ABSTRACT

Title of dissertation: ESTIMATION OF A FUNCTION OF A LARGE COVARIANCE MATRIX USING CLASSICAL AND BAYESIAN METHODS

Judith Law, Doctor of Philosophy, 2018

Dissertation directed by: Professor Partha Lahiri
Department of Mathematics

In this dissertation, we consider the problem of estimating a high dimensional covariance matrix in the presence of small sample size. The proposed Bayesian solution is general and can be applied to different functions of the covariance matrix in a wide range of scientific applications, though we narrowly focus on a specific application of allocation of assets in a portfolio where the function is vector-valued with components which sum to unity. While often there exists a high dimension of time series data, in practice only a shorter length is tenable, to avoid violating the critical assumption of equal covariance matrix of investment returns over the period.

Using Monte Carlo simulations and real data analysis, we show that for small sample size, allocation estimates based on the sample covariance matrix can perform poorly in terms of the traditional measures used to evaluate an allocation for portfolio analysis. When the sample size is less than the dimension of the covariance matrix, we encounter difficulty computing the allocation estimates because of

singularity of the sample covariance matrix. We evaluate a few classical estimators. Among them, the allocation estimator based on the well-known POET estimator is developed using a factor model. While our simulation and data analysis illustrate the good behavior of POET for large sample size (consistent with the asymptotic theory), our study indicates that it does not perform well in small samples when compared to our proposed Bayesian estimator. A constrained Bayes estimator of the allocation vector is proposed that is the best in terms of the posterior risk under a given prior among all estimators that satisfy the constraint. In this sense, it is better than all classical plug-in estimators, including POET and the proposed Bayesian estimator. We compare the proposed Bayesian method with the constrained Bayes using the traditional evaluation measures used in portfolio analysis and find that they show similar behavior. In addition to point estimation, the proposed Bayesian approach yields a straightforward measure of uncertainty of the estimate and allows construction of credible intervals for a wide range of parameters.

ESTIMATION OF A FUNCTION OF A LARGE COVARIANCE MATRIX USING CLASSICAL AND BAYESIAN METHODS

by

Judith Law

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Partha Lahiri, Chair/Advisor
Professor Abram Kagan
Professor Ben Kedem
Professor Yan Li
Professor Paul Smith

Acknowledgments

I want to thank my advisor, Professor Partha Lahiri, for this opportunity to work on a challenging problem and his assistance throughout the process. I also want to thank Dr. Paul Smith and Dr. Abram Kagan for invaluable guidance over the past few years.

I appreciate the help of the wonderful staff and everyone associated with the Mathematics department and the AMSC program. It has been a joy to work with you and I have enjoyed my time here and appreciate all of your help and support.

I owe my deepest thanks to my family, especially my mother, father, and little brother who have always stood by me and guided me.

Contents

Acknowledgements ii

Table of Contents iii

List of Figures v

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Outline of Thesis . . . 7

2 Two Classical Methods of Covariance Estimation 9
- 2.1 Thresholding Principal Orthogonal Complements . . . 9
  - 2.1.1 The Relationship Between PCA and High-Dimensional Factor Analysis . . . 12
  - 2.1.2 The POET Method . . . 14
  - 2.1.3 Weighted Quadratic Norm . . . 18
- 2.2 Constant Correlation Shrinkage Estimator . . . 19
  - 2.2.1 Estimation of the Optimal Shrinkage Intensity . . . 21
  - 2.2.2 Notes on the Method . . . 24

3 Allocation Using A Bayesian Method 26
- 3.1 Bayesian Exploratory Factor Model . . . 26
  - 3.1.1 Identification . . . 28
  - 3.1.2 Prior Specification . . . 29
    - 3.1.2.1 Prior Distributions on the Indicator Matrix . . . 29
    - 3.1.2.2 Prior Distributions on the Idiosyncratic Variances and Factor Loadings . . . 31
    - 3.1.2.3 Prior Distributions on the Factor Correlation Matrix 34

4 Applications to Simulated and Real Data 39
- 4.1 Simulated Data Study . . . 39
  - 4.1.1 Data Generation . . . 40
  - 4.1.2 Required Specifications . . . 43
  - 4.1.3 Parameters of Interest . . . 47
    - 4.1.3.1 Covariance Matrix, $\mathbf{\Sigma}$ . . . 47
    - 4.1.3.2 Allocation Vector, θ . . . 48

4.1.4 Comparison of Methods 53
4.2 Real Data Example 62
4.2.1 Returns on Portfolio From Standard & Poor's 500 Index . . . 62
4.2.2 Comparison of Estimates 63

5 Conclusion 69

Bibliography 73

List of Figures

4.1 V_1 With Six Methods 55
4.2 V_1 With Four Methods 56
4.3 V_2 With Six Methods 57
4.4 M_1 With Three Methods 57
4.5 M_2 Showing POET and B.POST 58
4.6 M_3 Showing POET and B.POST 59
4.7 Credible Regions for $\hat{\theta}^U$ Showing 100 Variables 60
4.8 Credible Regions for $\hat{\theta}^U$ Showing 50 Variables 61
4.9 Pie Chart of Periods When Method Was Superior 64
4.10 Box Plot of Four Methods 65
4.11 POET and B.POST Comparison 66
4.12 SFM and B.POST Comparison 67
4.13 Using Four Values of K 68

Chapter 1: Introduction

1.1 Background

Multivariate data arise when researchers measure several variables on each unit in the sample. Researchers in many disciplines collect data sets which are multivariate. For observed multivariate data, we will use the notation y_{mn} to indicate the particular value of the m-th variable that is observed on the n-th trial. We display N measurements on M variables as the $(M \times N)$ matrix $\boldsymbol{Y}$, where $\boldsymbol{Y} = (\mathrm{y}_1, \ldots, \mathrm{y}_N)$, and y_n, $n = 1, \ldots, N$, is an $(M \times 1)$ vector. Each column of the matrix is a multivariate observation of M variables.

When the set of measurements is one particular realization of what might have been observed, we say the data are a *sample* of size N from an M-variate population. Three basic descriptive statistics of the observed sample are the sample mean, the sample variance-covariance matrix and the sample correlation matrix.

We briefly review the concept of random samples in the multivariate context. To study the sampling variability of statistics such as the basic descriptive statistics with the ultimate aim of making inferences, we need to make assumptions about the variables whose observed values constitute the data set $\boldsymbol{Y}$; see Johnson and Wichern (1992), [22]. When we do not have the actual measurements, but intend to collect a

set of N measurements, we can treat the column vectors as random variables. The vectors $(\mathrm{Y}_1, \ldots, \mathrm{Y}_N)$ form a random sample if their joint density function is given by the product $f(\mathrm{y}_1)f(\mathrm{y}_2)\cdots f(\mathrm{y}_N)$, where $f(\mathrm{y}_n) = f(y_{1n}, y_{2n}, \ldots, y_{Mn})$ is the density function for the n-th column vector. Furthermore, in the case that the $f(\mathrm{y}_n)$ are the same for $n = 1, \ldots, N$, then $(\mathrm{Y}_1, \ldots, \mathrm{Y}_N)$ form a random sample and represent independent observations from a common joint distribution with density function $f(\mathrm{y})$.

If we assume the common joint distribution has mean vector μ and covariance matrix $\mathbf{\Sigma}$, then we can examine how the sample mean and covariance matrix estimators fare as point estimators of the population mean vector μ and covariance matrix $\mathbf{\Sigma}$. Covariance matrix estimation is fundamental in multivariate analysis. Many applications require an estimation of a covariance matrix when the number of dimensions M, is large compared to the number of observations N, including work with genetics, large scale hypothesis testing, validation of asset pricing theory, and risk management. Often it is the inverse of the covariance matrix that plays a central role. For example, in risk management and portfolio allocation, a covariance estimator is needed that does not excessively amplify the estimation error upon inversion. Given a portfolio of M assets, with covariance $\mathbf{\Sigma}$, the allocation vector w of the global minimum variance portfolio can be written in matrix form as

$$\mathrm{w} = \frac{\mathbf{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{1}}$$

where $\mathbf{1}$ denotes an $(M \times 1)$ column of ones. Clearly, a reliable estimate of $\mathbf{\Sigma}^{-1}$ is essential and noise in the estimate can lead to large errors in the estimation of the

risk of a portfolio.

It can be shown that with data sets where $M > N$ or with data sets where $M < N$, but M/N is larger than approximately 0.2, the sample covariance matrix is often an inappropriate estimator. In the case of $M > N$ it is singular and is not available for applications requiring the inverse. In the case of high M/N, the sample covariance can perform poorly and lead to invalid conclusions. When used to estimate asymptotic relative loss of the global minimum variance portolio, the loss of the estimator using the sample covariance matrix is relatively small up to $M/N = 0.2$ but thereafter, as $M/N \to 1$, it rises hyperbolically to infinity, see Bodnar *et al.* (2018) [6].

It is widely acknowledged that the sample correlation matrix works poorly in large dimensions. With reference to analyzing multivariate observations of stock returns, Ledoit and Wolf (2004) [24] noted that estimating the covariance matrix of stock returns has been problematic and the standard statistical method is to gather a history of past stock returns and compute the sample covariance matrix. However, when the number of stocks under consideration is large relative to the number of historical return observations available, the sample covariance matrix is estimated with a lot of error. The most extreme coefficients in the matrix thus estimated tend to take on extreme values because they contain an extreme amount of error. The mean-variance optimization functions will rely more on those extreme values which are the most unreliable values in the sample covariance matrix. Michaud (1989), [27] famously described portfolio optimization as "error maximization."

Dempster (1972), [11] showed that errors in the estimation of the covariance

matrix had large consequences on the inverse of the covariance matrix and questioned whether the sample covariance should ever be used.

In portfolio allocation and risk management, it is not unusual to have the number of stocks M, be on the same order as the sample size. For example, when analyzing 200 stocks over a three year period, using daily data, then $M = 200$ and $N = 750$, and the concentration ratio (M/N) is .27.

Factor models are widely used in economics and finance. If a few factors can completely capture the cross-sectional risks, the number of parameters to be estimated can be significantly reduced. For example in the Fama-French three-factor model, in Fama and French (1993) [13], the number of parameters to be estimated in the case with 200 stocks and 3 years of daily data drops from 20,100 to 800. The benefits of using the factor model over the sample covariance matrix, to estimate the covariance model, in the case of increasing dimension of both M and N and also the number of factors K was demonstrated in Fan *et al.* (2008) [14]. The authors showed that under certain assumptions, the advantage of the factor model lies in the estimation of the inverse of the covariance matrix. In fact, there is not much advantage to using the factor model to estimate the covariance matrix for computations that do not involve its inverse. When the problem does not involve the inverse of the covariance matrix, the two estimators behave roughly the same.

The approach by Markowitz in portfolio selection and optimization has had a profound impact on multivariate analysis and covariance estimation; see Markowitz (1952) [26]. As documented by Fan *et al.* (2012) [17], the Markowitz portfolio is very sensitive to errors in the estimates of the inputs, namely the expected return

and the covariance matrix. One of the problems is the computational difficulty associated with solving a large-scale quadratic optimization problem with a dense covariance matrix. A concern is that the allocation vector which hinges on the estimated covariance matrix can be very different from the theoretical allocation vector.

The sample covariance matrix has appealing properties, such as being maximum likelihood under normality. However, it is a general drawback of maximum likelihood that it can perform poorly in small sample. For the covariance matrix, small sample problems occur unless N is at least one order of magnitude larger than M; see Ledoit and Wolf (2003) [23].

The main goal of our dissertation is to introduce a Bayesian method for estimating any function of the covariance of observed multivariate random variables which are generated by a factor model, exact or approximate. Conti *et al.* (2014) [9] developed a Bayesian exploratory factor model and demonstrated the advantages of the model in dimensionality selection and identification of the true latent structure. The method performs dedicated factor analysis with a stochastic search on the structure of the factor loading matrix and simultaneously estimates the number of factors, the allocation of each measurement to a unique factor, the corresponding factor loadings, and the idiosyncratic variances.

We extend and adapt the model in order to estimate the posterior distribution of a parameter of interest, which is a function of the covariance matrix. We consider the case when the underlying factor model is exact and the case when the underlying factor model is approximate. With the Bayesian approach, we construct credible

regions quantifying uncertainty associated with a vector estimate.

We compare the Bayesian method to two classical methods which are applicable in the case of high dimension.

1) Constant Correlation Composite (CCC).

This approach uses a composite of the sample covariance estimator and the average correlations with a weighting parameter that minimizes the expected distance between the shrinkage estimator and the true covariance matrix. The method of Ledoit & Wolf, (2003) [23] seeks to reduce estimation error in the sample covariance matrix by reducing the most extreme correlations found in the sample. This method does not rely on the assumption of a factor structure.

2) Thresholding Principal Orthogonal Complements (POET).

In the case that the observed data are generated by a factor model, the common factor structure is incorporated to estimate the covariance matrix. Further, covariance matrix estimation is extended to the case when the covariance matrix of idiosyncratic components is not diagonal. There may exist significant cross-sectional correlations even after removing correlations explained by common factors. If the covariance matrix of the idiosyncratic components is sparse, the covariance matrix is estimated by adapting principal component analysis and applying a thresholding technique to the idiosyncratic covariance matrix, as presented in Fan *et al.* (2013) [15].

1.2 Outline of Thesis

In Chapter 2, we present and synthesize the two methods referred to above, the Constant Correlation Composite as presented in Ledoit & Wolf, (2003) [23] and the method of Thresholding Principal Orthogonal Complements as presented in Fan *et al.* (2013) [15].

In Chapter 3, we present and synthesize a Bayesian Exploratory Factor Analysis which estimates the components of a factor model at each iteration of an MCMC sample as presented in Conti *et al.* (2014) [9]. The iterations explore distributions based on stipulated priors and starting values. The model estimates the number of factors simultaneously with the factor loadings, factor correlations and idiosyncratic variances.

In Chapter 4, we simulate daily returns on a large stock portfolio over a 300 day period based on a model calibrated from an actual portfolio. We consider the impact of varying M from 10 to 600, which includes the case of M greater than N. We test several measurements of distance from the true covariance matrix according to the various methods. In addition, we compare our Bayesian estimate to estimates from the sample covariance matrix and the strict factor model, which is a special case of the POET model.

We present a comparison of the methods on a set of stocks randomly selected from the S&P 500 Index, with daily returns spanning ten years. A test is conducted on a rolling panel of four months, with the earliest month dropped and another month added at each step, until the covariances for 108 periods are estimated. The

global minimum variance portfolio is found under each method. The actual risk of the portfolio in the month following the test periods is calculated. For each method, we determine the proportion of test periods when the method resulted in the lowest risk and evaluate the gain in decreased risk.

On the generated data, we observe scenarios in which the Bayesian method results in more reliable estimates. Using the real data, we found that the Bayesian method outperformed the others in many of the periods. Unlike the method using principal components, the Bayesian method has an explicit noise model and may have the advantage of being better able to distinguish between the underlying latent factors and the idiosyncratic components. Furthermore, utilizing the posterior distribution of the quantity of interest has advantages over a plug-in method where the estimate of the covariance is used in the function.

Chapter 5 provides the conclusions and suggestions for future research.

Chapter 2: Two Classical Methods of Covariance Estimation

2.1 Thresholding Principal Orthogonal Complements

In the case of a data set in which both the number of measurements M, and the size of the sample N, are large, one useful method for summarizing information is the factor model. Assuming a sparse error covariance matrix in an approximate factor model, Fan *et al.* (2013) [15] threshold the covariance matrix of the estimated residuals. The choice of thresholding methods are many; we use the adaptive thresholding as in Cai and Liu (2011) [7]. Fan *et al.* (2013) [15] show that the following factor model admits consistent estimation of the covariance matrix of the observed data set when both M and N diverge to infinity while K, the number of factors, is assumed fixed.

$$\mathrm{y}_i = \mathbf{B}\mathrm{f}_i + \mathrm{u}_i, \qquad i = 1, \ldots, N \tag{2.1}$$

where $\mathrm{y}_i = (y_{1i}, \ldots, y_{Mi})'$,

$\mathbf{B} = (\mathrm{b}_1, \ldots, \mathrm{b}_M)'$,

$\mathrm{b}_m = (b_{1m}, \ldots, b_{Km})'$

$\mathrm{f}_i = (f_{1i} \ldots, f_{Ki})'$

$\mathrm{u}_i = (u_{1i}, \ldots, u_{Mi})'$,

$\mathbf{B}$ is an $(M \times K)$ matrix of factor loadings, f_i is a vector of common factors, and u_i denotes the idiosyncratic component of the model. The only observable random variable in the model is y_i. It is possible to have M much larger than N.

We assume that f_i is uncorrelated with u_i. If $i = 1, \ldots, N$ represents points in time, the covariance matrix is assumed to be time invariant. The factors f_i may include scaled factors to cover latent time-varying factor loadings resulting in a covariance matrix $\mathbf{B}\mathrm{Cov}(\mathrm{f}_i)\mathbf{B}'$ that is constant over time. We can write the $(M \times M)$ covariance matrix of y_i as:

$$\mathbf{\Sigma} = \mathbf{B}\mathrm{Cov}(\mathrm{f}_i)\mathbf{B}' + \mathbf{\Sigma}_{\mathrm{u}} \tag{2.2}$$

where $\mathbf{\Sigma}_{\mathrm{u}}$ is the covariance matrix of u_i.

As was shown by Chamberlain and Rothschild (1983), [8], in many applications of factor analysis it is desirable to allow cross-sectional dependence among the error terms. This gives rise to the approximate factor model, in which the covariance matrix $\mathbf{\Sigma}_{\mathrm{u}}$ is not diagonal. In addition, the diagonal entries may vary in a large range. As a result, efficiently estimating the factor model under both large M and large N must take into account both cross-sectional heteroskedasticity and cross-sectional dependence of the error terms u_{mi}.

An additional two constraints are placed on the model. First, we assume the factors are pervasive in the sense that a non-negligible fraction of factor loadings are non-vanishing as $M \to \infty$. The first K eigenvalues of $\mathbf{\Sigma}$ are assumed to diverge with respect to an increasing M; they are bounded away from both zero and infinity as $M \to \infty$. This allows us to identify the decomposition shown in Eq. (2.2)

asymptotically as $M \to \infty$.

Such an assumption might be justified in the case of investment portfolios consisting of a large number of stocks from an economic environment. For example, in the capital asset-pricing model (CAPM) of Sharpe (1964) [32], the market return is a common pervasive factor. There is a growing body of empirical evidence that stock returns are related to factors based on macroeconomic, market and firm-level characteristics which may indicate the existence of pervasive factors common to a large pool of investment alternatives. Another application is regression models with panel data, when the errors are assumed to be independent of known covariates but dependent on some unobservable factors. Generalized least squares can be used to produce consistent estimators of the coefficients; however, it relies on estimation of the inverse of the covariance matrix, which may be better estimated using the POET method.

Next, we assume that $\mathbf{\Sigma}_{\mathrm{u}}$ is approximately sparse as in Bickel and Levina (2008) [5]. Defining the spectral norm and the Frobenius norm of a matrix $\mathbf{A}$ as $\|\mathbf{A}\| = \lambda_{\max}^{1/2}(\mathbf{A}'\mathbf{A})$, and $\|\mathbf{A}\|_F = \mathrm{tr}^{1/2}(\mathbf{A}'\mathbf{A})$ respectively, and where $\lambda_{\max}$ denotes the maximum eigenvalue of the matrix $\mathbf{A}$, consider $\mathbf{\Sigma}_{\mathrm{u}} = (\sigma_{\mathrm{u},jk})_{M\times M}$, and $q \in [0, 1)$. Define m_M as:

$$m_M = \max_{j\leq M} \sum_{k\leq M} |\sigma_{\mathrm{u},jk}|^q. \tag{2.3}$$

The spectral norm of of $\mathbf{\Sigma}_{\mathrm{u}}$ will grow no faster than the rate of m_M,

$$\|\mathbf{\Sigma}_{\mathrm{u}}\| \leq \max_j \sum_{k=1} |\sigma_{\mathrm{u},jk}|^q (\sigma_{\mathrm{u},jj}\sigma_{\mathrm{u},kk})^{(1-q)/2} = \mathcal{O}(m_M) \tag{2.4}$$

Fan *et al.* restrict the growth of m_M as $M \to \infty$ to $o(M)$ which ensures that

$\|\mathbf{\Sigma}_{\mathrm{u}}\| = o(M)$. The intuition is that conditional on the common factors, many pairs of the cross-sectional units become weakly correlated, and only a few (possibly large) correlations remain. Thus, all of the eigenvalues of $\mathbf{\Sigma}_{\mathrm{u}}$ are bounded as $M \to \infty$.

2.1.1 The Relationship Between PCA and High-Dimensional Factor Analysis

Let $\{\lambda_j(\mathbf{B}\mathrm{Cov}(\mathrm{f}_i)\mathbf{B}')\}_{j=1}^{K}$ be the eigenvalues of $\mathbf{B}\mathrm{Cov}(\mathrm{f}_i)\mathbf{B}', i = (1, \ldots, N)$, in non-decreasing order.

We label the assumptions as follows:

Assumptions of POET

A. Both M and $N \to \infty$, while K is fixed

B. $\mathrm{Cov}(\mathrm{f}_i,\, \mathrm{u}_i) = 0, \quad i = 1, \ldots, N$

C. $\mathrm{Cov}(\mathrm{u}_{mi},\, \mathrm{u}_{ml}) = 0, \quad i \neq l, \quad i, l = 1, \ldots, N, \; m = 1, \ldots, M$

D. $\lambda_j(\mathbf{B}\mathrm{Cov}(\mathrm{f}_i)\mathbf{B}') = \mathcal{O}(M), \quad j = 1, \ldots, K, \; i = 1, \ldots, N$

E. $\|\mathbf{\Sigma}_{\mathrm{u}}\| = o(M)$

Fan *et al.* (2013) [15] show that factor analysis under the model is approximately the same as principal component analysis for high-dimensional data. The factor loading matrix $\mathbf{B}$ spans a K-dimensional space. As M and N increase, the first K principal components of $\mathbf{\Sigma}$ are close to the normalized columns of $\mathbf{B}$.

In the factor model shown in Eq. (2.1), without further conditions, the decomposition of $\mathbf{\Sigma}$ as shown in Eq. (2.2) is not identifiable. Without loss of generality,

conditions are imposed requiring the columns of $\mathbf{B}$ to be orthogonal and $\mathrm{Cov}(\mathrm{f}_i)$ to equal $\mathbf{I}_K$. The columns of $\mathbf{B}$ are ordered such that the eigenvalues of $\mathbf{BB}'$ are in non-increasing order.

One can write the symmetric population covariance matrix $\mathbf{\Sigma} = \mathbf{VDV}'$, where $\mathbf{D}$ is the diagonal matrix composed of the non-zero eigenvalues of $\mathbf{\Sigma}$ in non-increasing order, and $\mathbf{V}$ is a matrix with columns equal to the independent eigenvectors of $\mathbf{\Sigma}$ in corresponding order, scaled to length 1. Since $\mathbf{D}$ is diagonal it is clear that the covariance matrix can be further decomposed, as:

$$\mathbf{\Sigma} \;=\; \mathbf{V}_1\mathbf{D}_1\mathbf{V}_1' + \mathbf{V}_2\mathbf{D}_2\mathbf{V}_2' \tag{2.5}$$

where $\mathbf{D}_1$ is a diagonal matrix composed of the first K eigenvalues of $\mathbf{\Sigma}$ in non-increasing order, and $\mathbf{V}_1$ is a matrix with columns equal to the first K columns of $\mathbf{V}$. The authors show that the columns of $\mathbf{V}_1$ can be used as a proxy of the space spanned by the columns of $\mathbf{B}$.

Fan *et al.* (2013) [15] show that for large M, the normalized columns of $\mathbf{B}$ are close to the first K principal components of $\mathbf{\Sigma}$. In fact, the procedure is only valid in the case when M is large, and it is inconsistent if M is bounded; see Bai (2003) [2]. Weyl's eigenvalue theorem in Weyl (1912) [36] can be used to show that under the assumptions, the gaps between the first K eigenvalues of $\mathbf{\Sigma}$ and the corresponding eigenvalues of $\mathbf{BB}'$ are no larger than $\|\mathbf{\Sigma}_{\mathrm{u}}\|$, and these eigenvalues are non-vanishing per Assumption D. The next $M - K$ eigenvalues of $\mathbf{\Sigma}$ are bounded by $\|\mathbf{\Sigma}_{\mathrm{u}}\|$. Using the Sin θ theorem of Davis and Kahan (1970) [10], it is clear that the gap between the columns of $\mathbf{V}_1$ and the normalized columns of $\mathbf{B}$ is bounded by $\|\mathbf{\Sigma}_{\mathrm{u}}\| \times c/M$

for some constant c for large enough M, and due to the conditional sparsity of $\mathbf{\Sigma}_{\mathrm{u}}$ (Assumption E), the gap is vanishing as $M \to \infty$. Therefore, the model in Eq. (2.13) consistently estimates the space spanned by the eigenvectors of the true population covariance $\mathbf{\Sigma}$.

This provides the spectral consistency at the population level $\mathbf{\Sigma}$ in the case of high dimension. In addition, with Assumption D and without Assumption E, Bai (2003) [2] has shown that principal component analysis on the sample covariance matrix can consistently estimate the space spanned by the true factor loadings.

2.1.2 The POET Method

The estimator which the authors termed the (POET) estimator, $\widehat{\mathbf{\Sigma}}_{\mathrm{POET}}$, (Principal Orthogonal complEment Thresholding) is simple, optimization-free and uses the data only through the sample covariance matrix. The steps are as follows:

1. Estimate K if it is unknown.

2. Decompose the sample covariance matrix using singular value decomposition.

3. Keep the covariance matrix formed by the first K principal components.

4. Apply the thresholding procedure to the remaining covariance matrix.

When K is unknown, we estimate K from the data, selecting the value which minimizes a measurement of loss with a penalty term. Any consistent estimator of K can be adopted. We use the method proposed by Bai and Ng (2002) [3], in which K is chosen to minimize a data-driven objective function with a penalty for increasing

K. Let $\boldsymbol{Y}$ be the $(M \times N)$ matrix of observed values,

$$\widehat{K} = \underset{0 \le K_1 \le K_{\max}}{\arg\min} \left[\log \left\{ \frac{1}{MN} \| \boldsymbol{Y} - \frac{1}{N} \boldsymbol{Y} \widehat{\boldsymbol{F}}_{K_1} \widehat{\boldsymbol{F}}_{K_1}' \|_F^2 \right\} + K_1 g(M,N) \right] \tag{2.6}$$

where $g(M,N) = \dfrac{M+N}{MN} \log \left(\dfrac{MN}{M+N} \right)$,

$K_{\max}$ is a prescribed upper bound, and

$\widehat{\boldsymbol{F}}_{K_1}$ is an $(N \times K_1)$ matrix whose columns are $\sqrt{N}$ times the eigenvectors

corresponding to the K_1 largest eigenvalues of $\boldsymbol{Y}'\boldsymbol{Y}$.

Using the estimated K, we decompose the sample covariance matrix using singular value decomposition. Let $\boldsymbol{Y}$ be the $(M \times N)$ matrix of observed values and let $\mathbf{1}$ be a vector of ones of length N.

$$\widehat{\boldsymbol{\Sigma}}_{\text{SAM}} = (N-1)^{-1} \boldsymbol{Y}\boldsymbol{Y}' - \{N(N-1)\}^{-1} \boldsymbol{Y}\mathbf{1}\mathbf{1}'\boldsymbol{Y}' \tag{2.7}$$

Then the sample covariance has the following spectral decomposition:

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_{\text{SAM}} &= \sum_{j=1}^{\hat{K}} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j' + \sum_{j=\hat{K}+1}^{M} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j' \\ &= \sum_{j=1}^{\hat{K}} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j' + \widehat{\mathbf{R}}_{\hat{K}} \end{aligned} \tag{2.8}$$

where $\hat{\lambda}_1, \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_M$ are the ordered eigenvalues of the sample covariance matrix and $\{\hat{\xi}_j\}_{j=1}^M$ are their corresponding eigenvectors.

Because $\boldsymbol{\Sigma}_{\text{u}}$ is assumed to be sparse (Assumption E), we apply thresholding on $\widehat{\mathbf{R}}_{\hat{K}}$ with elements $(\hat{r}_{jk})_{(M \times M)}$ and denote the result as $\widehat{\boldsymbol{\Sigma}}_{\text{u},\hat{K}}^{\mathcal{T}}$ with elements $(\hat{r}_{jk}^{\mathcal{T}})_{(M \times M)}$. The model admits any procedure that restricts the maximum sum of

every row of $\widehat{\mathbf{R}}_{\hat{K}}$ in terms of absolute values to a specified value m_M, which can depend on M, but must grow more slowly than M.

A simple thresholding method is hard thresholding over a constant thresholding parameter. Each off-diagonal element smaller than a cutoff ω is replaced with zero, and can be written as:

$$\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{\mathrm{u},\hat{K}} \equiv \{\hat{r}^{\mathcal{T}}_{jk}\} = \begin{cases} \hat{r}_{jj} & ;\ j = k \\ \hat{r}_{jk} 1\{|\hat{r}_{jk}| > \omega\} & ;\ j \neq k \end{cases} \tag{2.9}$$

An alternative to using a universal upper bound is to use an entry-dependent thresholding rule which adapts to the variability of the individual entries in the covariance matrix to be thresholded.

We can write a wide class of thresholding rules as follows:

$$\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{u,\hat{K}} \equiv \{\hat{r}^{\mathcal{T}}_{jk}\} = \begin{cases} \hat{r}_{jj} & ;\ j = k \\ s_{jk}(\hat{r}_{jk}) 1\{|\hat{r}_{jk}| > \omega_{jk}\} & ;\ j \neq k \end{cases} \tag{2.10}$$

where $s_{jk}(\cdot)$ is a generalized shrinkage function and ω_{jk} is possibly an entry-dependent threshold. Examples of shrinkage functions include hard thresholding (where small elements are eliminated and large elements are unchanged) and soft thresholding (where small elements are eliminated and large elements are reduced on an absolute basis by the thresholding parameter). Other shrinkage functions reduce the larger elements according to a ratio of the element to the thresholding parameter. In particular, this notation includes hard thresholding over a constant thresholding parameter as shown in Eq. (2.8), in which $s_{jk}(x) = x1\{|\hat{r}_{jk}| > \omega_{jk}\}$ and $\omega_{jk} = \omega$.

An example of adaptive thresholding is given by Cai and Liu (2011) [7] who show that an entry-dependent method is particularly beneficial when the diagonal

elements vary over a wide range or no upper bound is apparent. Based on their method, we use the entry-dependent thresholding parameter:

$$\omega_{jk} = C\sqrt{\hat{\theta}_{jk}}\left(\sqrt{\frac{\log M}{N}} + \sqrt{\frac{1}{M}}\right) \quad j,k \in (1,\ldots,M) \tag{2.11}$$

where $C > 0$ is a regularization parameter,

$$\hat{\theta}_{jk} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\mathrm{u}}_{ji}\hat{\mathrm{u}}_{ki} - \hat{r}_{jk})^2$$

$$\widehat{\mathrm{u}}_i = Y_i - \widehat{\mathbf{B}}_{\hat{K}}\widehat{\mathrm{f}}_{\hat{K},i} \quad i = 1,\ldots,N$$

and $s_{jk}(\cdot)$ is chosen to be the soft shrinkage function:

$$s_{jk}(x) \;=\; \mathrm{sgn}(x)(|x| - \omega_{jk})_+ \tag{2.12}$$

where $(x)_+ = 0$ for $x \leq 0$.

It is important to choose C large enough to maintain the positive definiteness of $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{\mathrm{u},\hat{K}}$. One option for choosing C is to take the minimum value that results in positive definiteness. Another method is to minimize a multifold cross validation function over partitions of the N observations, selecting the value for each partition which results in the minimum squared differences between the elements of the thresholded partition and the non-thresholded partition. The average of these values gives $\hat{C}$; however, it is then increased if necessary until $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{\mathrm{u},\hat{K}}$ is positive definite.

With this shrinkage function, every non-zero element is penalized. The penalty increases as $\hat{\theta}_{jk}$ increases, placing greater penalty on those estimates with higher deviations about the estimate. Intuitively, the estimates which are consistently significant across the N observations are more likely to indicate true cross-sectional

correlation. The result is

$$\widehat{\mathbf{\Sigma}}_{\text{POET}} \;=\; \sum_{j=1}^{\hat{K}} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j' + \widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{\text{u},\hat{K}}\,. \tag{2.13}$$

The POET estimator encompasses many popular estimators as specific cases. Under the hard or soft thresholding shrinkage functions, when $\omega_{jk} = 0$, then thresholding has no effect and $\widehat{\mathbf{\Sigma}}_{\text{POET}} = \widehat{\mathbf{\Sigma}}_{\text{SAM}}$. When ω_{jk} is high enough, the estimator becomes the strict factor model with diagonal $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{\text{u},\hat{K}}$.

2.1.3 Weighted Quadratic Norm

With both M and N increasing to infinity, in the presence of very spiked eigenvalues as under assumption D, the covariance matrix $\mathbf{\Sigma}$ cannot be consistently estimated. The spectral norm $\|\widehat{\mathbf{\Sigma}}_{\text{POET}} - \mathbf{\Sigma}\|$ can diverge when $N = \mathcal{O}(M^2)$, see Fan *et al.* (2013) [15]. With this measure, the POET method provides no gain over using the sample covariance matrix.

We can estimate the precision matrix $\mathbf{\Sigma}^{-1}$ with a satisfactory rate under several norms. Here we see large gains from using $\widehat{\mathbf{\Sigma}}_{\text{POET}}$ over $\widehat{\mathbf{\Sigma}}_{\text{SAM}}$.

The sample covariance inverse is defined only when $M < N$. However, consider the weighted quadratic norm introduced by Fan *et al.* (2008) [14]:

$$\|\boldsymbol{A}\|_{\Sigma} = \frac{1}{\sqrt{M}} \|\mathbf{\Sigma}^{-1/2} \boldsymbol{A} \mathbf{\Sigma}^{-1/2}\|_F \tag{2.14}$$

We can find the convergence rate under the weighted quadratic norm of both the $\widehat{\mathbf{\Sigma}}_{\text{SAM}}$ and $\widehat{\mathbf{\Sigma}}_{\text{POET}}$ as M and N increase to infinity. The weighted quadratic norm of $\widehat{\mathbf{\Sigma}}_{\text{SAM}}$ can be estimated, but will not converge unless $M = o(N)$.

The weighted quadratic norm of $\widehat{\mathbf{\Sigma}}_{\text{POET}}$ converges at a much faster rate and will converge as long as $M = o(N^2)$.

The benefit stems from the decomposition.

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_{\text{POET}} &= \sum_{j=1}^{\hat{K}} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j' + \widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{\text{u},\hat{K}} \\ &= \widehat{\mathbf{B}}_{\hat{K}} \widehat{\mathbf{B}}_{\hat{K}}' + \widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{\text{u},\hat{K}} \end{aligned} \tag{2.15}$$

Using the Sherman-Morrison-Woodbury formula in Sherman *et al.* (1949) [33], the inverse of $\widehat{\mathbf{\Sigma}}_{\text{POET}}$ can be written as:

$$\left(\widehat{\mathbf{\Sigma}}_{\text{POET}}\right)^{-1} = \left(\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{\text{u},\hat{K}}\right)^{-1} - \left(\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{\text{u},\hat{K}}\right)^{-1} \widehat{\mathbf{B}}_{\hat{K}} \left[\mathbf{I}_K + \widehat{\mathbf{B}}_{\hat{K}}' \left(\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{\text{u},\hat{K}}\right)^{-1} \widehat{\mathbf{B}}_{\hat{K}}\right]^{-1} \widehat{\mathbf{B}}_{\hat{K}}' \left(\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{\text{u},\hat{K}}\right)^{-1} \tag{2.16}$$

The inverse of POET relies on the inverse of a sparse $(M \times M)$ matrix and the inverse of a $(K \times K)$ matrix. In contrast, the inverse of the sample covariance requires the inverse of a dense $(M \times M)$ matrix. In the case of very large M and especially if the inverse of the covariance matrix is to be utilized, if prior information justifies the assumptions, then $\widehat{\mathbf{\Sigma}}_{\text{POET}}$ may be a good estimator.

2.2 Constant Correlation Shrinkage Estimator

A method of Ledoit & Wolf, 2004 [24] seeks to reduce estimation error in the sample covariance matrix by reducing the most extreme correlations found in the sample. The estimator is consistent under very weak assumptions; however, it is not necessarily unbiased. The method is computationally inexpensive, and is applicable when $N > M$ and in high dimensional settings with $M \geq N$. Based in part on the

work of Efron and Morris (1977), [12] a shrinkage technique is used, resulting in a composite of two sample statistics.

The optimal shrinkage intensity $\delta^* > 0$ is found as the minimizer of the risk function using the squared Frobenius norm to measure loss. The resulting estimate, $\widehat{\boldsymbol{\Sigma}}_{shr}$, is a compromise between the unbiased but error-prone sample covariance matrix $\boldsymbol{S}$ and a constrained estimator $\boldsymbol{F}$. We take a convex linear combination, as follows:

$$\widehat{\boldsymbol{\Sigma}}_{shr} = \delta^* \boldsymbol{F} + (1 - \delta^*) \boldsymbol{S} \tag{2.17}$$

where $0 < \delta^* < 1$.

Consider a random sample $X_1, \ldots, X_N$ representing N observations of M measurements from a joint distribution which has covariance matrix $\boldsymbol{\Sigma}$. We require only the following assumptions, as shown in Ledoit and Wolf (2003) [23]:

(a) X_i are independent and identically distributed, $i = 1, \ldots, N$,

(b) X_i have finite fourth moments; for all i in $1, \ldots, N$, and for all h, j, k, l, in $\{1, \ldots, M\}$, $\mathbb{E}\left[|x_{ih} x_{ij} x_{ik} x_{il}|\right] < \infty$

(c) M is finite, and $N \to \infty$

(d) $\boldsymbol{\Phi} \neq \boldsymbol{\Sigma}$

The first three are sufficient to ensure the sample variances and covariances have limiting normal distributions. The fourth will be explained shortly.

We use the following notation,

$$
\begin{aligned}
\boldsymbol{\Sigma} &\equiv \text{the true covariance of the vectors } X_i \\
&\quad \sigma_{jk} \equiv \text{elements of } \boldsymbol{\Sigma} \\
\boldsymbol{S} &= \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})' \\
&\quad \text{where } \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i, \\
&\quad s_{jk} \equiv \text{elements of } \boldsymbol{S} \\
r_{jk} &= \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}} \\
\bar{r} &= \frac{2}{(M-1)M}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M} r_{jk} \\
\boldsymbol{F} &\equiv \{f_{jk}\} = \begin{cases} s_{jj} & ;\ j = k \\ \bar{r}\sqrt{s_{jj}s_{kk}} & ;\ j \neq k \end{cases}
\end{aligned}
\tag{2.18}
$$

The matrix $\boldsymbol{F}$ is the sample constant correlation matrix. The true population constant correlation matrix is denoted as $\boldsymbol{\Phi}$ with elements ϕ_{jk} and it is constructed analogously to $\boldsymbol{F}$ using the population variances σ_{jk} and population correlations.

2.2.1 Estimation of the Optimal Shrinkage Intensity

We define the loss function as the squared Frobenius norm and optimize with respect to δ, by finding the value δ^* which minimizes the expected loss, $R(\delta)$.

$$
\begin{aligned}
L(\delta) &= \|\widehat{\boldsymbol{\Sigma}}_{shr} - \boldsymbol{\Sigma}\|_F^2 \\
\delta^* &= \arg\min_{\delta} \mathbb{E}\, L(\delta) \\
&= \arg\min_{\delta} R(\delta)
\end{aligned}
\tag{2.19}
$$

$$
\begin{aligned}
R(\delta) &= \mathbb{E}\,\|\widehat{\boldsymbol{\Sigma}}_{shr} - \boldsymbol{\Sigma}\|_F^2 \\
&= \mathbb{E}\,\|\delta\boldsymbol{F} + (1-\delta)\boldsymbol{S} - \boldsymbol{\Sigma}\|_F^2 \\
&= \mathbb{E}\Big[\sum_{k=1}^{M}\sum_{j=1}^{M}\left(\delta f_{jk} + (1-\delta)s_{jk} - \sigma_{jk}\right)^2\Big] \\
&= \sum_{k=1}^{M}\sum_{j=1}^{M}\Big(\mathrm{Var}(\delta f_{jk} + (1-\delta)s_{jk} - \sigma_{jk}) + \Big[\mathbb{E}[\delta f_{jk} + (1-\delta)s_{jk} - \sigma_{jk}]\Big]^2\Big) \\
&= \sum_{k=1}^{M}\sum_{j=1}^{M}\Big(\delta^2\mathrm{Var}(f_{jk}) + (1-\delta)^2\mathrm{Var}(s_{jk}) + 2\delta(1-\delta)\mathrm{Cov}(f_{jk}, s_{jk}) \\
&\qquad\qquad + \left(\delta\mathbb{E}[f_{jk} - s_{jk}] + \mathbb{E}[s_{jk} - \sigma_{jk}]\right)^2\Big) \\
&= \sum_{k=1}^{M}\sum_{j=1}^{M}\Big(\delta^2\mathrm{Var}(f_{jk}) + (1-\delta)^2\mathrm{Var}(s_{jk}) + 2\delta(1-\delta)\mathrm{Cov}(f_{jk}, s_{jk}) \\
&\qquad\qquad + \left(\delta\mathbb{E}[f_{jk} - s_{jk}] + \mathrm{Bias}(s_{jk})\right)^2\Big) \\
&= \sum_{k=1}^{M}\sum_{j=1}^{M}\Big(\delta^2\mathrm{Var}(f_{jk}) + (1-\delta)^2\mathrm{Var}(s_{jk}) + 2\delta(1-\delta)\mathrm{Cov}(f_{jk}, s_{jk}) \\
&\qquad\qquad + \delta^2(\phi_{jk} - \sigma_{jk})^2\Big) \qquad (2.20)
\end{aligned}
$$

The last step is due to the fact that $\boldsymbol{S}$ is an unbiased estimator of $\boldsymbol{\Sigma}$. Setting the first derivative with respect to δ of $R(\delta)$ equal to zero, and noting that the risk function is convex with a positive second derivative everywhere, we find the global minimum,

$$
\begin{aligned}
\delta^* &= \frac{\sum_{k=1}^{M}\sum_{j=1}^{M}(\mathrm{Var}(s_{jk}) - \mathrm{Cov}(f_{jk}, s_{jk}))}{\sum_{k=1}^{M}\sum_{j=1}^{M}(\mathrm{Var}(f_{jk}) + \mathrm{Var}(s_{jk}) - 2\,\mathrm{Cov}(f_{jk}, s_{jk}) + (\phi_{jk} - \sigma_{jk})^2)} \\
&= \frac{\sum_{k=1}^{M}\sum_{j=1}^{M}(\mathrm{Var}(s_{jk}) - \mathrm{Cov}(f_{jk}, s_{jk}))}{\sum_{k=1}^{M}\sum_{j=1}^{M}(\mathrm{Var}(f_{jk} - s_{jk}) + (\phi_{jk} - \sigma_{jk})^2)} \qquad (2.21)
\end{aligned}
$$

Since the X_i are iid and have finite fourth moments, we invoke the Central Limit Theorem and Slutzky's Theorem to show that the optimal asymptotic shrink-

age intensity can be written as,

$$\begin{aligned}\delta^* &= \frac{\sum_{k=1}^{M}\sum_{j=1}^{M}(\text{AsyVar}(\sqrt{N}s_{jk}) - \text{AsyCov}(\sqrt{N}f_{jk}, s_{jk}))}{N\sum_{k=1}^{M}\sum_{j=1}^{M}(\phi_{jk} - \sigma_{jk})^2} + \mathcal{O}\Big(\frac{1}{N^2}\Big) \\ &= \frac{\kappa}{N} + \mathcal{O}\Big(\frac{1}{N^2}\Big) \end{aligned} \tag{2.22}$$

We see that κ/N is well defined when $\boldsymbol{\Phi} \neq \boldsymbol{\Sigma}$, as required by assumption (d). We then find a consistent estimator for κ and finally write the estimate of the optimal shrinkage intensity as,

$$\hat{\delta}^* = \max\Big\{0,\ \min\Big\{\frac{\hat{\kappa}}{N},\ 1\Big\}\Big\} \tag{2.23}$$

where

$$\begin{aligned}\hat{\kappa} &= \frac{(\hat{\pi} - \hat{\rho})}{\hat{\gamma}} \\ \hat{\pi} &= \sum_{j=1}^{M}\sum_{k=1}^{M}\frac{1}{N}\sum_{i=1}^{N}\big((x_{ij} - \bar{x}_{\cdot j})(x_{ik} - \bar{x}_{\cdot k}) - s_{jk}\big)^2 \\ \hat{\vartheta}_{jj,jk} &= \frac{1}{N}\sum_{i=1}^{N}\big((x_{ij} - \bar{x}_{\cdot j})^2 - s_{jj}\big)\big((x_{ij} - \bar{x}_{\cdot j})(x_{ik} - \bar{x}_{\cdot k}) - s_{jk}\big) \\ \hat{\rho} &= \sum_{j=1}^{M}\frac{1}{N}\sum_{i=1}^{N}\big((x_{ij} - \bar{x}_{\cdot j})^2 - s_{jj}\big)^2 + \sum_{j=1}^{M}\sum_{k=1,k\neq j}^{M}\frac{\bar{r}}{2}\Big(\sqrt{\frac{s_{kk}}{s_{jj}}}\hat{\vartheta}_{jj,jk} + \sqrt{\frac{s_{jj}}{s_{kk}}}\hat{\vartheta}_{kk,jk}\Big) \\ \hat{\gamma} &= \sum_{j=1}^{M}\sum_{k=1}^{M}(f_{jk} - s_{jk})^2 \end{aligned} \tag{2.24}$$

Note that $\hat{\pi}$ is the sum of all of the usual estimators for the asymptotic variance of each element of $\boldsymbol{S}$, scaled by $\sqrt{N}$. Each estimate in $\hat{\kappa}$ is consistent for the corresponding element of δ^* as N increases while M remains fixed.

The estimate $\hat{\kappa}/N$ may fall outside the interval $[0, 1]$ in which case it is truncated, and $\widehat{\boldsymbol{\Sigma}}_{shr}$ is equal to $\boldsymbol{F}$ if $\hat{\kappa}/N > 1$ or $\boldsymbol{S}$ if $\hat{\kappa}/N < 0$.

2.2.2 Notes on the Method

The influence of $\boldsymbol{F}$, the sample constant correlation matrix, diminishes as $\hat{\delta}^*$ decreases, which happens as:

(a) the variances of the elements of $\boldsymbol{S}$ decreases

(b) the correlation between the elements of $\boldsymbol{F}$ and the corresponding elements of $\boldsymbol{S}$ increases

(c) the mean squared difference between $\boldsymbol{F}$ and $\boldsymbol{S}$ increases

In an empirical study conducted on U.S. stock data, using $N = 239$, and $M = \{30, 50, 100, 225, 500\}$, Ledoit and Wolf (2004) [24], found that the constant correlation shrinkage estimator, $\widehat{\boldsymbol{\Sigma}}_{shr}$, outperformed the sample covariance matrix $\boldsymbol{S}$ in all cases and performed comparably and often only slightly worse than more structured methods designed for use on portfolio data.

In addition to comparing the constant correlation shrinkage estimator, $\widehat{\boldsymbol{\Sigma}}_{shr}$ and $\boldsymbol{S}$ on measures of performance, there are some practical advantages to using $\widehat{\boldsymbol{\Sigma}}_{shr}$. The matrix $\widehat{\boldsymbol{\Sigma}}_{shr}$ is always positive definite, whereas when $M \geq N$, the sample covariance matrix $\boldsymbol{S}$ is semi-positive definite and singular, $\boldsymbol{F}$ is positive definite, and therefore the convex combination, $\widehat{\boldsymbol{\Sigma}}_{shr}$, yields a positive definite matrix. Furthermore, unlike some of the more structured estimators, the method is simple to implement and can be computed nearly as quickly as $\boldsymbol{S}$. In contrast to a factor model, we do not assume that a small number of factors capture the cross-sectional correlations, and thus we do not need to estimate the number of factors, their values

and and residual errors. We include this estimator to see how it performs compared to more complicated methods.

Chapter 3: Allocation Using A Bayesian Method

3.1 Bayesian Exploratory Factor Model

We specify prior distributions on components of the factor model, set initial values, and generate a sample using the Markov chain Monte Carlo framework; see Robert and Casella (2000) [30]. Each iteration gives an estimate of allocation of minimum risk. The estimates of the minimum risk allocation are averaged to obtain the allocation implied by the method.

We use an approach developed in Conti *et al.* (2014) [9] to perform dedicated factor analysis with stochastic search on the structure of the factor loading matrix. The number of latent factors, as well as the allocation of the observed variables to the factors, is not fixed a priori but determined during MCMC sampling. The dedicated factor model where each measurement is allocated to at most one factor has been described as a "factor model with simple structure" in Thurstone (1947) [34].

The model is convenient, because conditionally-conjugate forms are chosen that lead to straightforward posterior computation by a Gibbs sampler. However, Conti *et al.* (2014) [9] note that specification of the hyperparameters in the prior may be difficult. Prior elicitation is particularly important in this model, because in the limiting case as the prior variance for the normal and inverse-gamma components

increases the posterior becomes improper. To address this problem often informative priors are chosen. In the absence of subject matter knowledge, sometimes the hyperparameters in the prior are chosen after an initial analysis of the data. Using the data twice in this manner could lead to an underestimation of uncertainty in model selection.

The model is specified as follows: For each observation $i = 1, \ldots, N$,

$$
\begin{aligned}
Y_i &= \boldsymbol{\alpha}\theta_i + \epsilon_i \\
\theta_i &\sim \mathcal{N}_K(0, \boldsymbol{R}) \\
\epsilon_i &\sim \mathcal{N}_M(0, \boldsymbol{\Sigma}_u) \\
\boldsymbol{\Sigma}_u &= \text{diag}(\sigma_1^2, ..., \sigma_M^2) \\
&(\theta_i, \epsilon_i) \text{ independent} \\
&(\epsilon_i, \epsilon_j) \text{ independent for } i \neq j
\end{aligned} \tag{3.1}
$$

where Y_i is the vector containing M observed continuous variables and is a function of the latent factors and the residual idiosyncratic terms. The vector θ_i contains the K latent factors, and $\boldsymbol{\alpha}$ is the $(M \times K)$ matrix of factor loadings. The vector ϵ_i consists of M residual idiosyncratic terms. The covariance matrix $\boldsymbol{R}$ of the latent factors is assumed to be a correlation matrix. The variances of the idiosyncratic errors are positive; each $\sigma_m^2 > 0$.

Each row of the factor loading matrix $\boldsymbol{\alpha}$ contains at most one non-zero element. The allocation of the manifest variables to the latent factors is indicated by the binary matrix $\boldsymbol{\Delta}$ with the same dimensions as $\boldsymbol{\alpha}$, such that each row Δ_m indicates which factor loading, if any, is different from zero. For example, $\Delta_m =$

$(0, .., 0, 1, 0, ..., 0) \equiv e_k$ indicates that variable m loads on the k-th factor, where e_k is a vector of length K that contains a 1 as its k-th element and the remaining elements are equal to 0.

3.1.1 Identification

We place the following additional restrictions on the model for purposes of identification:

We require $\boldsymbol{R} = \text{Cov}(\theta_i)$ to be of full rank with diagonal elements all equal to 1. This implies $N \geq K$. And, with $N \geq K$, any covariance matrix of K vectors each consisting of N linearly independent random elements with an underlying continuous distribution, will be of full rank with probability one. We rescale such that the variance of each θ_i equals 1.

Identifiability will be ensured if at least three measurements are allocated to each factor. This condition ensures that $\boldsymbol{\alpha}$ has the "row deletion property", see Anderson and Rubin (1956) [1]. This condition is incorporated in the sampling algorithm by requiring a reversion to the prior sample in the case that only one or two measurements are allocated to a particular factor.

Note that identification is achieved only with respect to the scale of the latent factors. Non-identifiability problems may affect the posterior sample due to column switching and sign switching of the factor loadings. The signs of the factor loadings in a given column of $\boldsymbol{\alpha}$ and the sign of the corresponding factor θ_i may be switched simultaneously. Further, the columns of $\boldsymbol{\alpha}$ may be permuted. However, these triv-

ial rotation problems can be addressed after conducting the MCMC sampling, by ordering the columns and switching the signs of the loadings in a consistent manner.

In order to calculate the minimum risk allocation, we need only identification with respect to the scale of the latent factors. This is sufficient for the purposes of the estimation of $\text{Cov}(Y_i) = \boldsymbol{\alpha R \alpha'} + \boldsymbol{\Sigma}_u$.

The identifying restrictions are not not required to conduct inference in a Bayesian approach; however, Conti *et al.* (2014) [9] chose to constrain the sampler to stay in regions of the parameter space where only classically identified models are generated in order to aid in interpretation of the factors and avoid the generation of spurious factors. Further, the assumption of normality on the latent factors and on the error terms, θ_i and ϵ_i, is to ease Bayesian inference. The method simultaneously selects the number of factors and the allocation of measurements to those factors using a log odds ratio for which a complete distributional specification of the model is required.

3.1.2 Prior Specification

3.1.2.1 Prior Distributions on the Indicator Matrix

The allocation of the measurements to groups of dedicated measurements can be interpreted as a mixture problem with an unknown, but finite, number of components. Let τ_k denote the probability that a measurement loads on factor k. Further, let τ_0 denote the probability that the measurement does not load on any of the factors.

The indicators are assumed to have the following probabilities, for $k = 1, \ldots, K$:

$$\Pr(\Delta_m = e_k \mid \tau_k) = \tau_k$$

$$\Pr(\Delta_m = (0, \ldots, 0) \mid \tau_0) = \tau_0$$

$$\tau = (\tau_0, \tau_1, \ldots, \tau_K) = [\tau_0, (1-\tau_0)\tau_1^*, \ldots, (1-\tau_0)\tau_K^*] \tag{3.2}$$

where $\sum_{k=0}^{K} \tau_k = 1$ and $\sum_{k=1}^{K} \tau_k^* = 1$. To conduct Bayesian inference, prior distributions are assigned to the parameters τ_0 and τ^* as follows:

$$\tau_0 \sim \mathcal{B}eta(\kappa_0, \xi_0)$$

$$\tau^* = (\tau_1^*, \ldots, \tau_K^*) \sim \mathcal{D}ir(\kappa) \tag{3.3}$$

with κ_0, ξ_0 and κ as hyperparameters to be specified. The Beta distribution for τ_0 can be selected so as to obtain more or less mass toward 0 or 1, depending on our prior knowledge about the number of measurements that should be discarded from the analysis. For example, a small κ_0 and a large ξ_0 implies a belief that more of the M measurements should be utilized (i.e., allocated to a factor). A Dirichlet distribution on the weights has been used extensively in mixture modeling.

Let $n_k(\boldsymbol{\Delta}) = \sum_{m=1}^{M} \mathbf{1}[\Delta_m = \epsilon_k]$, for $k = 0, \ldots, K$. Here $n_k(\boldsymbol{\Delta})$ is the number of non-zero elements in the k-th column of the loading matrix $\boldsymbol{\alpha}$. In addition, we also compute $n_0(\boldsymbol{\Delta})$ which is the number of measurements that are not allocated to any

factor. The marginal likelihood for the indicator matrix is as follows:

$$
\begin{aligned}
p(\boldsymbol{\Delta}) &= \int\int p(\boldsymbol{\Delta}|\tau_0,\tau^*)p(\tau_0)p(\tau^*)\; d\tau_0 d\tau^* \\
&= \int\int \tau_0^{n_0(\boldsymbol{\Delta})}(1-\tau_0)^{M-n_0(\boldsymbol{\Delta})}\prod_{k=1}^{K}\tau_k^{*n_k(\boldsymbol{\Delta})} \\
&\qquad \times\frac{\Gamma(\kappa_0+\xi_0)}{\Gamma(\kappa_0)\Gamma(\xi_0))}\tau_0^{\kappa_0-1}(1-\tau_0)^{\xi_0-1} \\
&\qquad \times\frac{\Gamma(\sum_{k=1}^{K}\kappa_k)}{\prod_{k=1}^{K}\Gamma(\kappa_k)}\prod_{k=1}^{K}\tau_k^{*\kappa_k-1}d\tau_0 d\tau^* \\
&= \frac{\Gamma(\kappa_0+\xi_0)}{\Gamma(\kappa_0)\Gamma(\xi_0))}\frac{\Gamma(\kappa_0+n_0(\boldsymbol{\Delta}))\Gamma(\xi_0+M-n_0(\boldsymbol{\Delta}))}{\Gamma(\kappa_0+\xi_0+M)} \\
&\qquad \times\frac{\prod_{k=1}^{K}\Gamma(\kappa_k+n_k(\boldsymbol{\Delta}))}{\Gamma\Big(M-n_0(\boldsymbol{\Delta})+\sum_{k=1}^{K}\kappa_k\Big)}
\end{aligned}
\tag{3.4}
$$

3.1.2.2 Prior Distributions on the Idiosyncratic Variances and Factor Loadings

Continuing with the model developed in Conti *et al.* (2014) [9], an Inverse-Gamma prior distribution is assumed on the idiosyncratic variances, $\sigma_1^2,\ldots,\sigma_M^2$. A normal prior distribution is assumed on the non-zero factor loadings, $\alpha_1^{\Delta},\ldots,\alpha_M^{\Delta}$ conditional on σ_m^2, where α_m^{Δ} denotes the only non-zero loading in α_m, the m-th row of $\boldsymbol{\alpha}$.

$$
\begin{aligned}
\sigma_m^2 &\sim \mathcal{I}nv\text{-}\mathcal{G}amma(c_0,\; C_m^0) \\
\alpha_m^{\Delta}\mid\sigma_m^2 &\sim \mathcal{N}(a_m^0,\; A_m^0\sigma_m^2)
\end{aligned}
\tag{3.5}
$$

Let $Y_m = (Y_{1m}, ..., Y_{Nm})'$, then,

$$\begin{aligned} Y_m &= \boldsymbol{\theta}\alpha_m + \epsilon_m \\ &= \theta_k \alpha_m^{\Delta} + \epsilon_m \\ Y_m &\sim \mathcal{N}_N(\theta_k \alpha_m^{\Delta},\ \sigma_m^2 \mathbf{I}) \end{aligned} \tag{3.6}$$

where $\theta_k = (\theta_{1k}, \ldots, \theta_{Nk})'$ is the k-th column of $\boldsymbol{\theta}$, the $(N \times K)$ matrix of latent factors.

The marginal likelihoods of the observed variables can be expressed as follows. Two expressions are obtained to cover the following two cases: (i) the case that no measurements are allocated to any factors (the null model) and (ii) the case that at least three measurements are allocated to a factor (the dedicated case).[1]

First, in case (i), when no measurements are allocated to any factors (the null model):

$$\begin{aligned} p(Y_m|\Delta_m, \boldsymbol{\theta}) &= \int p(Y_m|\boldsymbol{\theta}, \sigma_m^2) p(\sigma_m^2)\ d\sigma_m^2 \\ &= \int_0^{\infty} (2\pi\sigma_m^2)^{-\frac{N}{2}} \exp\left\{-\frac{Y_m' Y_m}{2\sigma_m^2}\right\} \\ &\qquad \times \frac{(C_m^0)^{c_0}}{\Gamma(c_0)} {\sigma_m^2}^{(-c_0-1)} \exp\left\{-\frac{C_m^0}{\sigma_m^2}\right\}\ d\sigma_m^2 \\ &= (2\pi)^{-\frac{N}{2}} \frac{(C_m^0)^{c_0}}{\Gamma(c_0)} \\ &\qquad \times \int_0^{\infty} {\sigma_m^2}^{(-c_N-1)} \exp\left\{-\frac{C_m^{Nnull}}{\sigma_m^2}\right\}\ d\sigma_m^2 \\ &= (2\pi)^{-\frac{N}{2}} \frac{(C_m^0)^{c_0}}{\Gamma(c_0)} \frac{\Gamma(c_N)}{(C_m^{Nnull})^{c_N}} \end{aligned} \tag{3.7}$$

where $c_N = c_0 + \frac{N}{2}$, and

[1] The cases where only one or only two measurements are allocated to a factor are not considered, in order to ensure identifiability of the dedicated factors up to sign and column switching, see Section 3.1.1.

$C_m^{Nnull} = C_m^0 + \frac{Y_m' Y_m}{2}$.

The quantities c_0, and C_m^0 specify the first two moments of the prior distributions of σ_m^2 . We use the symbols c_N and C_m^{Nnull} in order to write the marginal distribution of Y_m more compactly in terms of the conditional moments of $\sigma_m^2|Y_m, \boldsymbol{\theta}, \alpha_m^{\Delta}$. We can write the conditional distributions of the idiosyncratic variances as:

$$\sigma_m^2|Y_m, \boldsymbol{\theta}, \alpha_m^{\Delta} \;\sim\; \mathcal{I}nv\text{-}\mathcal{G}amma(c_N,\; C_m^{Nnull}) \tag{3.8}$$

Second, in case (ii), when at least three measurements are allocated to a factor (the dedicated case):

$$\begin{aligned}
p(Y_m|\Delta_m, \boldsymbol{\theta}) &= \int\int p(Y_m|\boldsymbol{\theta}, \alpha_m^{\Delta}, \sigma_m^2) p(\alpha_m^{\Delta}|\sigma_m^2) p(\sigma_m^2)\; d\alpha_m^{\Delta} d\sigma_m^2 \\
&= \int_0^{\infty} \int_{-\infty}^{\infty} (2\pi\sigma_m^2)^{-\frac{N}{2}} \exp\left\{ -\frac{(Y_m - \theta_k \alpha_m^{\Delta})'(Y_m - \theta_k \alpha_m^{\Delta})}{2\sigma_m^2} \right\} \\
&\qquad \times (2\pi A_m^0 \sigma_m^2)^{-\frac{1}{2}} \exp\left\{ -\frac{(\alpha_m^{\Delta} - a_m^0)^2)}{2A_m^0 \sigma_m^2} \right\} \\
&\quad \times \frac{(C_m^0)^{c_0}}{\Gamma(c_0)} {\sigma_m^2}^{(-c_0-1)} \exp\left\{ -\frac{C_m^0}{\sigma_m^2} \right\}\; d\alpha_m^{\Delta} d\sigma_m^2 \\
&= (2\pi)^{-\frac{N}{2}} (2\pi A_m^0)^{-\frac{1}{2}} \frac{(C_m^0)^{c_0}}{\Gamma(c_0)} \\
&\quad \times \int_{-\infty}^{\infty} {\sigma_m^2}^{-\frac{1}{2}} \exp\left\{ -\frac{(\alpha_m^{\Delta} - A_m^N a_m^N)^2}{2A_m^N \sigma_m^2} \right\}\; d\alpha_m^{\Delta} \\
&\quad \times \int_0^{\infty} {\sigma_m^2}^{(-c_N-1)} \exp\left\{ -\frac{C_m^N}{\sigma_m^2} \right\}\; d\sigma_m^2 \\
&= (2\pi)^{-\frac{N}{2}} \frac{(C_m^0)^{c_0}}{\Gamma(c_0)} \frac{(A_m^N)^{\frac{1}{2}}}{(A_m^0)^{\frac{1}{2}}} \frac{\Gamma(c_N)}{(C_m^N)^{c_N}}
\end{aligned} \tag{3.9}$$

where $c_N = c_0 + \frac{N}{2}$,

$C_m^N = C_m^0 + \frac{1}{2}\Big(Y_m' Y_m + \frac{(a_m^0)^2}{A_m^0} - A_m^N (a_m^N)^2 \Big)$,

$A_m^N = \frac{1}{1/A_m^0 + \theta_k' \theta_k}$, and

$a_m^N = a_m^0 / A_m^0 + \theta_k' Y_m$.

In the second case, in addition to the conditional distribution of the idiosyncratic variances, we have the conditional distribution of the non-zero factor loadings:

$$\begin{aligned} \sigma_m^2|Y_m, \boldsymbol{\theta}, \alpha_m^{\Delta} &\sim \mathcal{I}nv\text{-}\mathcal{G}amma(c_N,\ C_m^N) \\ \alpha_m^{\Delta}|Y_m, \boldsymbol{\theta}, \sigma_m^2 &\sim \mathcal{N}(A_m^N a_m^N,\ A_m^N \sigma_m^2) \end{aligned} \tag{3.10}$$

3.1.2.3 Prior Distributions on the Factor Correlation Matrix

We require $\boldsymbol{R} = \text{Cov}(\theta_i)$ to be of full rank with diagonal elements all equal to 1 (see Section 3.1.1). In the MCMC procedure, the latent factors are sampled and the covariance $\boldsymbol{\Omega}$ is generated. Then, $\boldsymbol{\Omega}$ is transformed to a correlation matrix using $\boldsymbol{\Omega} = \boldsymbol{\Lambda}^{\frac{1}{2}} \boldsymbol{R}\ \boldsymbol{\Lambda}^{\frac{1}{2}}$, where $\boldsymbol{\Lambda} = \text{diag}(\Lambda_1, \ldots, \Lambda_K)$. As shown in Zhang *et al.* (2006) [37], for $\nu - K + 1 > 0$:

$$\boldsymbol{\Omega} \sim \mathcal{I}nv\text{-}\mathcal{W}ishart(\nu,\ \boldsymbol{S}) \tag{3.11}$$

where $\boldsymbol{S}$ is a positive definite $(K \times K)$ matrix.

We find the joint distribution of $\boldsymbol{\Lambda}$ and $\boldsymbol{R}$, which includes a factor equal to the Jacobian of the transformation from $\boldsymbol{\Omega}$ to $\boldsymbol{\Lambda}$ and $\boldsymbol{R}$. The Jacobian of the transformation $(\boldsymbol{\Omega} \to \boldsymbol{\Lambda}, \boldsymbol{R})$ is the determinant of a vector of length $K(K-1)/2$ the number of unique off-diagonal elements in the $(K \times K)$ matrices $\boldsymbol{\Omega}$ and $\boldsymbol{R}$. Each element has the relationship: $\omega_{i,j} = r_{i,j}(\Lambda_i \Lambda_j)^{1/2}$. Then,

$$\begin{aligned} \mathcal{J}_{(\boldsymbol{\Omega} \to \boldsymbol{\Lambda}, \boldsymbol{R})} &= \left| \frac{\partial \omega_{1,2}}{\partial r_{1,2}}, \frac{\partial \omega_{1,3}}{\partial r_{1,3}}, ..., \frac{\partial \omega_{2,3}}{\partial r_{2,3}}, ..., \frac{\partial \omega_{K-1,K}}{\partial r_{K-1,K}} \right| \\ &= \left| (\Lambda_1 \Lambda_2)^{1/2}, (\Lambda_1 \Lambda_3)^{1/2}, ..., (\Lambda_2 \Lambda_3)^{1/2}, ..., (\Lambda_{K-1} \Lambda_K)^{1/2} \right| \\ &= |\boldsymbol{\Lambda}|^{\frac{K-1}{2}} \end{aligned} \tag{3.12}$$

Then, the joint distribution of $\boldsymbol{\Lambda}$ and $\boldsymbol{R}$ is:

$$\begin{aligned} p(\boldsymbol{\Lambda}, \boldsymbol{R}|\boldsymbol{S}) &= \mathcal{J}_{(\boldsymbol{\Omega}\to\boldsymbol{\Lambda},\boldsymbol{R})}\ p(\boldsymbol{\Omega}) \\ &= |\boldsymbol{\Lambda}|^{\frac{K-1}{2}} c\ |\boldsymbol{S}|^{\frac{\nu}{2}} |\boldsymbol{\Omega}|^{-\frac{(\nu+K+1)}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{S}\boldsymbol{\Omega}^{-1})\right\} \\ &= c\ |\boldsymbol{S}|^{\frac{\nu}{2}} |\boldsymbol{\Lambda}|^{-\frac{\nu}{2}-1} |\boldsymbol{R}|^{-\frac{(\nu+K+1)}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{S}\boldsymbol{\Lambda}^{-1}\boldsymbol{R}^{-1})\right\} \end{aligned} \quad (3.13)$$

where $c = (2^{\nu/2}\Gamma_K(\nu/2))^{-1}$ and $\Gamma_K(\cdot)$ is the multivariate gamma function. Further, we note that $|\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{R}\ \boldsymbol{\Lambda}^{\frac{1}{2}}| = |\boldsymbol{\Lambda}||\boldsymbol{R}|$, since $|\boldsymbol{A}\boldsymbol{B}| = |\boldsymbol{A}||\boldsymbol{B}|$ for any same sized square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

As shown in Barnard *et al.* (2000) [4] and Zhang *et al.* (2006) [37], in the special case where $\boldsymbol{S}$ is a diagonal matrix, whether $\boldsymbol{S}$ is considered fixed or random, the marginal distribution of $\boldsymbol{R}$ has a closed form as follows:

$$\begin{aligned} p(\boldsymbol{R}|\boldsymbol{S}) &= \int p(\boldsymbol{\Lambda}, \boldsymbol{R}|\boldsymbol{S})\ d\boldsymbol{\Lambda} \\ &= c\ |\boldsymbol{R}|^{-\frac{(\nu+K+1)}{2}} \prod_{k=1}^{K} \left(s_k^{\frac{\nu}{2}} \int_0^\infty \Lambda_k^{-\frac{\nu}{2}-1} \exp\left\{-\frac{s_k r^{kk}}{2\Lambda_k}\right\} \right) d\boldsymbol{\Lambda} \\ &= c\ |\boldsymbol{R}|^{-\frac{(\nu+K+1)}{2}} \prod_{k=1}^{K} \left(s_k^{\frac{\nu}{2}} s_k^{-\frac{\nu}{2}} (r^{kk})^{\frac{\nu}{2}} \Gamma\big(\frac{\nu}{2}\big) 2^{\frac{\nu}{2}} \right) \\ &= \frac{\prod_{k=1}^{K} \Gamma(\frac{\nu}{2})}{\Gamma_k(\frac{\nu}{2})}\ |\boldsymbol{R}|^{-\frac{(\nu+K+1)}{2}} \prod_{k=1}^{K} (r^{kk})^{\frac{\nu}{2}} \\ &= p(\boldsymbol{R}) \end{aligned} \quad (3.14)$$

where r^{kk} is the k-th diagonal element of $\boldsymbol{R}^{-1}$.

Whether $\boldsymbol{S}$ is random or whether it is a fixed parameter, $p(\boldsymbol{R})$ is not affected, leaving Bayesian inference invariant to the choice of the prior on $\boldsymbol{S}$. However, as noted in Conti *et al.* (2014) [9], assuming a prior distribution of $\boldsymbol{S}$ improves the marginal data augmentation algorithm used for inference.

Continuing with the special case that $\boldsymbol{S}$ is a diagonal matrix, $\boldsymbol{S} = \text{diag}(s_1, \ldots, s_K)$, from Eq. (3.14), and using:

$$\begin{aligned} p(\boldsymbol{\Lambda}|\boldsymbol{R}, \boldsymbol{S}) &= \frac{p(\boldsymbol{\Lambda}, \boldsymbol{R}|\boldsymbol{S})}{p(\boldsymbol{R}|\boldsymbol{S})} \\ &= \frac{p(\boldsymbol{\Lambda}, \boldsymbol{R}|\boldsymbol{S})}{p(\boldsymbol{R})} \end{aligned} \tag{3.15}$$

we can see that no matter the prior on $\boldsymbol{S}$, one can write a closed form expression for each variance $\Lambda_k|\boldsymbol{R}, s_k$ as follows:

$$\Lambda_k|\boldsymbol{R}, s_k \sim \mathcal{I}nv\text{-}\mathcal{G}amma\Big(\frac{\nu}{2},\ \frac{s_k r^{kk}}{2}\Big) \tag{3.16}$$

If one assumes the following hyper-prior $p(\boldsymbol{S})$ following the approach of Huang and Wand (2013) [21], then one can draw samples of s_k at each iteration, using the prior distribution,

$$s_k \sim \mathcal{G}amma\Big(\frac{1}{2},\ \frac{1}{2zA_k^2}\Big) \tag{3.17}$$

where $z = \nu - K + 1$.

In other words, to achieve the desired result, one could replace ν in the prior of $\boldsymbol{\Omega}$ shown in Eq. (3.11) with z. The re-parameterization makes s_k dependent on the dimension K, which is necessary to achieve a uniform distribution for the marginal distribution of each correlation, $r_{i,j}$, the off-diagonal elements of $\boldsymbol{R}$. We can show that setting $\nu = K + 1$ yields a "flat prior", under which the marginal distribution of each correlation is uniform on $(-1, 1)$.

Changing ν to $z - K + 1$ in Eq. (3.13) results in

$$\begin{aligned} p(\boldsymbol{\Lambda}, \boldsymbol{R}|\boldsymbol{S}) &= c_z\ |\boldsymbol{S}|^{\frac{(z+K-1)}{2}} |\boldsymbol{\Lambda}|^{-\frac{(z+k+1)}{2}} |\boldsymbol{R}|^{-\frac{(z+2k)}{2}} \\ &\quad \times \exp\left\{-\frac{1}{2}\text{tr}(\boldsymbol{S}\boldsymbol{\Lambda}^{-1}\boldsymbol{R}^{-1})\right\} \end{aligned} \tag{3.18}$$

where c_z is the normalizing constant. Again, in the special case where $\boldsymbol{S}$ is a diagonal matrix and we have the prior of s_k as shown in Eg. (3.17), then,

$$p(\boldsymbol{\Lambda}, \boldsymbol{R}|\boldsymbol{S}) \propto |\boldsymbol{R}|^{-\frac{(z+2k)}{2}} \prod_{k=1}^{K} \left(s_k^{\frac{(z+K-2)}{2}} \Lambda_k^{-\frac{(z+k+1)}{2}} \exp\left\{ -\frac{s_k r^{kk}}{2\Lambda_k} - \frac{s_k}{2zA_k^2} \right\} \right) \tag{3.19}$$

It suffices to consider only the marginal distribution in a (2×2) covariance matrix. For any (2×2) correlation term in $\boldsymbol{\Omega}$, we can apply the sub-covariance matrix property. Under our parameterization, the marginal distribution of any sub-covariance matrix in $\boldsymbol{\Omega}$ has the same distributional form as $\boldsymbol{\Omega}$ itself; see Huang *et al.* (2013) [21]. That is,

$$\begin{aligned} p(\Lambda_1, \Lambda_2, s_1, s_2, r_{1,2}) &\propto |\boldsymbol{R}|^{-\frac{(z+4)}{2}} \prod_{k=1}^{2} \left(s_k^{\frac{z}{2}} \Lambda_k^{-\frac{(z+3)}{2}} \exp\left\{ -\frac{s_k r^{kk}}{2\Lambda_k} - \frac{s_k}{2zA_k^2} \right\} \right) \\ &\propto (1 - r_{1,2}^2)^{-\frac{(z+4)}{2}} \prod_{k=1}^{2} \left(s_k^{\frac{z}{2}} \Lambda_k^{-\frac{(z+3)}{2}} \exp\left\{ -\frac{s_k}{2\Lambda_k(1 - r_{1,2}^2)} - \frac{s_k}{2zA_k^2} \right\} \right) \end{aligned} \tag{3.20}$$

where $s_1, s_2, \Lambda_1, \Lambda_2 > 0, \quad -1 < r_{1,2} < 1$.

The terms involving s_1 form the kernel of a Gamma distribution with parameters $(z/2 + 1,\ 1/(2\Lambda_1(1 - r_{1,2}^2)) + 1/2zA_1^2)$, and similarly for s_2. Integrating with respect to s_1 and s_2 gives the marginal distribution of $(\Lambda_1, \Lambda_2, r_{1,2})$:

$$p(\Lambda_1, \Lambda_2, r_{1,2}) \propto (1 - r_{1,2}^2)^{-\frac{(z+4)}{2}} \prod_{k=1}^{2} \left[\Lambda_k^{-\frac{(z+3)}{2}} \left(\frac{1}{\Lambda_k(1 - r_{1,2}^2)} + \frac{1}{zA_k^2} \right)^{-\frac{(z+2)}{2}} \right] \tag{3.21}$$

Integrating with respect to Λ_1 and Λ_2 gives the marginal distribution of $r_{1,2}$, and

using the substitution $u_k = \left(\frac{1}{\Lambda_k(1-r_{1,2}^2)} + \frac{1}{zA_k^2}\right)$.

$$\begin{aligned} p(r_{1,2}) &\propto \int_0^\infty (1-r_{1,2}^2)^{-\frac{(z+4)}{2}} \prod_{k=1}^{2} \left[\Lambda_k^{-\frac{(z+3)}{2}} \left(\frac{1}{\Lambda_k(1-r_{1,2}^2)} + \frac{1}{zA_k^2}\right)^{-\frac{(z+2)}{2}}\right] \, d\Lambda_k \\ &\propto \int_0^\infty (1-r_{1,2}^2)^{-\frac{(z+4)}{2}} \prod_{k=1}^{2} \left[(1-r_{1,2}^2)^{\frac{(z+1)}{2}} (u_k - \frac{1}{zA_k^2})^{\frac{(z-1)}{2}} u_k^{-\frac{(z+2)}{2}}\right] \, du_k \\ &\propto (1-r_{1,2}^2)^{\frac{(z-2)}{2}} \end{aligned} \tag{3.22}$$

When z is chosen to be 2 (or equivalently, ν is chosen to equal $K+1$), the marginal distribution of the correlation parameter $r_{1,2}$ in $\boldsymbol{R}$ is uniform on $(-1,1)$. Any sub-covariance of a matrix distributed as Inverse-Wishart is also distributed as Inverse-Wishart, so the result holds for any correlation parameter $r_{i,j}$ in $\boldsymbol{R}$.

Chapter 4: Applications to Simulated and Real Data

4.1 Simulated Data Study

We carry out simulations to test the methods of covariance estimation. Assume that the matrix $\boldsymbol{Y}$ is observed, where $\boldsymbol{Y} = (\mathrm{y}_1, \ldots, \mathrm{y}_N)$, y_n is an $(M \times 1)$ vector, $n = 1, \ldots, N$, and that the observations are the logarithms of the daily excess returns over the risk-free rate of M stocks over a period of N days.

A parameter of interest is the variance-covariance matrix $\mathrm{Cov}(\mathrm{y}_n)$ where y_n is an $(M \times 1)$ random vector. We assume the distribution of y_n has a constant covariance matrix $\boldsymbol{\Sigma}$ over the N days and we seek to estimate $\boldsymbol{\Sigma}$ using the N observations, y_n, $n = 1, \ldots, N$. As discussed below, we assume a factor model with factors that may cover latent time-varying factor loadings resulting in a covariance matrix $\mathbf{B}\mathrm{Cov}(\mathrm{f}_n)\mathbf{B}'$ that is constant over time.

Another parameter of interest is the allocation of the M assets that results in portfolio returns with the lowest variance. In other words, we seek the optimum portfolio allocation vector, say θ, under the criterion of minimizing the variance of the portfolio returns. We estimate θ using data $\boldsymbol{Y}$ and refer to the estimate as $\hat{\theta}$. Then we compute the true variance of the returns using the portfolio allocation $\hat{\theta}$ and compare it to the true global minimum variance σ_g^2, which will be described

in detail below. Utilizing the unweighted Bayes estimator, we construct credible regions for the unknown parameter vector θ.

We choose this application because it provides an opportunity to evaluate particular details of covariance estimation and to apply and compare several Bayesian methods. However, we contend that the methods could be generalized to many other applications involving covariance matrices and functions of covariance matrices.

The Bayesian methods, the evaluation functions and the comparisons are described in detail in Section 4.1.3. First, we describe the method of data generation and the specifications needed to implement the methods.

4.1.1 Data Generation

We examine the performance of the methods on simulated data. In order to find an estimator that performs well at estimating the parameter of interest which is the true global minimum variance σ_g^2, we simulate returns on investment alternatives and assume a goal of minimizing portfolio risk (i.e., minimizing the variance of the returns of the selected portfolio.) The number of investment alternatives is fixed at $M = 100$, and the time span N increases from 20 to 300. We assume that the natural logarithms of the excess returns over the risk-free interest rate of the M stocks follow a factor model with six factors,

$$y_{mn} = \sum_{j=1}^{6} b_{mj} f_{jn} + u_{mn} \tag{4.1}$$

where $m = 1, \ldots, M$, and $n = 1, \ldots, N$,

$$\mathrm{b}_m \sim \mathcal{N}_6(\mu_B, \mathbf{\Sigma}_B),\ \mathrm{b}_m = (b_{m1}, \ldots, b_{m6})'$$

$$\mathrm{u}_n \sim \mathcal{N}_M(0, \mathbf{\Sigma}_u),\ \mathrm{u}_n = (u_{1n}, \ldots, u_{Mn})'$$

$\mathbf{B} = (\mathrm{b}_1, \ldots, \mathrm{b}_M)'$ is an $(M \times 6)$ matrix of factor loadings, $\mathrm{f}_n = (f_{1n}, \ldots, f_{6n})'$ is a vector of common factors, and u_n denotes the idiosyncratic component of the model.

We assume that f_n is uncorrelated with u_n. Further, we assume that $\mathbf{\Sigma}$ is invariant with respect to the dimension N. The factor returns f_n are assumed to follow a stationary vector autoregressive model of order 1, denoted as VAR(1),

$$\mathrm{f}_n = \mu + \Phi \mathrm{f}_{n-1} + \epsilon_n \tag{4.2}$$

To generate realistic values for $\mu_B, \mathbf{\Sigma}_B, \mathbf{\Sigma}_u$, and the parameters of the VAR(1) process, we calibrate the model using actual log excess returns from a randomly drawn portfolio of 100 companies from the Standard & Poor's 500 stock market index, an index of more than 500 large American companies having common stock listed on the NYSE or NASDAQ. The daily returns for a 500 day period are reduced by the risk-free return to obtain the daily excess return. The risk-free return is assumed to be the daily return on the U.S. Department of the Treasury auctioned 13-week Treasury bills. The natural logarithms of the daily excess returns are utilized. The calibration steps are similar to those in Section 6 of Fan *et al.* (2013) [15].

A six factor model is fit to the input data using the principal components method described in Section 2.1, resulting in a (100×6) matrix $\widetilde{\mathbf{B}}$, and a (500×6) matrix $\widetilde{\mathbf{F}}$.

The sample mean vector of the the fitted $\widetilde{\mathbf{B}}$ is assumed to be the population

mean vector. The sample covariance matrix of the fitted $\widetilde{\mathbf{B}}$ is assumed to be the population covariance matrix. They are shown in Table 4.1.

Table 4.1: Parameters to generate b

μ_b	$\mathbf{\Sigma}_B$					
-1.673	0.480	0.192	0.051	0.113	0.035	0.033
0.114	0.192	0.300	-0.003	-0.008	-0.002	-0.002
0.030	0.051	-0.003	0.162	-0.002	-0.001	-0.001
0.067	0.113	-0.008	-0.002	0.154	-0.001	-0.001
0.021	0.035	-0.002	-0.001	-0.001	0.119	0.000
0.019	0.033	-0.002	-0.001	-0.001	0.000	0.102

For Model Design 1, generate $\mathbf{B}$ which remains fixed over 100 replications, where $\mathbf{B} = (\mathrm{b}_1, \ldots, \mathrm{b}_M)'$, and $\{\mathrm{b}_m\}_{m=1}^{M} \sim \mathcal{N}_6(\mu_B, \mathbf{\Sigma}_B)$, independently.

A VAR(1) model is fit to $\widetilde{\mathbf{F}}$ and the multivariate least squares estimators are found for μ and Φ. To ensure the model is stationary, the eigenvalues of Φ are examined to ensure that they fall within the unit circle. The estimated parameters are shown in Table 4.2.

Table 4.2: Parameters to generate f_n

μ	$\mathrm{Cov}(\mathrm{f}_n)$						Φ					
-0.001	1.010	-0.064	0.008	0.012	0.049	-0.041	-0.108	-0.057	0.027	0.034	-0.014	-0.003
0.000	-0.064	1.165	-0.472	0.096	0.281	-0.042	-0.252	-0.116	0.008	-0.037	-0.033	0.031
0.009	0.008	-0.472	1.830	0.090	0.113	0.018	0.189	0.010	-0.001	-0.065	-0.109	0.026
0.000	0.012	0.096	0.090	1.171	-0.005	0.022	0.015	-0.064	-0.071	0.002	-0.068	0.084
0.000	0.049	0.281	0.113	-0.005	1.862	-0.070	-0.225	-0.129	-0.015	-0.042	-0.082	0.091
-0.012	-0.041	-0.042	0.018	0.022	-0.070	1.512	0.256	0.047	-0.011	0.039	-0.031	-0.025

Let $\widetilde{\mathbf{\Sigma}}_u = \mathrm{Cov}(\mathrm{y}_n - \widetilde{\mathbf{B}}\tilde{\mathrm{f}_n})$. This matrix is not diagonal in general, but we consider only $\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_M^2$, the diagonal elements of $\widetilde{\mathbf{\Sigma}}_u$. We fit a gamma distribution to this vector using the maximum likelihood estimates of the first two moments, $\alpha = (\overline{\tilde{\sigma}^2})^2/\mathrm{Var}(\tilde{\sigma}_m^2)$, and $\beta = \mathrm{Var}(\tilde{\sigma}_m^2)/\overline{\tilde{\sigma}^2}$, where $\overline{\tilde{\sigma}^2}$ is the mean of the $\tilde{\sigma}_m^2$. We assume $\mathbf{\Sigma}_u$ is diagonal and can be written as $\mathrm{diag}(\sigma_1^2, \ldots, \sigma_M^2)$. Each σ_m^2 is generated

independently from $\mathcal{G}amma(\alpha, \beta)$.

Two Additional Models

Model Design Two: B_2 is using only the highest two values in each row of $\widetilde{\mathbf{B}}$. A corresponding F_2 is found, using least squares regression. Then F_2 is fit to a VAR(1) process.

Model Design Three: $\mathrm{Cov}(\mathrm{b}_m)$ is assumed to be an identity matrix. B_3 is drawn, and only the largest absolute value in each row is retained. All other values in the row are replaced with zero. A corresponding F_3 is found, using least squares regression, and F_3 is fit to a VAR(1) process.

Simulations

For each of the three models, M is fixed at 100. N ranges from 20 to 300, in increments of 10. We first generate $\mathbf{B}$ which remains fixed over 100 replications.

For each fixed N,

1. Generate $\{\mathrm{u}_n\}_{n=1}^N$ independently, for each replication.
2. Generate $\{\mathrm{f}_n\}_{n=1}^N$ independently, for each replication.
3. Calculate$\{\mathrm{y}_n\}_{n=1}^N$ using $\mathrm{y}_n = \mathbf{B}\mathrm{f}_\mathrm{n} + \mathrm{u}_\mathrm{n}$, for each replication.

4.1.2 Required Specifications

Specifications for SFM Method

We assume K is an unknown parameter to be estimated. We estimate K by using one of the methods proposed in Section 2.1.2, presented in Bai and Ng

(2002) [3], in which K is chosen to minimize a data-driven objective function with a penalty for increasing K.

Specifications for the POET Method

K is chosen using the same procedure as under the SFM method.

The value of the thresholding constant is determined using the cross-validation procedure described in Section 2.1.2. The cross-validation procedure indicates a thresholding constant of .5 for the generated data.

Specifications for the B.POST Method

The prior distribution on the idiosyncratic variances σ_m^2 is assumed to be inverse-Gamma with shape and scale parameters c_0 and C_m^0. When $N > M$, we can use a data-driven prior that makes use of the observed covariance matrix to specify the hyperparameters; see Frühwirth-Schnatter and Lopes (2010) [18].

$$C_m^0 = \begin{cases} (c_0 - 1)/(\widehat{\boldsymbol{\Sigma}}_{\text{SAM}}^{-1})_{mm} & ;\ N > M \\ 0.5 & ;\ M \geq N \end{cases} \tag{4.3}$$

An Inverse Gamma prior for the idiosyncratic variances ensures that no estimates will be negative. However, other issues may occur that render the estimates nonsensical. For example, a Heywood case may arise in which the constraint $\sigma_m^2 \leq 1/((\boldsymbol{\alpha}\boldsymbol{R}\boldsymbol{\alpha}' + \boldsymbol{\Sigma})^{-1})_{mm}$ is violated, see Heywood (1931) [20]. Such violation will occur when C_m^0 is very large, and will be avoided with very small C_m^0. However, small C_m^0 introduces downward bias and thus a middle ground is sought. When $N > M$, the sample covariance can be utilized to include a rough estimate of the relative size of σ_m^2. The posterior distribution has a positive mean which is a composite of the prior and likelihood estimates. The choice of c_0 also has an affect on

the estimate of σ_m^2. For the case of $N > M$, when c_0 is large, more weight is given to the estimate $\frac{1}{(\widehat{\mathbf{\Sigma}}_{\text{SAM}}^{-1})_{mm}}$. When c_0 is small, or equivalently, as it approaches 1, more weight is given to the estimate $\left(Y_m' Y_m - A_m^N (a_m^N)^2\right)$, see Eq. (3.9), noting that $a_m^0 = 0$. We assume the value of 2.5 as suggested in Frühwirth-Schnatter and Lopes (2010) [18]. Lower values may not be sufficient to bound the prior away from zero. Lopes and West (2004) [25] showed that a value of 1.1 resulted in estimates that were too close to zero in their simulations.

The quantity A_m^0 affects both the estimate of σ_m^2 and the estimate of $\alpha_m^{\mathbf{\Delta}} | \sigma_m^2$. An increase in A_m^0 will decrease the expected value of $\hat{\sigma}_m^2$ and increase A_m^N. The higher the value of A_m^0 the more the estimate of $\text{Cov}(Y_i)$ is shifted from the idiosyncratic component to the communality component; more of the variance is explained by the factors.

The choice of κ will affect the prior probability that a particular number of factors is utilized. A lower value of κ is appropriate as the true number of factors increases. Therefore, in the simulations, we carefully select κ for the different designs in order to generate plausible prior probabilities for the number of factors. Table 4.3 shows the probabilities conditional on the inclusion of the measurement in the model, simulated from an accept-reject sampling scheme for three values of κ. We specify $\kappa = 1$ for all three models. The probability of a measurement's inclusion is determined by the Beta distribution on τ_0 and we set τ_0 with an expectation of .5 for all designs.

The prior specifications are summarized in Table 4.4.

We run 20,000 iterations for each of the 100 Monte Carlo replications, discard-

Table 4.3: Prior Distribution of the Number of Factors

M	$K_{\max}$	κ	# Factors and Probability							True K
			6	7	8	9	10	11	12	
100	12	1	0.00	0.05	0.52	0.41	0.02	0.00	0.00	9
100	12	0.8	0.00	0.03	0.41	0.51	0.05	0.00	0.00	9
100	12	0.5	0.00	0.01	0.21	0.56	0.21	0.01	0.00	9

Table 4.4: Specification of Prior Parameter Values

Indicator matrix		
	κ_0	0.1
	ξ_0	0.1
	κ	1
Idiosyncratic variances		
	c_0	2.5
	C_m^0	.5
Factor loadings		
	a_m^0	0
	A_m^0	3
Factor correlation matrix		
	ν	$K_{\max}+1$
	$\boldsymbol{S}$	$\mathbf{I}_{K_{\max}}$
	ν^*	2
	A_k^2	0.5

ing the first 10,000 to accommodate a burn-in period.

4.1.3 Parameters of Interest

4.1.3.1 Covariance Matrix, $\mathbf{\Sigma}$

As mentioned previously, we assume the distribution of y_n has a constant covariance matrix $\mathbf{\Sigma}$ over the N days and we seek to estimate $\mathbf{\Sigma}$ using the N observations, y_n, $n = 1 \ldots, N$. We apply the following methods to estimate $\mathbf{\Sigma}$:

1. Sample Covariance Estimator, $\widehat{\mathbf{\Sigma}}_{\mathrm{SAM}}$
2. Constant Correlation Composite Estimator, $\widehat{\mathbf{\Sigma}}_{\mathrm{CCC}}$
3. Strict Factor Model Estimator, $\widehat{\mathbf{\Sigma}}_{\mathrm{SFM}}$
4. POET Estimator, $\widehat{\mathbf{\Sigma}}_{\mathrm{POET}}$
5. B.POST Estimator, $\widehat{\mathbf{\Sigma}}_{\mathrm{B.POST}}$

The first is the ordinary sample covariance estimator and its inverse is only available when $N > M$. Methods 2, 3, and 4 are described in detail in Chapter 2.

A fifth method is based on the model described in Chapter 3, but needs more explanation since there are several ways to obtain an estimate of covariance using the model suggested in Conti *et al.* (2014) [9], the Bayesian Exploratory Factor Model. We specify values for prior distribution parameters as detailed in Section 4.1.2 and allow the method to simultaneously estimate the number of factors on which the observations depend and estimate the covariance of the observations. For

the B.POST Estimator, we run 20,000 iterations for each of the 100 Monte Carlo replications, discarding the first 10,000. Further, it is often the case that a small percentage of the iterations result in a model that is not identifiable and those iterations are discarded. A covariance matrix $\widehat{\mathbf{\Sigma}}_{\text{B.POST}_s}$ is generated for each saved iteration, $s = 1, \ldots, S$. Finally, we calculate $\mathbb{E}\left[\mathbf{\Sigma} \mid \text{data}\right] = \widehat{\mathbf{\Sigma}}_{\text{B.POST}}$, the mean of the S covariance matrices. We utilize the R packages "POET" by Fan *et al.* (2016) [16] and "BayesFM" by R. Piatek (2017) [28].

4.1.3.2 Allocation Vector, θ

A parameter of interest which we designate as θ is the allocation vector that results in portfolio returns with the lowest variance. We seek the optimum portfolio allocation vector under the criterion of minimizing the variance of the portfolio returns.

This parameter is meaningful when a selection of the measurements is made based on estimated composite risk. For example, in portfolio analysis it is often necessary to combine assets to meet a risk criterion. We assume that $\boldsymbol{Y}$ represents returns on M assets over N time periods, and that we have a goal of estimating the portfolio allocation of the M assets which results in the smallest variance of returns.

Each data generating process used in the simulation study depends on a known M dimensional non-random positive definite covariance matrix $\text{Cov}(\text{y}_n)$, denoted as $\mathbf{\Sigma}$. The global minimum variance portfolio is the unique value of the function

$$\underset{\text{w}|\ \text{w}'\mathbf{1}=1}{\arg\min}\ \ \text{w}'\mathbf{\Sigma}\text{w}$$

where $\mathrm{w} = (w_1, \ldots, w_M)'$ denotes the vector of portfolio weights, and $\mathbf{1}$ is an $(M \times 1)$ vector of ones. The solution is easily obtained using the method of Lagrange multipliers and is given by $\theta = (1/\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1})\boldsymbol{\Sigma}^{-1}\mathbf{1}$. We are interested in estimating θ.

The global minimum variance is denoted σ_g^2 and is equal to $\theta'\boldsymbol{\Sigma}\theta = 1/(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1})$.

We find an estimate of the portfolio weights under each method considered. For each method "a", a $\in$ {SAM, CCC, SFM, POET, B.POST},

$$\hat{\theta}_{\mathrm{a}} = \frac{\widehat{\boldsymbol{\Sigma}}_{\mathrm{a}}^{-1}\mathbf{1}}{\mathbf{1}'\widehat{\boldsymbol{\Sigma}}_{\mathrm{a}}^{-1}\mathbf{1}} \tag{4.4}$$

We estimate the frequentist's risk of $\hat{\theta}$ by the expected value of Eq. (4.5) for an estimator $\hat{\theta}$ under the model.

$$\begin{aligned} L(\hat{\theta}, \theta) &= \|\boldsymbol{\Sigma}^{1/2}(\hat{\theta} - \theta)\|^2 \\ &= (\hat{\theta} - \theta)'\boldsymbol{\Sigma}(\hat{\theta} - \theta) \\ &= \hat{\theta}'\boldsymbol{\Sigma}\hat{\theta} - 2\theta'\boldsymbol{\Sigma}\hat{\theta} + \theta'\boldsymbol{\Sigma}\theta \\ &= \hat{\theta}'\boldsymbol{\Sigma}\hat{\theta} - \theta'\boldsymbol{\Sigma}\theta \\ &= \hat{\theta}'\boldsymbol{\Sigma}\hat{\theta} - \sigma_g^2 \end{aligned} \tag{4.5}$$

since

$$\begin{aligned} \theta'\boldsymbol{\Sigma}\hat{\theta} &= \frac{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}}{(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1})(\mathbf{1}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{1})} \\ &= \frac{1}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}} \\ &= \theta'\boldsymbol{\Sigma}\theta \end{aligned} \tag{4.6}$$

Note that $\hat{\theta}'\boldsymbol{\Sigma}\hat{\theta} \geq \sigma_g^2$.

Under a Bayesian approach, we minimize the posterior risk, that is, the expected value of the loss given the data and the prior. The Bayesian approach will

depend on the loss function. We consider the loss function:

$$L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)' W(\boldsymbol{\Sigma})(\hat{\theta} - \theta) \tag{4.7}$$

where $W(\boldsymbol{\Sigma})$ is an $(M \times M)$ nonsingular matrix. In particular, when $W(\boldsymbol{\Sigma}) = \mathbf{I}$, we have an unweighted loss function.

Utilizing the iterations of the Bayesian model in Conti *et al.* (2014) [9], the Bayesian Exploratory Factor Model, we obtain three Bayesian estimators of the vector parameter θ.

(1) Unconstrained Bayes Estimator under Weighted Loss, $\hat{\theta}^B$

The Bayes estimator of θ is obtained by minimizing the posterior risk $\mathbb{E}\left[L(\hat{\theta}, \theta) \mid \text{data}\right]$, where the expectation is over the posterior distribution of $\boldsymbol{\Sigma}$ under the model and the prior. We will show that the Bayes estimator of θ under the above loss, model and prior is given by:

$$\hat{\theta}^B = \left(\mathbb{E}\left[W(\boldsymbol{\Sigma}) \mid \text{data}\right]\right)^{-1} \mathbb{E}\left[W(\boldsymbol{\Sigma})\, \theta \mid \text{data}\right]. \tag{4.8}$$

To show this, note that

$$\begin{aligned}
\mathbb{E}\left[L(\hat{\theta},\theta) \mid \text{data}\right] &= \mathbb{E}\left\{\left[(\hat{\theta}-\hat{\theta}^B)+(\hat{\theta}^B-\theta)\right]' W(\boldsymbol{\Sigma})\left[(\hat{\theta}-\hat{\theta}^B)+(\hat{\theta}^B-\theta)\right] \mid \text{data}\right\} \\
&= \mathbb{E}\left[(\hat{\theta}-\hat{\theta}^B)'W(\boldsymbol{\Sigma})(\hat{\theta}-\hat{\theta}^B) \mid \text{data}\right] \\
&\quad + \mathbb{E}\left[(\hat{\theta}^B-\theta)'W(\boldsymbol{\Sigma})(\hat{\theta}^B-\theta) \mid \text{data}\right] \\
&\quad + 2\mathbb{E}\left[(\hat{\theta}-\hat{\theta}^B)'W(\boldsymbol{\Sigma})(\hat{\theta}^B-\theta) \mid \text{data}\right] \\
&= \mathbb{E}\left[(\hat{\theta}-\hat{\theta}^B)'W(\boldsymbol{\Sigma})(\hat{\theta}-\hat{\theta}^B) \mid \text{data}\right] \\
&\quad + \mathbb{E}\left[(\hat{\theta}^B-\theta)'W(\boldsymbol{\Sigma})(\hat{\theta}^B-\theta) \mid \text{data}\right] \\
&= (\hat{\theta}-\hat{\theta}^B)'\mathbb{E}\left[W(\boldsymbol{\Sigma}) \mid \text{data}\right](\hat{\theta}-\hat{\theta}^B) \\
&\quad + \mathbb{E}\left[(\hat{\theta}^B-\theta)'W(\boldsymbol{\Sigma})(\hat{\theta}^B-\theta) \mid \text{data}\right],
\end{aligned}$$

since

$$\begin{aligned}
&\mathbb{E}\left[(\hat{\theta}-\hat{\theta}^B)'W(\boldsymbol{\Sigma})(\hat{\theta}^B-\theta) \mid \text{data}\right] \\
&\quad = (\hat{\theta}-\hat{\theta}^B)'\mathbb{E}\left[W(\boldsymbol{\Sigma})(\hat{\theta}^B-\theta) \mid \text{data}\right] \\
&\quad = (\hat{\theta}-\hat{\theta}^B)'\left[\mathbb{E}\left\{W(\boldsymbol{\Sigma}) \mid \text{data}\right\}\hat{\theta}^B - \mathbb{E}\left\{W(\boldsymbol{\Sigma})\,\theta \mid \text{data}\right\}\right] \\
&\quad = (\hat{\theta}-\hat{\theta}^B)'\left[\mathbb{E}\left[W(\boldsymbol{\Sigma}) \mid \text{data}\right]\left[\mathbb{E}\left(W(\boldsymbol{\Sigma}) \mid \text{data}\right)\right]^{-1}\mathbb{E}\left[W(\boldsymbol{\Sigma})\,\theta \mid \text{data}\right] - \mathbb{E}\left\{W(\boldsymbol{\Sigma})\,\theta \mid \text{data}\right\}\right] \\
&\quad = 0
\end{aligned}$$

The result is found by setting the setting the first derivative with respect to $\hat{\theta}$ equal to zero, and it can be seen from the fact that the second term in the last equality does not involve $\hat{\theta}$.

The estimator does not in general satisfy the necessary constraint $\mathbf{1}'\hat{\theta}^B = 1$, so we will not include $\hat{\theta}^B$ in our evaluation.

(2) Constrained Bayes Estimator under Weighted Loss, $\hat{\theta}^{CB}$

For the constrained optimization it is enough to consider

$$f(\hat{\theta}) = (\hat{\theta} - \hat{\theta}^B)'\mathbb{E}\left[W(\boldsymbol{\Sigma}) \mid \text{data}\right](\hat{\theta} - \hat{\theta}^B) + 2\lambda(\mathbf{1}'\hat{\theta} - 1). \tag{4.9}$$

Now $\frac{\partial f(\hat{\theta})}{\partial \hat{\theta}} = 0$ implies

$$\hat{\theta} = \hat{\theta}^B - \lambda \left[\mathbb{E}\ (W(\boldsymbol{\Sigma}) \mid \text{data})\right]^{-1} \mathbf{1}.$$

$\frac{\partial f(\hat{\theta})}{\partial \lambda} = 0$ implies

$$\mathbf{1}'\hat{\theta} = 1.$$

Hence

$$\lambda = \frac{\mathbf{1}'\hat{\theta}^B - 1}{\mathbf{1}' \left[\mathbb{E}\ (W(\boldsymbol{\Sigma}) \mid \text{data})\right] \mathbf{1}}$$

so that the solution to the constrained optimization is given by $\hat{\theta}^{CB}$ given by

$$\begin{aligned} \hat{\theta}^{CB} &= \hat{\theta}^B - \frac{\left[\mathbb{E}\ (W(\boldsymbol{\Sigma}) \mid \text{data})\right]^{-1} \mathbf{1}}{\mathbf{1}' \left[\mathbb{E}\ (W(\boldsymbol{\Sigma}) \mid \text{data})\right]^{-1} \mathbf{1}} (\mathbf{1}'\hat{\theta}^B - 1) \\ &= \frac{\left[\mathbb{E}\ (W(\boldsymbol{\Sigma}) \mid \text{data})\right]^{-1} \mathbf{1}}{\mathbf{1}' \left[\mathbb{E}\ (W(\boldsymbol{\Sigma}) \mid \text{data})\right]^{-1} \mathbf{1}} \end{aligned} \tag{4.10}$$

We have shown that $\hat{\theta}^{CB}$ is a plug-in estimator with the constrained optimality property. In other words, when $W(\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$ we find that $\hat{\theta}^{CB}$ is obtained by replacing $\widehat{\boldsymbol{\Sigma}}_{\text{a}}$ in Eq. (4.4) with $\mathbb{E}\ [\boldsymbol{\Sigma} \mid \text{data}]$. Unlike other plug-in methods, this is an optimal method.

(3) Bayes Estimator under Unweighted Loss, $\hat{\theta}^U$

When $W(\boldsymbol{\Sigma}) = \mathbf{I}$, the method is under an unweighted loss. We show that the Unconstrained Bayesian Estimator under unweighted loss is equivalent to the Constrained Bayesian Estimator under unweighted loss.

To see this, note that in this case Eq. (4.8) becomes

$$\hat{\theta}^B = \mathbb{E}\left[\theta \mid \text{data}\right] \tag{4.11}$$

and

$$\mathbf{1}'\hat{\theta}^B = \mathbf{1}'\mathbb{E}\left[\frac{\mathbf{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{1}} \mid \text{data}\right] = \mathbb{E}\left[\frac{\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{1}} \mid \text{data}\right] = 1$$

For each iteration in the MCMC sampler, after the burn-in phase, we calculate the minimum variance allocation vector. We then average the iteration allocation vectors to find the allocation vector $\hat{\theta}^U$.

We investigate the Bayes estimator in greater detail. We have discussed the point estimator of the vector θ under this unweighted loss. The approach can give us a measure of accuracy as the matrix $\text{Var}(\theta \mid \text{data})$ or some summary measure such as trace of $\text{Var}(\theta \mid \text{data})$ or determinant of the matrix. We can also produce credible regions of θ. While this Bayes estimator, $\hat{\theta}^U$, is not optimal with respect to the weighted loss in Eq. (4.5), it is of interest to study the robustness of the performance of this Bayes estimator by comparing the perceived variance of the portfolio found under the method with the perceived variance of the portfolios found under other rival estimators.

4.1.4 Comparison of Methods

We evaluate the estimate of θ for each method "a", a $\in$ {SAM, CCC, SFM, POET, $\hat{\theta}^B$, $\hat{\theta}^{CB}$, $\hat{\theta}^U$}. We compare the following:

$$V_1 = \hat{\theta}'\mathbf{\Sigma}\hat{\theta} - \theta'\mathbf{\Sigma}\theta \tag{4.12}$$

$$V_2 = 1 - \hat{\theta}'\widehat{\mathbf{\Sigma}}\hat{\theta}/\hat{\theta}'\mathbf{\Sigma}\hat{\theta} \tag{4.13}$$

For plug-in estimators, $\hat{\theta}$ entirely depends on a plug-in estimate of $\mathbf{\Sigma}$. For the Unweighted Bayesian Estimator $\hat{\theta}^U$ (which is not a plug-in method) we obtain simultaneous credible regions for each parameter in the vector, allowing for dependence among pairs of elements in the parameter vector.

We calculate V_1 and V_2 for different estimators of θ for each replication generated and compute the average. This provides a summary of the replications (or expected value with respect to the model given $\mathbf{\Sigma}$).

Defining the Frobenius norm of a matrix $\mathbf{A}$ as $\|\mathbf{A}\|_F = \mathrm{tr}^{1/2}(\mathbf{A}'\mathbf{A})$, we evaluate the estimate of $\mathbf{\Sigma}$ under the following measurements:

1. Frobenius norm of the precision error matrix

$$M_1 = \|\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\|_F \tag{4.14}$$

2. Frobenius norm of the low rank matrix

$$M_2 = \|\mathrm{Cov}(\widehat{\mathbf{FB}}') - \mathrm{Cov}(\mathbf{FB}')\|_F \tag{4.15}$$

3. Frobenius norm of the variance of the idiosyncratic components

$$M_3 = \|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\|_F \tag{4.16}$$

Figure 4.1 illustrates how poorly the sample covariance serves as an estimate compared to the other methods. Note that $\widehat{\mathbf{\Sigma}}_{\mathrm{CCC}}$ does not assume a factor model and still well outperforms the $\widehat{\mathbf{\Sigma}}_{\mathrm{SAM}}$ on this measurement, for the levels of N considered. In all cases, M is fixed at 100.

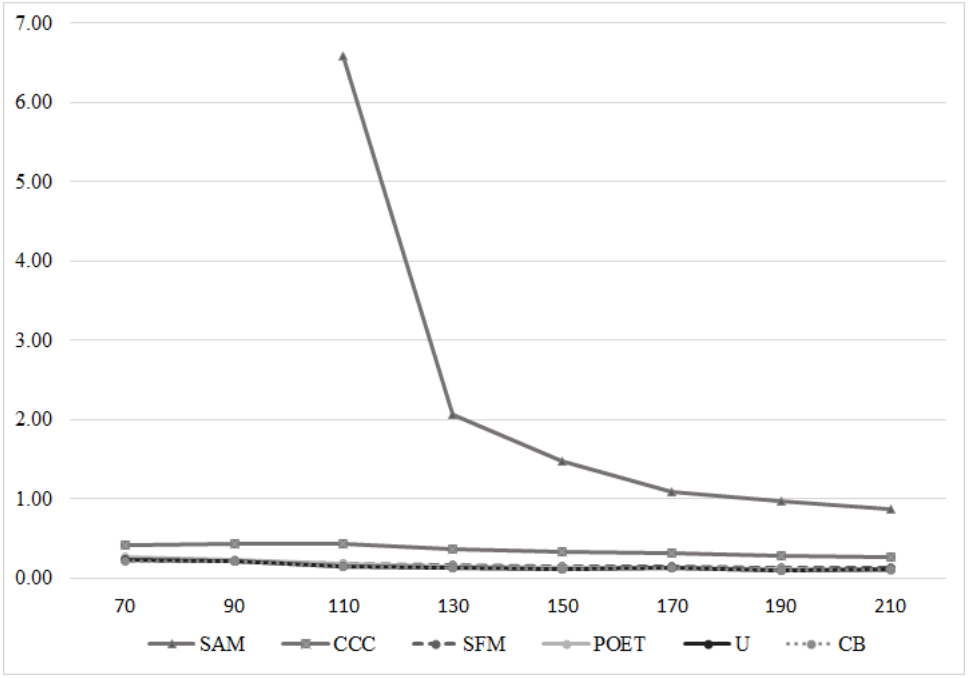

Figure 4.1: ***x-axis***: the number of days, N ***y-axis***: $\hat{\theta}'\boldsymbol{\Sigma}\hat{\theta} - \theta'\boldsymbol{\Sigma}\theta$

Figure 4.2 is similar but has the two methods with the highest values of V_1 removed, which were $\widehat{\boldsymbol{\Sigma}}_{\text{SAM}}$ and $\widehat{\boldsymbol{\Sigma}}_{\text{CCC}}$. It is clear that the four methods behave similarly, and decrease over the range of N considered. The POET method is superior to the SFM method for larger values of N. We see that the Bayes unweighted estimator $\hat{\theta}^U$ performs well, and is very similar to the constrained Bayes estimator $\hat{\theta}^{CB}$, however, it is rarely lower. The difference between the two Bayesian methods, $\hat{\theta}^U$ and $\hat{\theta}^{CB}$, and the POET method is larger when the sample size is small.

Figure 4.3 illustrates the performance of the method in comparison to the truth, given the selected portfolio. Again, the sample covariance is the worst method, and is only available for $N > M$. On this measurement, we see that the Bayes Unweighted Estimator $\hat{\theta}^U$ performs the best, only slightly better than the plug-in

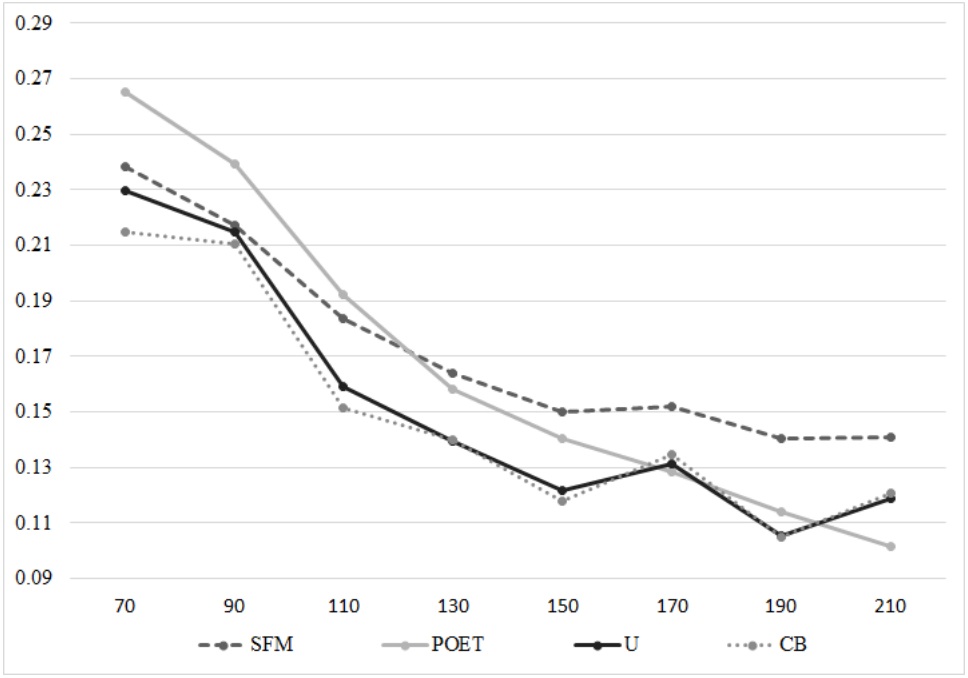

Figure 4.2: ***x-axis***: the number of days, N ***y-axis***: $\hat{\theta}'\boldsymbol{\Sigma}\hat{\theta} - \theta'\boldsymbol{\Sigma}\theta$

method $\hat{\theta}^{CB}$.

Figure 4.4 illustrates the behavior of three of the estimators on the Frobenius norm of the difference in the estimated and true precision matrix. We see that the B.POST method is superior under our model for the levels of N selected. This graph illustrates samples drawn under Design 2 whereby the factor loading matrix $\mathbf{B}$ has non-zero entries in at most two elements of each row. When $\mathbf{B}$ is more dense, then the POET method may result in a closer estimate.

Figures 4.5 and 4.6 serve to illuminate the performance of POET and B.POST on the two components of the covariance matrix $\boldsymbol{\Sigma}$ under the factor model. In particular, an advantage of the B.POST method may be an ability to better estimate the matrix of idiosyncratic variances $\boldsymbol{\Sigma}_u$. Unlike the SFM and POET method, the

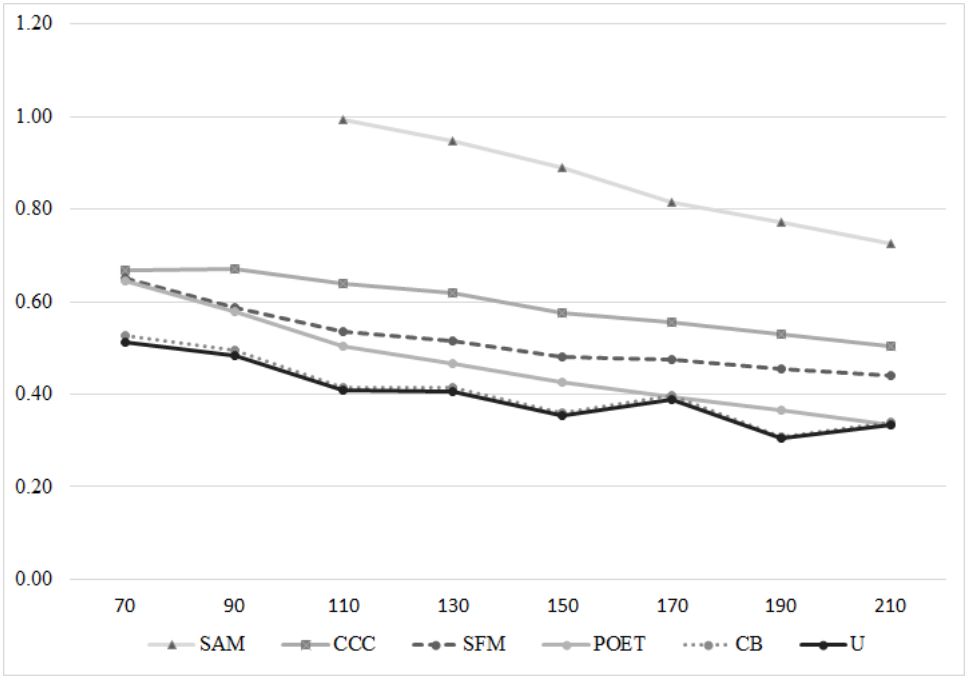

Figure 4.3: ***x-axis***: the number of days, N ***y-axis***: $1 - \hat{\theta}'\widehat{\mathbf{\Sigma}}\hat{\theta}/\hat{\theta}'\mathbf{\Sigma}\hat{\theta}$

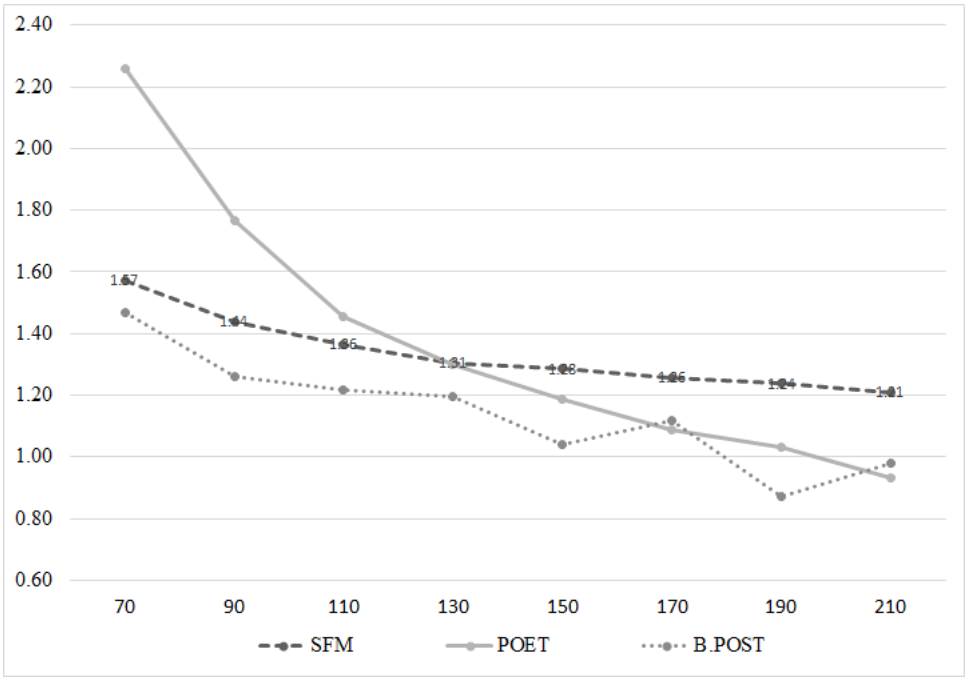

Figure 4.4: ***x-axis***: the number of days, N ***y-axis***: $\|\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\|_F$

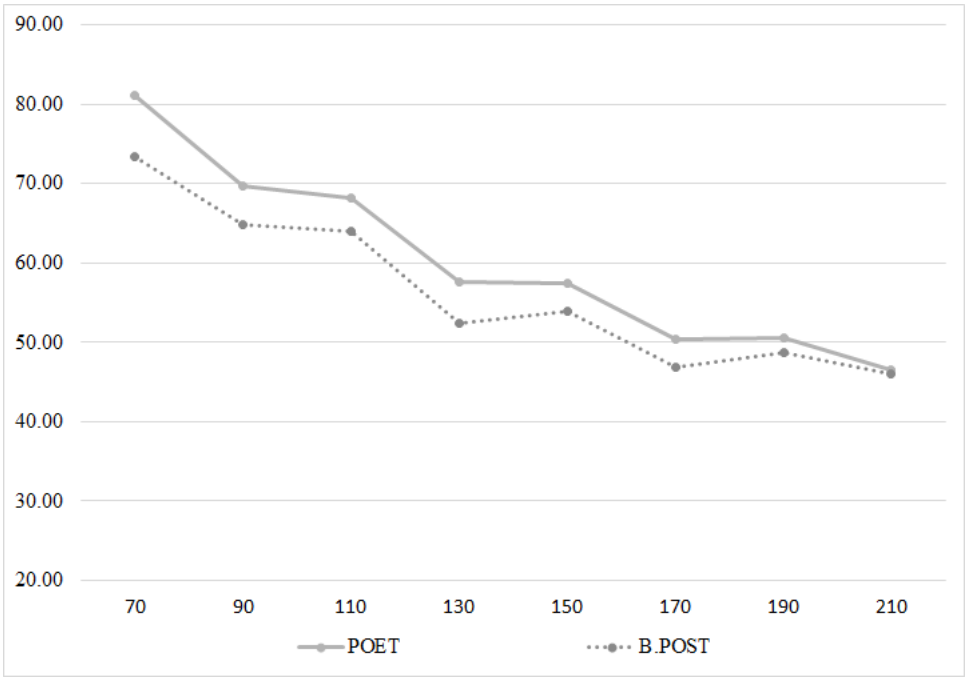

Figure 4.5: ***x-axis***: the number of days, N ***y-axis***: $\|\text{Cov}(\widehat{\mathbf{FB}}') - \text{Cov}(\mathbf{FB}')\|_F$

Bayesian approach enlists a separate model to estimate these parameters. The graph shows that even as N increases, the Bayesian model seems to maintain an improved estimate of this component of $\mathbf{\Sigma}$.

The Unweighted Bayesian Estimator, $\hat{\theta}^U$ allows the construction of credible regions for the vector parameter θ. We utilize a method presented in Held (2013) [19], in which simultaneous contour probabilities are constructed based on samples from a multivariate posterior distribution. Figures 4.7 and 4.8 show the credible regions constructed for one replication of the generated data, when $N = 110$ and $M = 100$. The regions are formed from the 20,000 iterations generated by the Bayesian model. Figure 4.8 limits the graph to only the first 50 assets to give a clearer view.

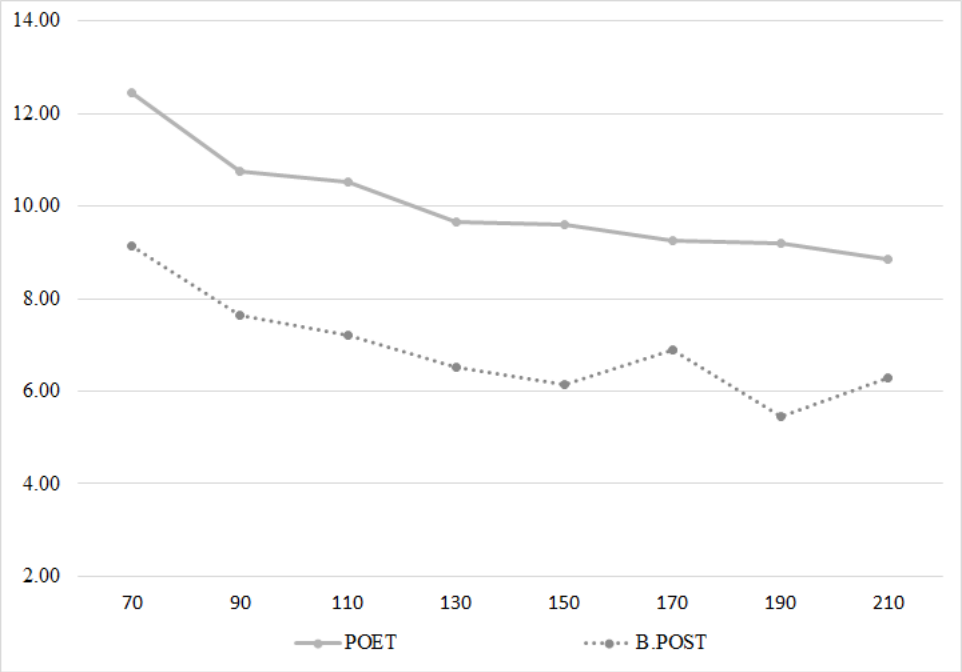

Figure 4.6: ***x-axis***: the number of days, N ***y-axis***: $\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\|_F$

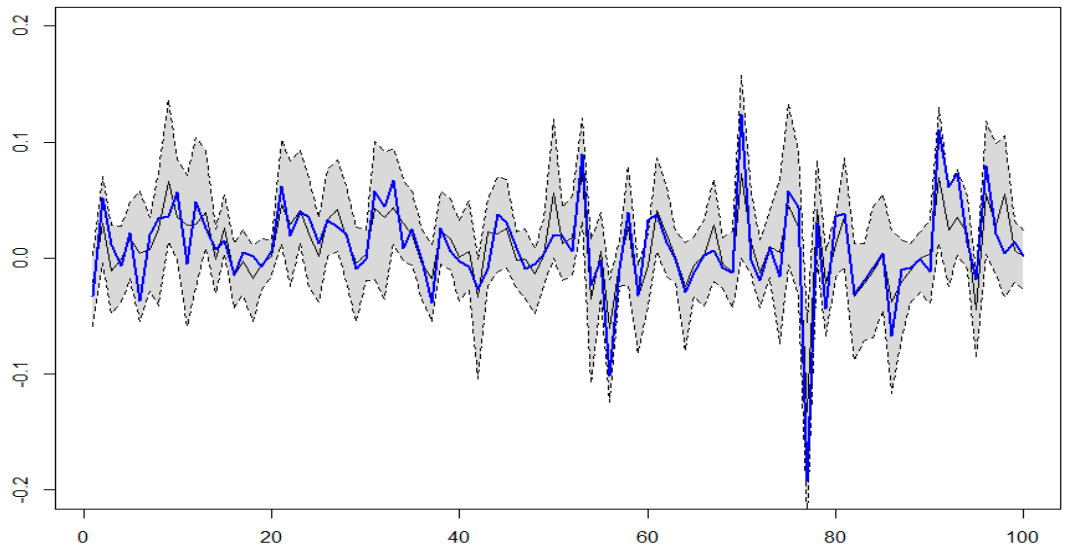

Figure 4.7: ***x-axis***: each asset in the portolio ***y-axis***: the percentage of the portfolio invested in the asset. The gray region indicates the credible region for θ using the method $\hat{\theta}^U$. The heavy line shows the true value. The black line shows the estimate $\hat{\theta}^U$.

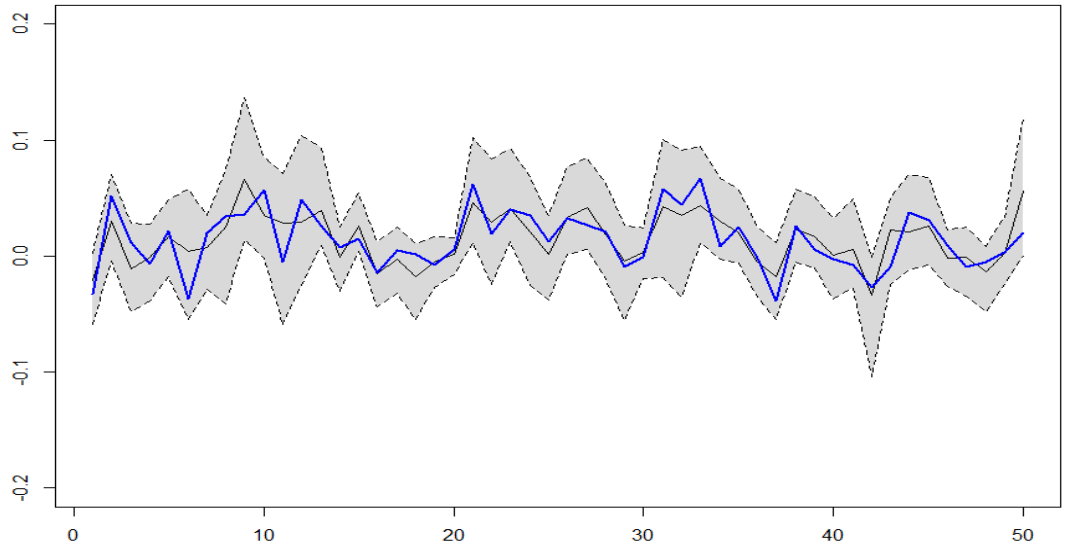

Figure 4.8: ***x-axis***: each asset in the portolio, showing 50 of the 100 assets ***y-axis***: the percentage of the portfolio invested in the asset. The gray region indicates the credible region for θ using the method $\hat{\theta}^U$. The heavy line shows the true value. The black line shows the estimate $\hat{\theta}^U$.

4.2 Real Data Example

4.2.1 Returns on Portfolio From Standard & Poor's 500 Index

One hundred companies are randomly selected from those included in the Standard & Poor's 500 stock market index, an index of more than 500 large American companies having common stock listed on the NYSE or NASDAQ. The daily returns for the period from January 1, 2007, through January 31, 2017, are reduced by the risk-free return to obtain the daily excess return. The risk-free return is assumed to be the daily return on the U.S. Department of the Treasury auctioned 13-week Treasury bills. The natural logarithms of the daily excess returns are utilized. Stock symbols, daily prices and Treasury bill indices are downloaded using the package "fportolio" in R [29].

On the first of each month, beginning with January, 2007, the minimum variance portfolio allocation θ is estimated under each of four methods, utilizing the subsequent 84 trading days, approximately four months of trading days. The procedure is repeated for the next 115 months, resulting in a total of 116 allocation vectors for each method: $\{\hat{\mathrm{w}}_{\mathrm{CCC}_r},\ \hat{\mathrm{w}}_{\mathrm{SFM}_r},\ \hat{\mathrm{w}}_{\mathrm{POET}_r},\ \hat{\mathrm{w}}_{\mathrm{B.POST}_r}\},\ r = 1, \ldots, 116$.

For the methods SFM and POET, the number of factors is estimated by using a method proposed by Bai and Ng (2002) [2], in which K is chosen to minimize a data-driven objective function with a penalty for increasing K, as described in Section (2.1.2). The number of factors for the 116 periods varied over the range including 2 and 7, with an average of 3.3. Three factors were selected.

For purposes of the POET method, the value of the threshold is determined using the cross-validation procedure described in Section 2.1.2. The cross-validation procedure indicates a thresholding constant of 0.5.

B.POST is calculated using 10,000 iterations and the same prior specifications as shown in Table 4.4 with a value of .1 for κ. The maximum number of factors was set for each period at four more than the number of factors suggested by the Bai and Ng method.

For each 84 day period, the subsequent 21 days (one month of trading days) is used for an out-of-sample test of the method. Implied actual risk of the four portfolios are calculated and compared. For each method "a", a $\in$ {CCC, SFM, POET, B.POST}, we obtain:

$$R_{\mathrm{a}_r} = \hat{\mathrm{w}}'_{\mathrm{a}_r} \mathrm{Cov}(y_{t_r}) \hat{\mathrm{w}}_{\mathrm{a}_r} \tag{4.17}$$

where y_{t_r} is the vector of the natural logarithm of the daily excess returns of the 100 selected companies on day t in period r, $t = 1, \ldots, 21,\ r = 1, \ldots, 116$.

4.2.2 Comparison of Estimates

This example relies on four key assumptions in order for Eq. (4.17) to be a reliable estimate of the actual risk of the portfolio, as noted in Scherer (2015) [31]. (1) The sample period is long enough to provide a reliable sample. (2) The serial correlation in the excess returns is stationary, and thus the factors and factor loadings can be written as a multivariate autocorrelation process. (3) The autocorrelations do not vary over the period. (4) The in-sample optimization applies to the

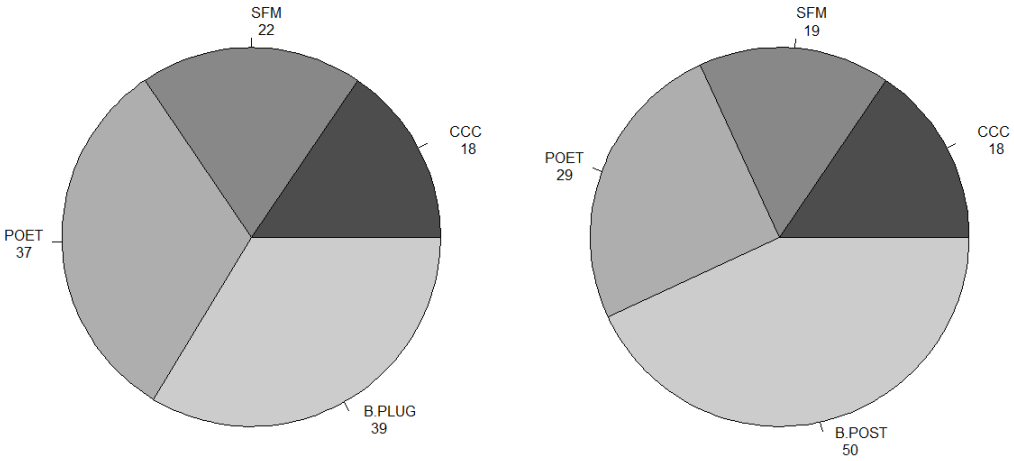

Figure 4.9: Number of periods when method resulted in lowest actual risk. Left side: Using the plug-in method as the Bayesian method. Right side: Using B.POST as the Bayesian method.

out-of-sample period. While it can be argued that none of the assumptions is met, the example is instructive considering the large advantage indicated by using the POET and B.POST methods over the SFM and CCC methods.

Figure 4.9 shows the number of periods when the estimator resulted in the portfolio allocation with the lowest actual risk. The methods that impose a factor structure result in a lower risk in all but 18 of the 116 periods. The results using "B.PLUG" are included to illustrate the advantage of "B.POST" over "B.PLUG".

The quantiles of the risk calculations for all periods are shown in Figure 4.10, which indicates that the estimates using the POET and B.POST methods have lower variability than the other two methods. Further, the method B.POST generates the lowest estimate, and the estimate with the lowest variability.

A comparison of the portfolios created under the B.POST and POET methods

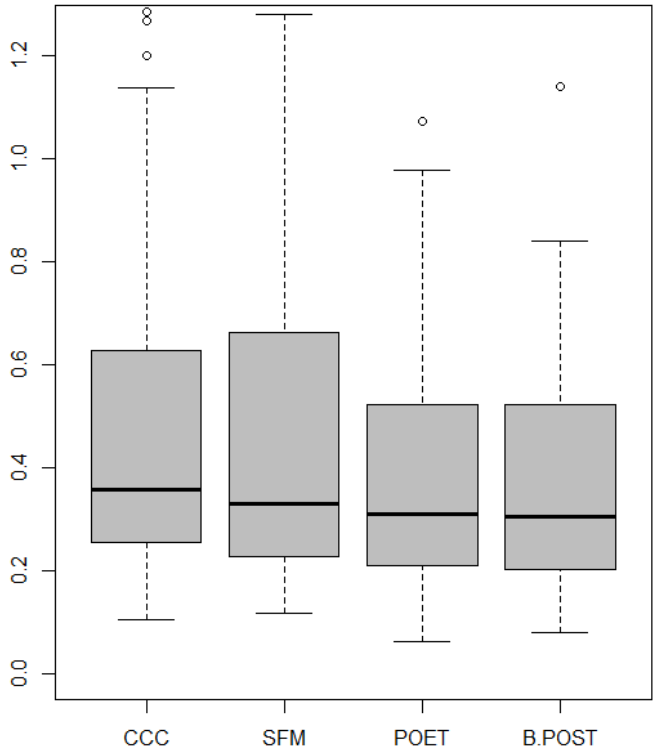

Figure 4.10: Quantiles of 116 estimates of actual risk under the methods.

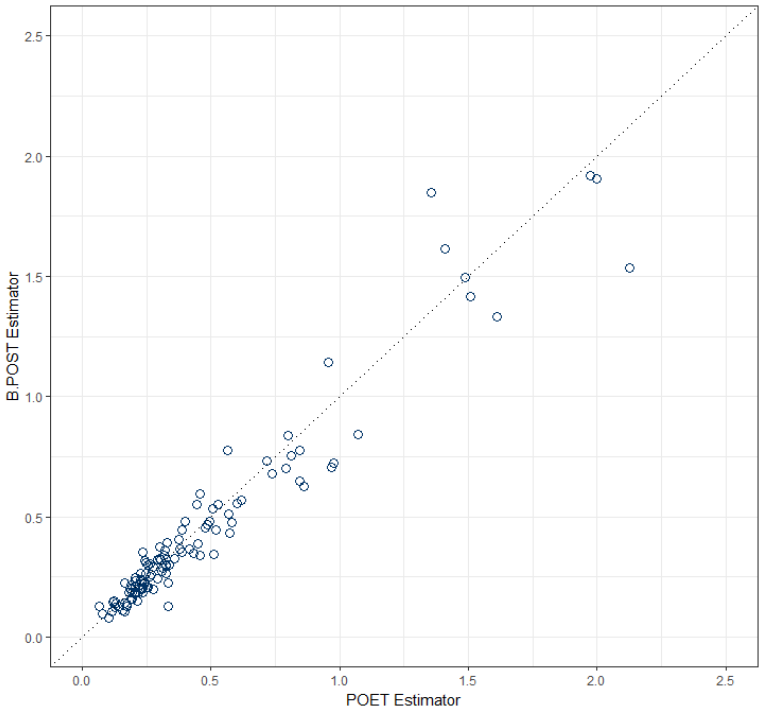

Figure 4.11: Risk of portfolios created with POET and B.POST.

reveals that the B.POST method results in a lower risk in 70 of the 116 periods, which is 60.3% of the periods. The total combined risk of the periods is only slightly lower under the B.POST method. Figure 4.11 illustrates that there is very little difference between the two methods in this example. The probability of having 70 or fewer successes according to a binomial distribution with a probability of success of .5 is lower than .01, so the difference can be considered significant. However, the advantage gained by using the B.POST method is small.

When the B.POST and SFM methods are compared, B.POST has a lower risk in 84 or 72% of the periods, and the gain is larger than the B.POST gain over POET. More of the points are more clearly in the lower right quadrant in Figure 4.12. The

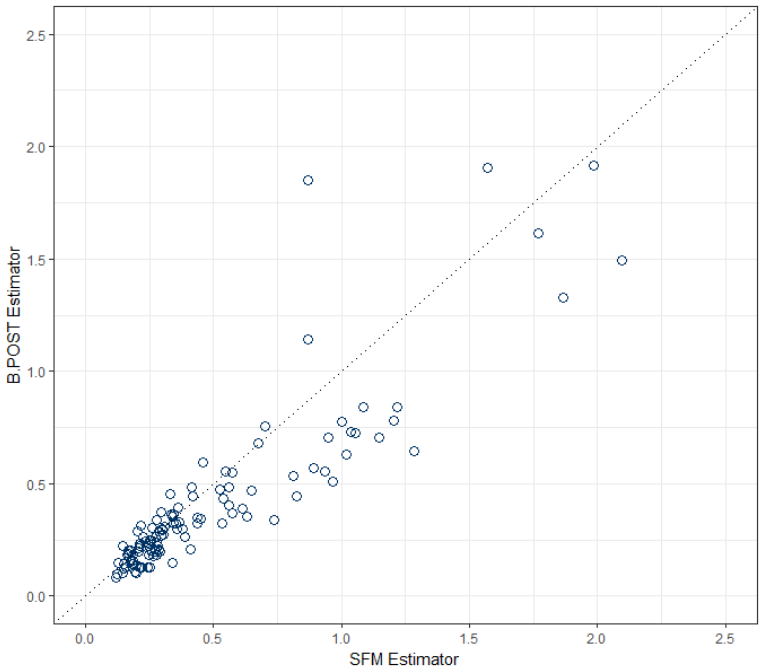

Figure 4.12: Risk of portfolios created with SFM and B.POST.

results of comparing the B.POST method to the CCC method are very similar, B.POST has a lower risk in 74% of the periods. In these cases the probability of achieving the results by random chance is very small.

We chose a different set of 100 stocks from the S&P 500 Index and found nearly identical results. Next, we examined the effect of changing K in the POET and SFM methods and the effect of changing the threshold constant in the POET method and found that the SFM and POET methods were optimized with the choices made.

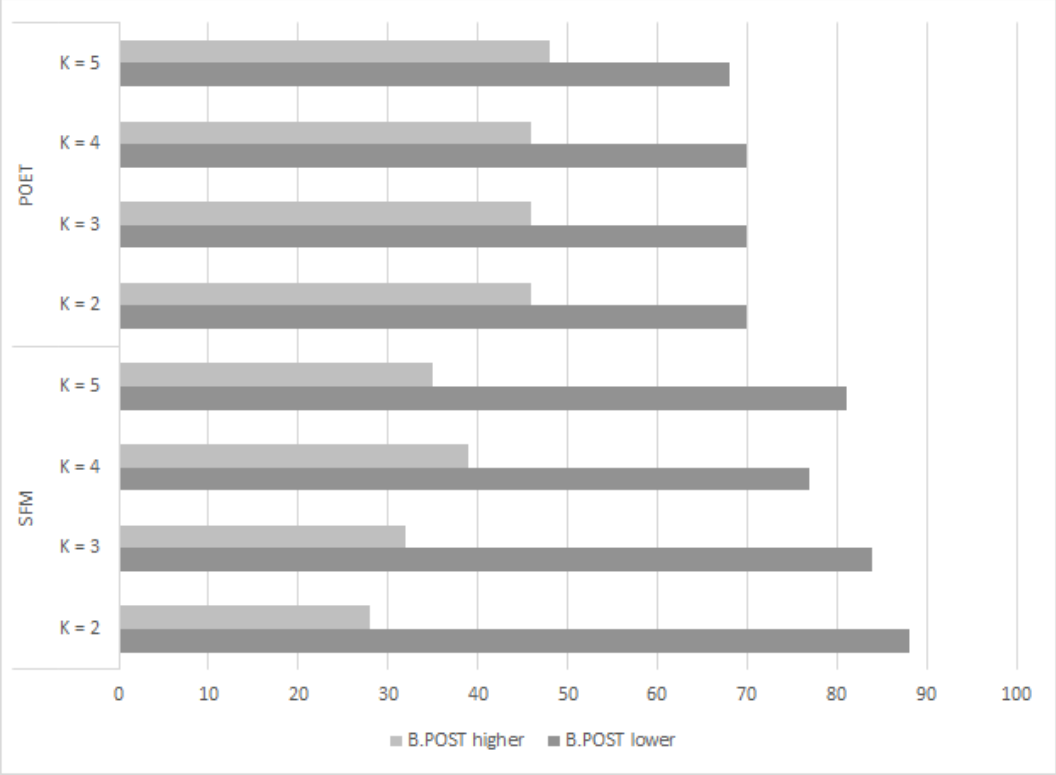

Figure 4.13: Risk of portfolios created with SFM and POET compared to B.POST using four values of K.

Chapter 5: Conclusion

Using Monte Carlo simulations and real data analysis, we showed that for small sample size, allocation estimates based on the sample covariance matrix can perform poorly in terms of the traditional measures used to evaluate an allocation for portfolio analysis. When the sample size is less than the dimension of the covariance matrix, we encountered difficulty even computing the allocation estimates due to the singularity of the sample covariance matrix. We evaluated a few classical estimators. Among them, the allocation estimator based on the well-known POET estimator, (see Fan *et al.* (2013) [15]), is developed using a factor model. While our simulation and data analysis support good behavior of POET for large sample size (consistent with the asymptotic theory), it did not perform well in small samples when compared to our proposed Bayesian estimator, $\hat{\theta}^U$.

In the case that a high-dimensional covariance matrix can be assumed to be sparse, conditional on pervasive factors, the POET estimator can take advantage of the structure. In practice, the strict factor model can be too restrictive and the POET method allows for the presence of cross-sectional correlation even after taking out the common factors.

Bayesian inference allows informative priors so that prior knowledge or results

of a previous model can be used to inform the current model. In cases where there is no prior knowledge of the number of factors, and no prior indication of sparsity, the Bayesian model can be used to explore and estimate a factor structure. Using both simulated data and the real data analysis, the Bayesian method was shown to have significant advantages over that approach traditionally used in the the applied literature.

A disadvantage to the POET model is that when components are extracted, error in the measurements is not considered. Component extraction is based on an eigenvalue decomposition of the raw correlation matrix, while factor extraction is applied on the "reduced" correlation matrix of the factor variance, ignoring the variance of the measurement error or idiosyncratic components. The Bayesian model explicitly models the variance of the idiosyncratic components $\mathbf{\Sigma}_u$.

The Bayesian model, as presented in Conti *et al.* (2014) [9] performs several steps in one process that must be performed sequentially in the POET method. It selects the dimensionality of the latent structure (the number of factors) and extracts the factors and performs rotation to search for a simple structure, where each measurement is dependent on at most one factor. The Bayesian procedure estimates the dimension jointly with the allocation of the measurements to the factors and the estimation of corresponding factor loadings.

Although the Bayesian model requires some a priori judgments, the process is then completed with no further input. In comparison, both principal components analysis and factor analysis requires choices by analysts at several sequential steps in the process.

On simulated data, we observed scenarios in which the B.POST method resulted in the most accurate estimate, and performed especially well relative to the other estimators when the sample size N was small. Using the real data, we found that the B.POST method outperformed the others in many of the periods. Unlike the method using principal components, the B.POST method has an explicit noise model and may have the advantage of being better able to distinguish between the underlying latent factors and the idiosyncratic components. Furthermore, utilizing the posterior distribution of the quantity of interest has advantages over a plug-in method where the estimate of the covariance is used in the function.

A constrained Bayes estimator of the allocation vector $\hat{\theta}^{CB}$ was proposed that is the best in terms of the posterior risk under a given prior among all estimators that satisfy the constraint. In this sense, it is superior to all classical plug-in estimators, including POET and the proposed Bayesian estimator. We compared the proposed Bayesian method with the constrained Bayes using the traditional evaluation measures used in portfolio analysis and found that they show similar behavior. In addition to point estimation, the proposed Bayesian approach yields a straightforward measure of uncertainty of the estimate and allows construction of credible intervals for a wide range of parameters. We demonstrated the construction of credible intervals on the vector parameter.

Encouraged by the impressive results of our proposed Bayesian method, we would like to investigate whether the method can be further improved by other choices of the prior distribution for the parameters of the factor model. This could lead to interesting results related to the propriety of the resulting posterior distribu-

tion, identifiability issues and the challenging problem of developing fast algorithms to solve the problem. In our simulation and data analysis, the constrained Bayesian method is very competitive to our proposed Bayesian method in terms of traditional criterion used in portfolio analysis. However, the methodology for measuring its uncertainty and the related problem of simultaneous interval estimation does not seem to be a straightforward problem. In the future, we hope to address these challenging issues.

Bibliography

[1] T.W. Anderson and H. Rubin, "Statistical Inference in Factor Analysis," Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. 5, University of California Press, Berkeley, 111-115 (1956).

[2] J. Bai, "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135-171 (2003).

[3] J. Bai and S. Ng, "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191-221, (2002).

[4] J. Barnard, R.E. McCulloch, and X.-L. Meng, "Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, with Applications to Shrinkage," *Statistica Sinica*, 10, 1281-1311 (2000).

[5] P. Bickel and E. Levina, "Covariance Regularization by Thresholding," *Annals of Statistics*, 36, 2577-2604 (2008).

[6] T. Bodnar, N. Parolya, and W. Schmid, "Estimation of the Global Minimum Variance Portfolio in High Dimensions," *European Journal of Operational Research*, 266, 371-390, (2018).

[7] T. Cai and W. Liu, "Adaptive Thresholding for Sparse Covariance Matrix Estimation," *Journal of the American Statistical Assocation*, 106, 672-684 (2011).

[8] G. Chamberlain and M. Rothschild, "Arbitrage, Factor Structure and Mean-Variance Analysis in Large Asset Markets," *Journal of Econometrics*, 51, 1305-1324 (1983).

[9] G. Conti, S. Frühwirth-Schnatter, J.J. Heckman and R. Piatek, "Bayesian Exploratory Factor Analysis," *Journal of Econometrics*, 183(1), 31-57 (2014).

[10] C. Davis and W. Kahan, "The Rotation of Eigenvectors by a Perturbation III," *SIAM Journal of Numerical Analysis*, 7, 1-46, (1970).

[11] A.P. Dempster, "Covariance Selection," *Biometrics*, 28, 157-175 (1972).

[12] B. Efron and C. Morris, "Stein's Paradox in Statistics," *Scientific American*, 236(5), 119-127 (1977).

[13] E. Fama and K. French, "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics*, 33, 3-56 (1993).

[14] J. Fan, Y. Fan, and J. Lv, "High Dimensional Covariance Matrix Estimation Using a Factor Model," *Journal of Econometrics*, 147, 186-197 (2008).

[15] J. Fan, Y. Liao, and M. Mincheva, "Large Covariance Estimation by Thresholding Principal Orthogonal Complements," *Journal of the Royal Statistical Society Series B*, 75, 603-680 (2013).

[16] J. Fan, Y. Liao, and M. Mincheva, "POET: Principal Orthogonal ComplEment Thresholding Method," R package version 2.0, https://CRAN.R-project.org/package=POET, (2016).

[17] J. Fan, J Zhang, and K. Yu, "Vast Portfolio Selection With Gross-Exposure Constraints," *Journal of the American Statistical Association*, 107:498, 592-606 (2012).

[18] S. Frühwirth-Schnatter and H.F. Lopes, "Parsimonious Bayesian Factor Analysis When the Number of Factors is Unknown," Working Paper, University of Chicago Booth School of Business (2010).

[19] L. Held, "Simultaneous Posterior Probability Statements from Monte Carlo Output," *Journal of Computational and Graphical Statistics*, 13, 20-35 (2004).

[20] H.B. Heywood, "On Finite Sequences of Real Numbers," *Proceedings of the Royal Society A; Math. Phys. Eng. Sciences"*, 134, 486-501 (1931).

[21] A. Huang and M.P. Wand, "Simple Marginally Noninformative Prior Distributions for Covariance Matrices," *Bayesian Analysis*, 8, 439-452 (2013).

[22] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis* (Prentice Hall, Englewood Cliffs, New Jersey, 1992).

[23] O. Ledoit and M. Wolf, "Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection," *Journal of Empirical Finance*, 10, 603-621 (2003).

[24] O. Ledoit and M. Wolf, "Honey, I Shrunk the Sample Covariance Matrix," *Journal of Portfolio Management*, 30, 4, 110-119 (2004).

[25] H.F. Lopes and M. West, "Bayesian Model Assessment in Factor Analysis," *Statistica Sinica*, 14, 41-67 (2004).

[26] H.M. Markowitz, "Portfolio Selection," *Journal of Finance*, 7, 77-91 (1952)

[27] R. Michaud, "The Markowitz Optimization Enigma: Is Optimized Optimal?" *Financial Analysts Journal*, 45, 31-42 (1989).

[28] R. Piatek, "BayesFM: Bayesian Inference for Factor Modeling," R package version 0.1.2, https://CRAN.R-project.org/package=BayesFM, (2017).

[29] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org, (2017).

[30] C.P. Robert and G. Casella, *Monte Carlo Statistical Methods* (Springer, New York, 2000).

[31] B. Scherer, *Portfolio Construction and Risk Budgeting* (5th Edition, Risk Books, London, 2015).

[32] W. Sharpe, "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," *Journal of Finance*, 19, 425-442 (1964).

[33] J. Sherman, and W. Morrison, "Adjustment of an Inverse Matrix Corresponding to Changes in the Elements of a Given Column or a Given Row of the Original Matrix," *Annals of Mathematical Statistics*, 20: 621 (1949).

[34] L.L. Thurstone, *Multiple Factor Analysis* (University of Chicago Press, Chicago, 1947).

[35] W.F. Velicer, "Determining the Number of Components from the Matrix of Partial Correlations," *Psychometrika*, 41, 321-327 (1976).

[36] H. Weyl, "Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen," *Math. Ann.*, 71, 441-479, (1912).

[37] X. Zhang, W.J. Boscardin, and T.R. Belin, "Sampling Correlation Matrices in Bayesian Models with Correlated Latent Variables," *Journal of Computational and Graphical Statistics*, 15(4), 880-896 (2006).