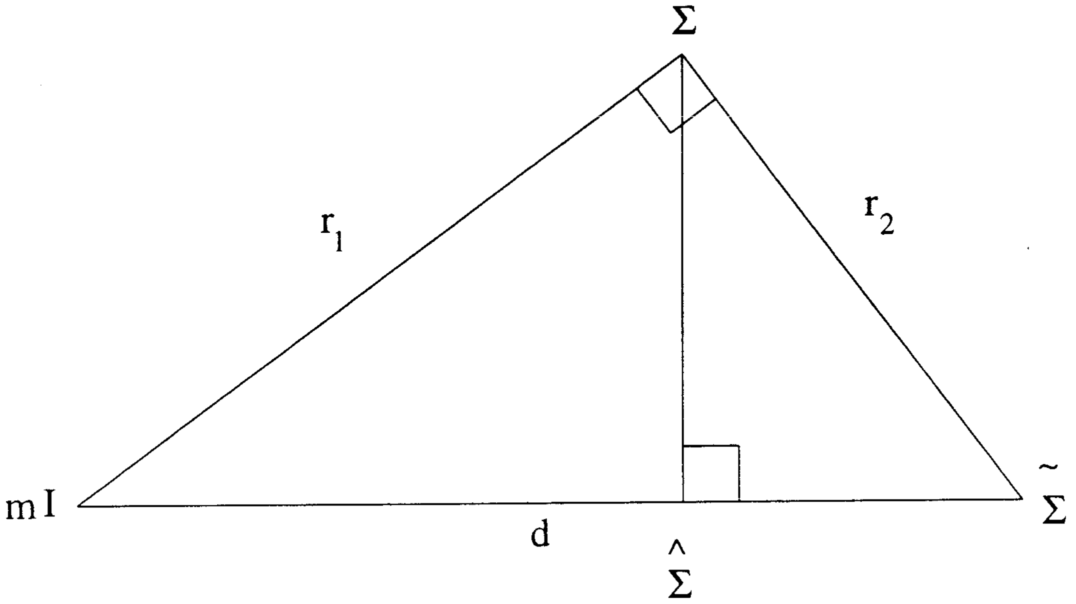


Figure 2: Geometric Interpretation of Theorem 5. Σ is the true covariance matrix, mI the scalar multiple of the identity closest to Σ, and $\tilde{\Sigma}$ the sample covariance matrix. r_1, r_2 and d denote the distances between these three matrices (see Theorem 5). The errors on mI and $\tilde{\Sigma}$ are orthogonal by Theorem 3. $\hat{\Sigma}$ is the weighted average of mI and $\tilde{\Sigma}$ with minimum mean squared error. It is the orthogonal projection of Σ onto the line between mI and $\tilde{\Sigma}$.

Bayesian Interpretation

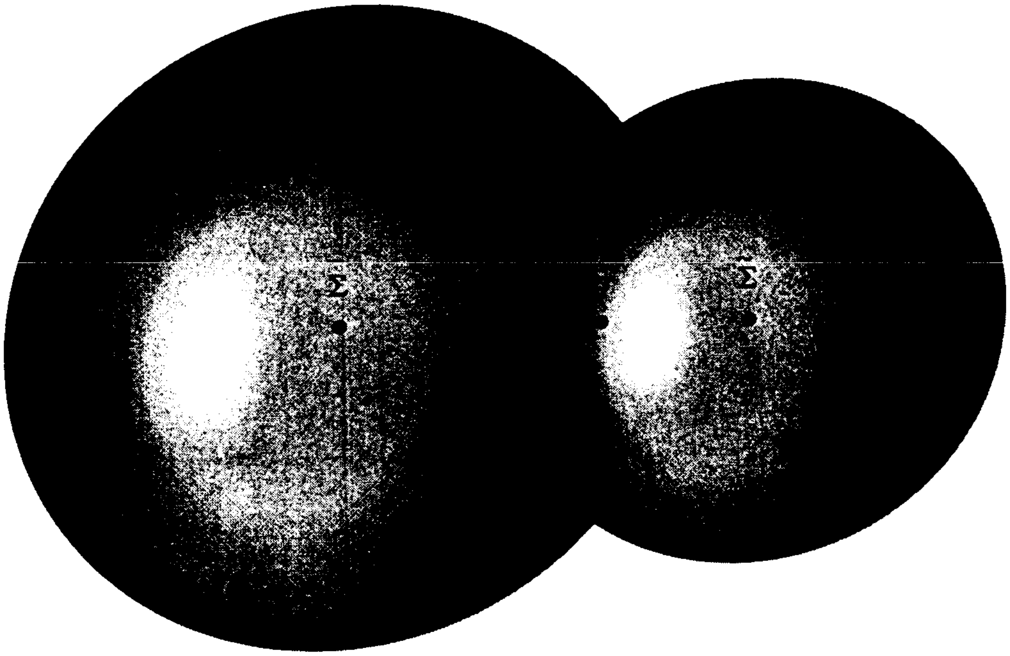


Figure 3: Bayesian Interpretation. The left sphere has center $\bar{\Sigma}$ and radius $\hat{r}_1$. The right sphere has center $\tilde{\Sigma}$ and radius $\hat{r}_2$. The distance between sphere centers is $\hat{d}$. If all we knew was that the true covariance matrix Σ lies on the left sphere, our best guess would be its center: the structured estimator $\bar{\Sigma}$. 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 $\tilde{\Sigma}$. 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 improved estimator $\hat{\hat{\Sigma}}$.

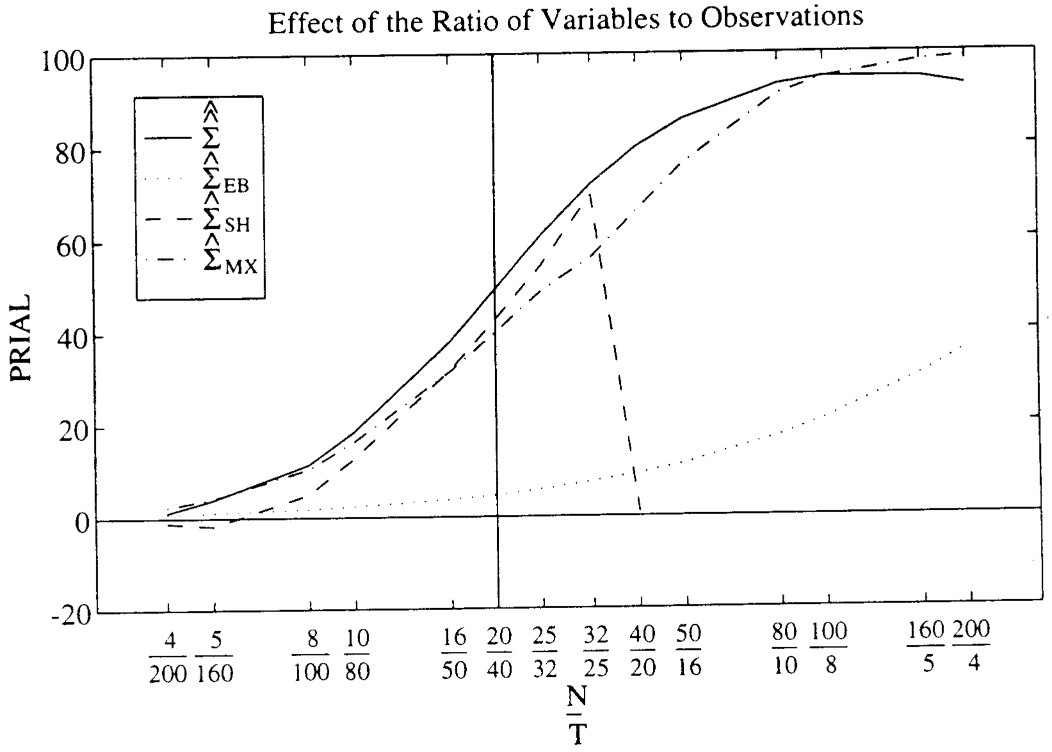


Figure 4: Effect of the Ratio of Number of Variables to Number of Observations on the Percentage Relative Improvement in Average Loss (PRIAL). Estimators and parameters are described in Section 4.1. Based on 1,000 Monte-Carlo simulations. $\hat{\Sigma}_{SH}$ is not defined when $N/T > 2$ because the isotonic regression does not converge.

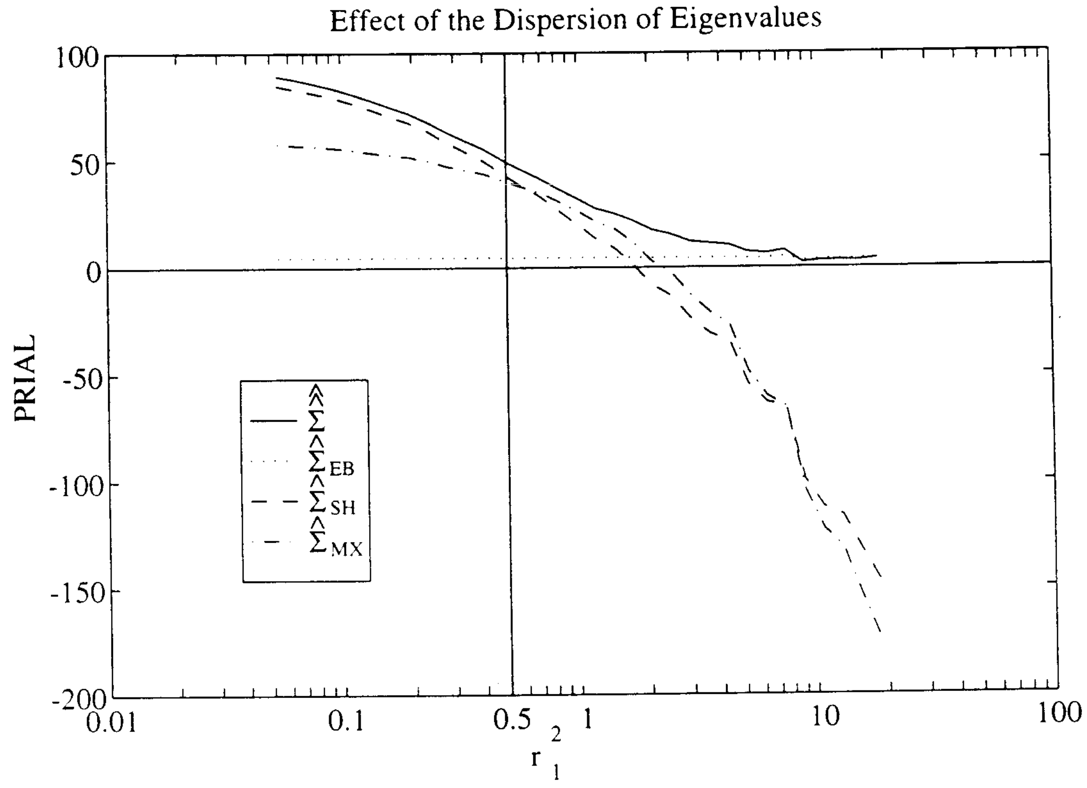


Figure 5: Effect of the Dispersion of Eigenvalues on the Percentage Relative Improvement in Average Loss (PRIAL). Estimators and parameters are described in Section 4.1. Based on 1,000 Monte-Carlo simulations.

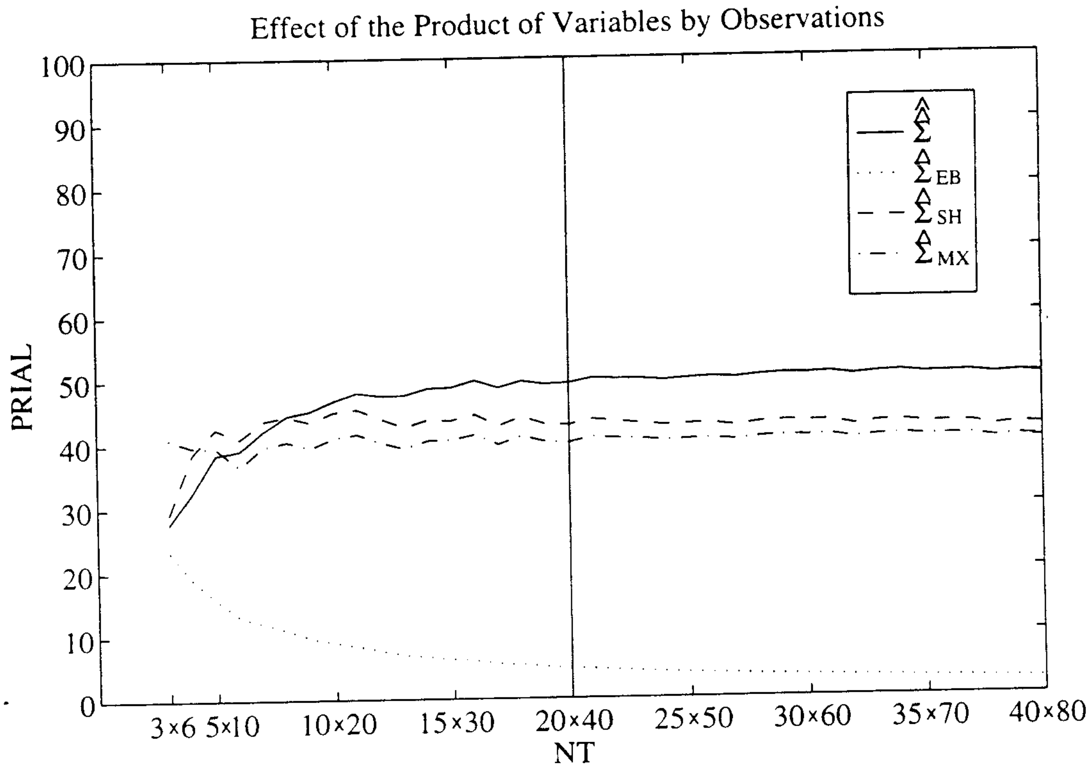


Figure 6: Effect of the Product of Variables by Observations on the Percentage Relative Improvement in Average Loss (PRIAL). Estimators and parameters are described in Section 4.1. Based on 1,000 Monte-Carlo simulations.

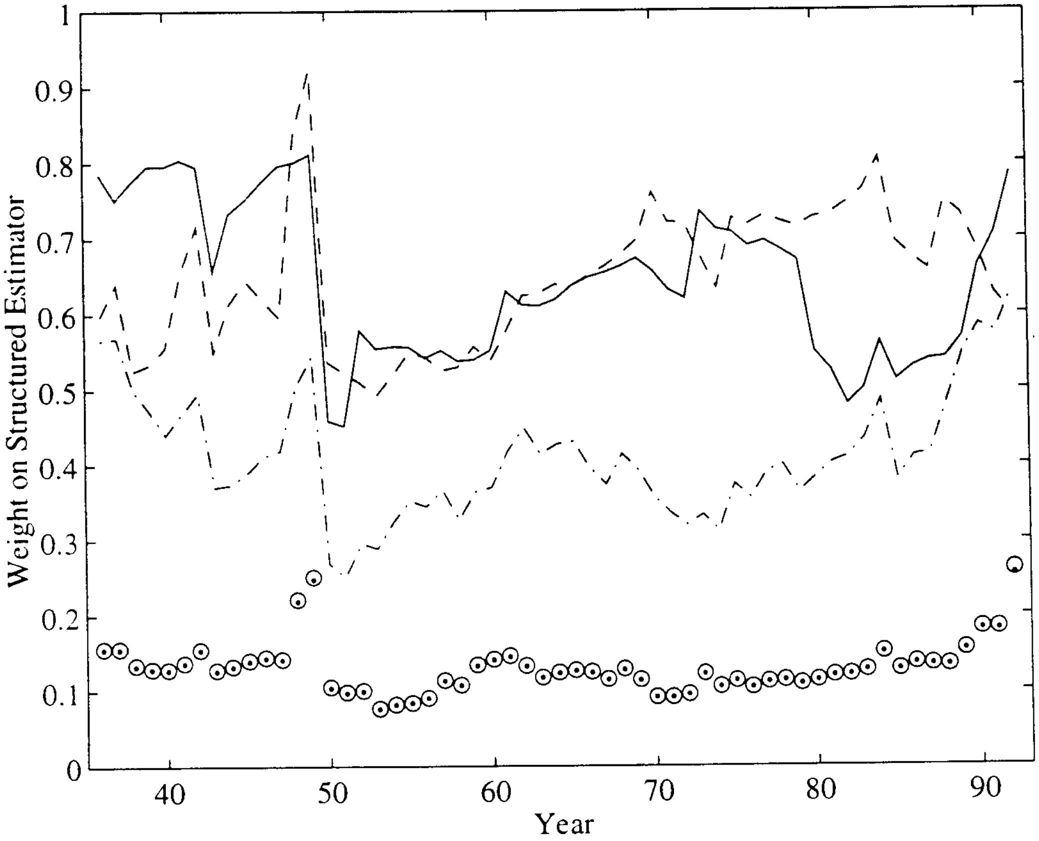


Figure 7: Weights on Structured Estimators. These weights are equal to $(\widehat{r}_2^2 - \widehat{\varphi})/\widehat{d^2}$, see Theorem 10. Dots correspond to the structured estimator $\overline{\Sigma} = \widehat{m}I$; circles, to the structured estimator of Appendix B.2; the dashed-dotted line, to Appendix B.1; the dashed line, to Appendix B.3; and the solid line, to Appendix B.4.

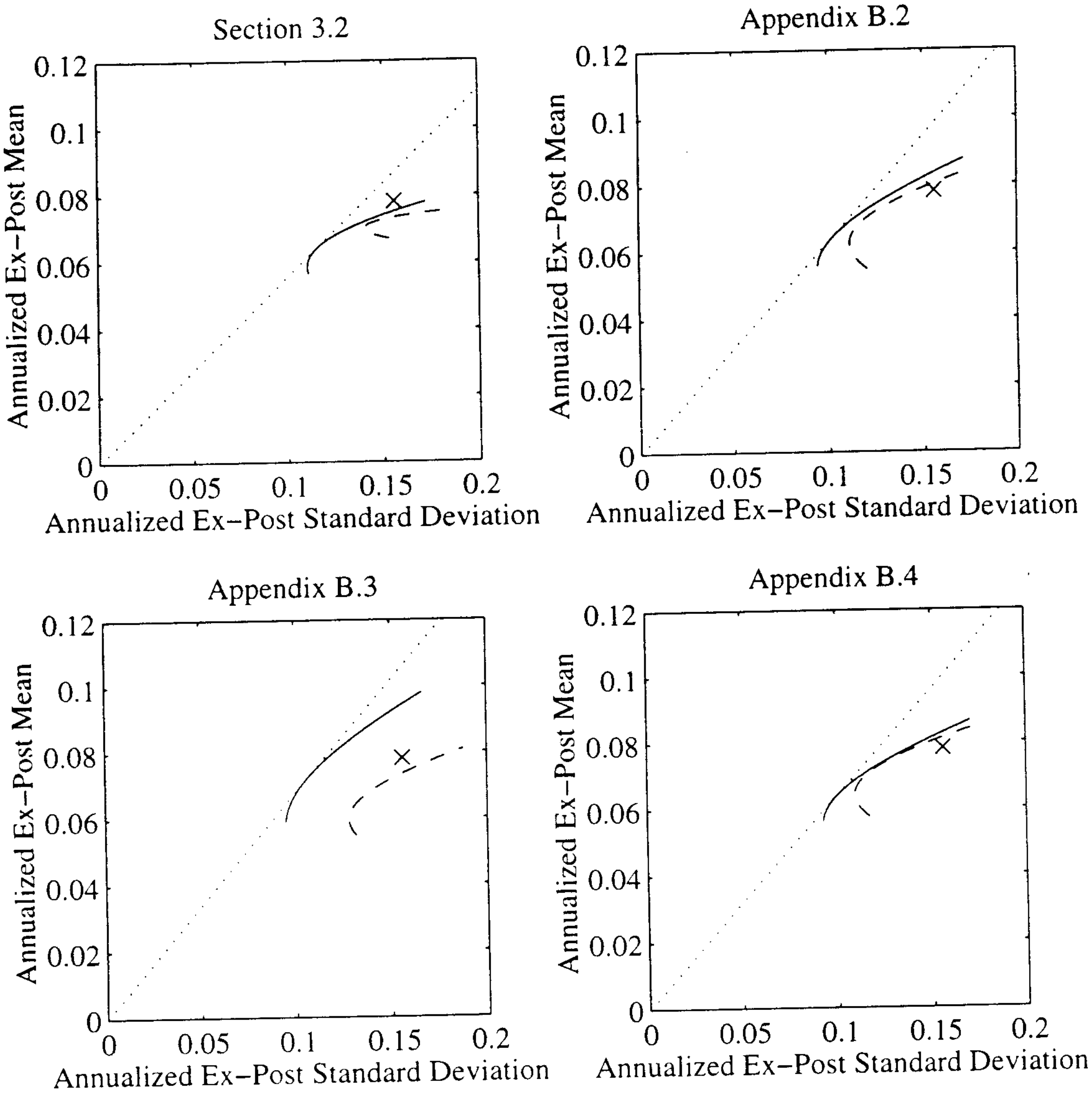


Figure 8: Ex-Post Characteristics of Ex-Ante Constrained Minimum Variance Portfolios. Portfolios are constrained to have a specified beta between zero and one, and size zero. On each graph, portfolios obtained from a structured estimator are plotted as a dashed line, together with portfolios from the corresponding shrinkage estimator as a solid line. The title of each graph gives the section where the structured estimator is described. In the interest of space, the graph corresponding to Appendix B.1 is not shown. It closely resembles the one corresponding to Section 3.2. The symbol × represents the CRSP value-weighted index, for reference. Shrinkage improves the risk-return tradeoff, moderately for the graphs on the left, and very slightly for the ones on the right. The interpretation is that the structured estimators from the graphs on the left are not very suitable for portfolio selection.

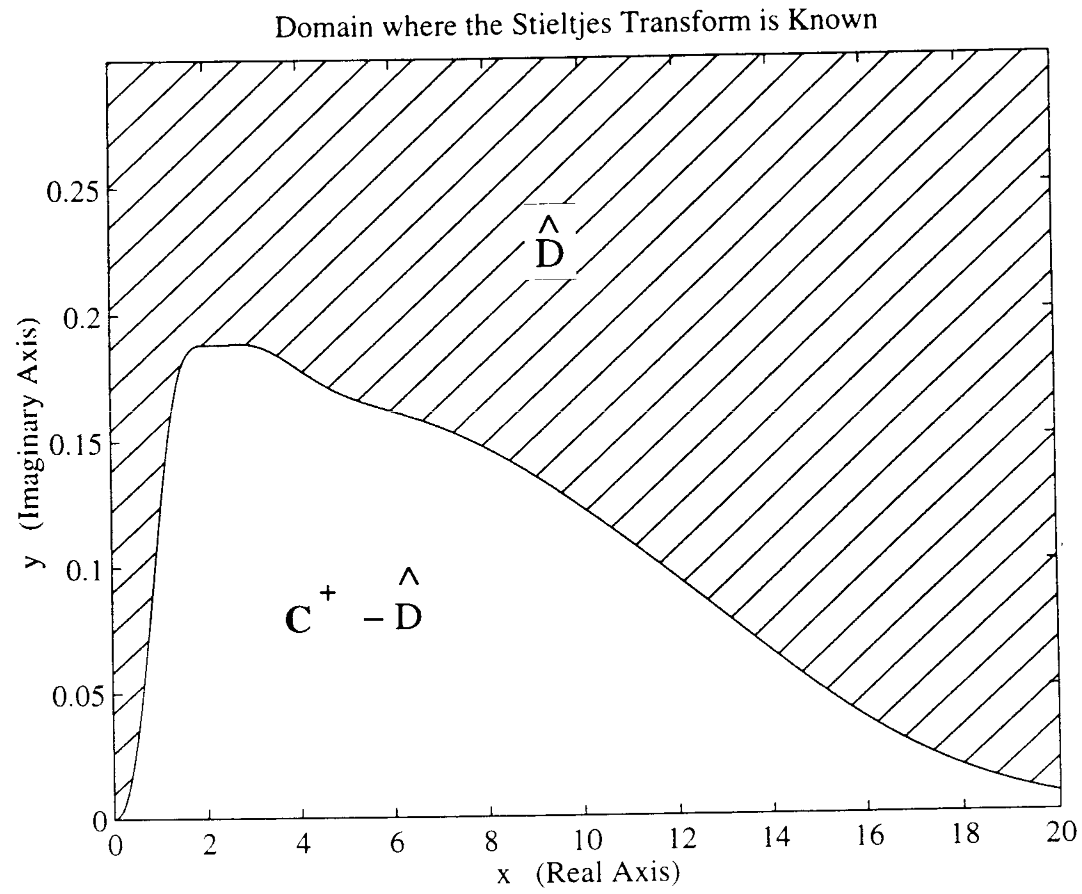


Figure 9: Domain where the Value of $s_{L\widehat{H}}$ is Known from Equation (17). The hatched domain represents a typical domain $\widehat{D}$, cf. Appendix A. $\widehat{D}$ is the domain where an estimate $s_{L\widehat{H}}$ of the Stieltjes transform of the true spectral c.d.f. H is known from Equation (17). The value of $s_{L\widehat{H}}$ is not shown in this figure. The Stieltjes inversion formula ties the density $h(x)$ of true eigenvalues to the imaginary part of $s_{L\widehat{H}}(x+i\varepsilon)$ for small $\varepsilon > 0$. Therefore we must extend Im$[s_{L\widehat{H}}]$ from the hatched domain $\widehat{D}$ towards the real line. It means solving a Laplace equation with free boundary. This is an ill-posed problem. The degree of ill-posedness is proportional to how far the hatched domain is from the real line. In this simulation, ill-posedness is less severe around large eigenvalues (large x) than small ones (small x). This figure is generated from $T = 1000$ observations on $N = 100$ variables. The true spectral c.d.f. is the standard lognormal distribution. It has many small, clustered eigenvalues and a few large, more isolated ones. This is the same general shape as the eigenvalues of the covariance matrix of the returns on all stocks traded in the stock market.