ANNALS OF ECONOMICS AND FINANCE **12-2**, 199–215 (2011)

Estimating High Dimensional Covariance Matrices and its Applications

Jushan Bai*

Department of Economics, Columbia University, New York, NY 10027
CEMA, the Central University of Finance and Economics, Beijing, China

and

Shuzhong Shi

Department of Finance, Guanghua School of Management, Beijing, China

Estimating covariance matrices is an important part of portfolio selection, risk management, and asset pricing. This paper reviews the recent development in estimating high dimensional covariance matrices, where the number of variables can be greater than the number of observations. The limitations of the sample covariance matrix are discussed. Several new approaches are presented, including the shrinkage method, the observable and latent factor method, the Bayesian approach, and the random matrix theory approach. For each method, the construction of covariance matrices is given. The relationships among these methods are discussed.

Key Words: Factor analysis; Principal components; Singular value decomposition; Random matrix theory; Empirical Bayes; Shrinkage method; Optimal portfolios; CAPM; APT; GMM.

JEL Classification Numbers: C33, C38.

1. INTRODUCTION

Estimating covariance matrices is an important part of portfolio selection, risk management, and asset pricing. The sample covariance matrix is often used for these purposes, but the sample covariance matrix has a number of undesirable properties when the dimension of the matrix is large.

*Financial support from the NSF (grants SES-0551275 and SES-0962410) is acknowledged.

1529-7373/2011

First, when the number of assets (N) is larger than the number of observations (T), the sample covariance matrix is not of full rank, so its inverse will not exist. Second, even if the sample covariance matrix is invertible, the expected value of its inverse is a biased estimator for the theoretical inverse. Third, the sample covariance can be volatile in the sense that the constructed weights for the mean-variance efficient portfolios may give rise to high turnover rates over time. Also, the out-of-sample portfolio risks usually far exceed the desired risks. In this paper, we review some of the new methodologies that overcome these deficiencies. These new methods are classified into four categories (not necessarily mutually exclusive). They include

(i) The shrinkage method. The shrinkage estimator is a linear combination of the sample estimator and another estimator. The latter can be the covariance matrix implied by the CAPM theory. How to determine the optimal shrinkage will be discussed.

(ii) Factor models. We consider covariance matrices implied by the large dimensional factor models, either observable or latent factor models. For the latter, we discuss the principal components method and the maximum likelihood method.

(iii) Bayesian and empirical Bayes estimators. These estimators are related to the shrinkage estimator. They provide alternative interpretations for the shrinkage method.

(iv) The method based on random matrix theory. This method aims to attenuate the randomness of the sample covariance S using the theory of random matrices of high dimension.

In portfolio selection, the inverse matrix is needed. Bruce-force inversion of a large dimensional matrix can be difficult and inaccurate. However, taking into account the structure of the proposed estimators, inversion can be easily performed. We also discuss the merit of each estimator in terms of the easiness of finding the inverse.

The emphasis is on issues arising from a high cross-sectional dimension. The methods are suitable under the assumption that the number of variables (assets) goes to infinity in contrast to the usual assumption that the number of observations goes to infinity. The large N asymptotics provides a good approximation for emerging markets financial data, where the time series dimension is small. Even as time goes by, the large N relative to T environment is likely to persist as a result of emergence of new firms, mergers and acquisitions. At any point in time, the number of firms with a long history may be small. To include as many firms as possible, we must be content with data sets having short time spans. Also, it may be desirable to use more recent data. In any case, the methods presented below also work well under large T.

2. THE SAMPLE COVARIANCE

Let $X_t = (X_{1t}, X_{2t}, ..., X_{Nt})'$ be an $N \times 1$ vector of random variables. For example, X_{it} can be the return for asset i in period t, $t = 1, 2, ..., T$. In the following, N is referred to as the number of variables, or the number of series, and T is referred to as the number of observations, or the sample size. Suppose $E(X_t) = \mu$ and $E[(X_t - \mu)(X_t - \mu)'] = \Sigma$. We assume Σ has a full rank N. The sample mean and sample covariance are defined respectively as

$$\bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t \text{ and } S = \frac{1}{T-1}\sum_{t=1}^{T}(X_t - \bar{X})(X_t - \bar{X})'.$$

The sample covariance S is a natural estimator for the population covariance Σ. The estimator has a number of advantages: simple to construct, unbiased, and intuitively appealing as it is the sample analogue of the theoretical central moment. Unbiasedness means its expected value is equal to the true covariance matrix, that is, $E(S) = \Sigma$. However, the sample covariance matrix has a number of disadvantages. When the number of observations (T) is less than the number of variables (N), the rank of S is at most T, so it is not invertible, even though the underlying true covariance matrix is invertible. Even when T is comparable to or larger than N, the sample covariance S has a significant amount of sampling error, and its inverse is a poor estimator for Σ^{-1}. For example, under normality assumption, the expected value of the inverse

$$E(S^{-1}) = \frac{T}{T-N-2}\Sigma^{-1}.$$

While S is unbiased for Σ, S^{-1} is highly biased for Σ^{-1} if N is close to T. In particular, for $N = T/2 + 2$, we have $E(S^{-1}) = 2\Sigma^{-1}$. It is possible, however, to directly estimate the inverse of Σ^{-1}, as in Fan et al. (2008).

These undesirable properties of S have led to many alternative and improved estimators. In the sections to follow, we review recent advances in this area. All these estimators can be viewed as some kind of shrinkage estimators, differing in the targets to which the sample covariance matrix is shrunk. The target matrices are considered to have some structures associated with some statistical or economic theory. For example, the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965) implies a simple factor structure. In general, a structured matrix has much fewer parameters to estimate than the unstructured matrix Σ, and therefore can be easily estimated with little estimation error. But a structured matrix can be highly biased if the underlying theory governing the structure is in-

correct. The shrinkage estimator seeks an optimal trade-off between biases and estimating variability.

For any matrix A, $n \times n$, we define its norm as $\|A\| = (\sum_i \sum_j a_{ij}^2)^{1/2}$. This is the standard Euclidean norm when A is viewed as an element in the n^2 dimensional Euclidean space.

3. SHRINKAGE ESTIMATOR

In this section, we consider shrinkage estimators in the context of asset returns, particularly the estimator proposed by Ledoit and Wolf (2003). For asset returns, there exists a natural target toward which the covariance matrix can be shrunk. Sharpe (1963)'s single index model postulates

$$X_{it} = \alpha_i + \beta_i X_{0t} + \varepsilon_{it}$$

where X_{it} is the stock i's return, X_{0t} is the market return, and ε_{it} is idiosyncratic return for stock i in period t and is uncorrelated with the market return. This implies the covariance matrix

$$\Phi = \beta\beta'\sigma_{00}^2 + \Omega_\epsilon$$

where $\beta = (\beta_1, ..., \beta_N)'$ is $N \times 1$, and σ_{00}^2 is the variance of the market portfolio. An estimator for Φ is

$$\hat{\Phi} = BB'\hat{\sigma}_{00}^2 + \hat{\Omega}_\epsilon$$

where $B = (b_1, ..., b_N)'$ and b_i is the least squares estimator for β_i, and $\hat{\Omega}_\epsilon = \text{diag}(\hat{\sigma}_{1,\varepsilon}^2, ..., \hat{\sigma}_{N,\varepsilon}^2)$, and each $\hat{\sigma}_{i,\varepsilon}^2$ is based on the OLS residuals. More specifically,

$$b_i = (\sum_{t=1}^{T} X_{0t}^2)^{-1} \sum_{t=1}^{T} X_{0t}X_{it}, \quad (i = 1, 2, ..., N),$$

$$\hat{\sigma}_i^2 = \frac{1}{T-1} \sum_{t=1}^{T} \hat{\epsilon}_{it}^2$$

where $\hat{\epsilon}_{it}$ is the regression residual

$$\hat{\epsilon}_{it} = X_{it} - b_i X_{0t}.$$

Finally, $\hat{\sigma}_{00}^2$ is the sample variance of the market returns.

One shrinkage estimator proposed by Ledoit and Wolf (2003) is

$$\hat{\Sigma}(\alpha) = \alpha\hat{\Phi} + (1-\alpha)S$$

a linear combination of $\hat{\Phi}$ and S, where $\hat{\Phi}$ is defined earlier. This estimator shrinks S toward the covariance matrix implies by the CAPM model. To derive the optimal shrinkage intensity α, they considered the following mean squared error criterion

$$L(\alpha) = E\|\hat{\Sigma}(\alpha) - \Sigma)\|^2.$$

Solving from the first order condition, the optimal solution for α that minimizes the mean squared error loss is

$$\alpha^* = \frac{\sum_{i=1}^N \sum_{j=1}^N [\text{var}(s_{ij}) - \text{cov}(\hat{\phi}_{ij}, s_{ij})]}{\sum_{i=1}^N \sum_{j=1}^N [\text{var}(\hat{\phi}_{ij} - s_{ij}) + (\phi_{ij} - \sigma_{ij})^2]}$$

where s_{ij} is the (i,j)th element in S, and all other entries are similarly defined. The value of α^* must be estimated. A consistent estimator $\hat{\alpha}^*$ for α^* is derived in Ledoit and Wolf (2003), the details are omitted. The final estimator for Σ is given by

$$\hat{\Sigma} = \hat{\alpha}^* \hat{\Phi} + (1 - \hat{\alpha}^*) S.$$

In a separate study, Ledoit and Wold (2004) considered an estimator that shrinks the sample variance S toward the identity matrix, and show the resulting matrix possesses a well behaved conditional number (the ratio of the maximum eigenvalue to the smallest eigenvalue). For portfolio selection, shrinking toward the market index model is intuitively appealing.

4. OBSERVABLE FACTOR MODELS

The CAPM model is a single factor model. When the market portfolio is proxied by the valued-weighted or equal-weighted index, the model becomes an observable factor model. This single index model can be easily extended to multiple factors:

$$X_{it} = \mu_i + \beta_{i1} Z_{1t} + \cdots \beta_{ik} Z_{kt} + \epsilon_{it} \tag{1}$$

$$i = 1, 2, ..., N; t = 1, ..., T$$

where $Z_t = (Z_{1t}, ..., Z_{kt})'$ is an observable vector. Chen, Roll, and Ross (1986) used macroeconomic variables as factors, for example, inflation, output growth gap, interest rate, risk premia, and term premia. The factors by Fama and French (1993) are portfolios based on firm characteristics. Lettau and Ludvigson (2001) use cointegration residuals (from a regression of consumption on income and wealth) as observable factors. Also see Gao

and Huang (2008). The BARRA's risk model includes a host number of industrial dummy variables, as well as other observable factors. Campbell, Lo, and MacKinlay (1997) provide a more extensive review on the topic.

Under model (1), the covariance matrix will be

$$\Sigma = \beta \Omega_Z \beta' + \Omega_\epsilon \tag{2}$$

where β is an $N \times k$ matrix consisting of the coefficients β_{ij} and Ω_Z is a $k \times k$ covariance matrix for the vector Z_t, and Ω_ϵ is diagonal and is the variance matrix of ϵ. The covariance matrix Σ depends on unknown coefficients, but they can be easily estimated by the least squares method equation by equation. Finally, analogous to the market model, we have

$$\hat{\Sigma} = \hat{\beta} \hat{\Omega}_Z \hat{\beta}' + \hat{\Omega}_\epsilon$$

where $\hat{\Omega}_Z$ is the sample covariance matrix of $Z_1, ..., Z_T$, and $\hat{\Omega}_\epsilon$ is an estimate of Ω, consisting of the residual variances. The theoretical properties of the estimator is studied by Fan et al. (2008). One advantage of observable factor models is that the model requires much fewer parameters than the latent factor models to be discussed below.

5. LATENT FACTOR MODELS

Factor models have both theoretical and empirical appeals. The single index model of Sharpe (1994) and Lintner (1965) is derived from an equilibrium consideration. The arbitrage pricing theory of Ross (1976) assumes asset returns have a factor structure so that risk premia can be expressed as a linear function of factor loadings. In addition to finance, factor models have been widely used in economics because factor models provide an efficient way to aggregate and synthesize information for large data sets. Bai and Ng (2008) provide more detailed discussion on the models' use in economics.

The previous section assumes observable factors, an ideal but not necessarily feasible assumption. Latent factor models relax this assumption and can be expressed as:

$$X_{it} = \mu_i + \lambda_i' f_t + \varepsilon_{it}$$

where both the factors f_t $(r \times 1)$ and factor loadings λ_i $(r \times 1)$ are unobservable. Here r represents the number of factors, which is also unknown. In vector form

$$X_t = \mu + \Lambda f_t + \varepsilon_t$$

where $X = (X_{i1}, ..., X_{iN})'$ and $\Lambda = (\lambda_1, ..., \lambda_N)'$; μ and ε_t are similarly defined. The implied covariance matrix is

$$\Sigma = \Lambda \Omega_f \Lambda' + \Omega_\epsilon$$

with $\Omega_f = \text{var}(f_t)$ and $\Omega_\epsilon = \text{var}(\varepsilon_t)$. Because both λ_i and f_t are unobservable and they enter the model in a multiplicative way, they cannot be identified separately without restrictions. This follows from the sample fact that $\alpha_i' f_t = \alpha_i' A A^{-1} f_t$ for an arbitrary invertible matrix. So normalization is made such that $\Omega_f = I$, implying $\Sigma = \Lambda\Lambda' + \Omega_\epsilon$. If Ω_ϵ is non-diagonal but its maximum eigenvalue is bounded, then the model is known as an approximate factor model, see Chamberlain and Rothschild (1983). Here we assume Ω_ϵ is diagonal. The model may be estimated by the maximum likelihood method, e.g., Lawley and Maxwell (1971) and Anderson (1984). The properties of the maximum likelihood estimator, under large N, is studied by Bai and Li (2010).

An alternative and simpler estimation method is that of the principal components. The principal components estimator of Σ makes use of the matrix spectral decomposition:

$$S = \sum_{i=1}^{N} b_i^2 h_i h_i'$$

where b_i^2 is the ith largest eigenvalue of S and h_i is the corresponding eigenvector. The above decomposition can be easily computed via the singular value decomposition. The estimator for Λ is defined as

$$\hat{\Lambda} = (b_1 h_1, ..., b_r h_r)$$

and the estimator of Ω_ϵ is defined as

$$\hat{\Omega}_\epsilon = \text{diag}(S - \hat{\Lambda}\hat{\Lambda}')$$

This gives

$$\hat{\Sigma} = \hat{\Lambda}\hat{\Lambda}' + \hat{\Omega}_\epsilon. \tag{3}$$

Connor and Korajczyk (1986, 1988) and Stock and Watson (2002), Bai and Ng (2002, 2011), and Bai (2003) studied the theoretical properties of the principal components estimators. Comparison between the principal components method and the maximum likelihood method, in terms of relative efficiency, is given in Bai and Li (2010). An application of latent factor models to testing the arbitrage pricing theory (APT) is performed by Lehmann and Modest (1988). Also using a latent factor framework,

Harvey, Solnik, and Zhou (1988) study the determinants of expected international asset returns.

The principal components method can also be applied to the sample correlation matrix. Let D be the $N \times N$ diagonal matrix formed by the diagonal elements of S. That is, $D = \text{diag}(S_{11}, ..., S_{NN})$. Define $C = D^{-1/2}SD^{-1/2}$, so that C is the sample correlation matrix. By the spectral decomposition, C can be written as

$$C = \tau_1\xi_1\xi_1' + \cdots \tau_k\xi_k\xi_k' + \cdots + \tau_N\xi_N\xi_N'$$

where $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_N$ are the eigenvalues of C and $\{\xi_k\}$ are the corresponding eigenvectors. Consider the "reduced" correlation matrix

$$\overline{C} = \sum_{i=1}^{r} \tau_i\xi_i\xi_i' + \text{diag}\Big(I_N - \sum_{i=1}^{r} \tau_i\xi_i\xi_i'\Big)$$

This is analogous to (3), except that the sample correlation matrix C is used instead of the sample covariance matrix S. Note that $\overline{C}$ is a correlation matrix because it is positive definite and the diagonal elements are all 1. Finally, the covariance matrix estimator is defined as

$$\hat{\Sigma}_c = D^{1/2}\overline{C}D^{1/2}. \tag{4}$$

This estimator takes into account heteroskedasticity over the cross section. If heteroskedasticity is heavy, this estimator will perform better than the principal components estimator directly applied to the sample covariance matrix. This estimator is more closely examined in Bai (2010). Also, see Jones (2001).

In practice, the number of factors r is unknown and has to be estimated. Bai and Ng (2002) propose information criteria to estimate r and establish consistency. Bai (2003, 2004) shows the Λ and $(f_1, f_2, ..., f_T)$, up to a rotation, can be consistently estimated and derives the rates of convergence and the limiting distributions.

Factor models in which f_t and ϵ_{it} are GARCH processes can also be considered. Related issues can be found in Bollerslev (1987), Engle (2002), Engle and Kroner (1995), Engle et al. (1990), and Engle and Sheppard (2001), Ledoit et al. (2003), and Tsay (2002). A generalized dynamic factor model is studied by Forni et al. (2000).

Bayesian estimation of factor models is considered by Chib et al. (2002), Han (2003), Nardari and Scruggs (2003). They also allow the disturbances to have stochastic volatility. The Bayesian analysis to be discussed in the next section does not impose a factor structure, but can incorporate a factor structure as prior information.

It is also feasible to shrink the sample covariance S towards the principle component estimator, as in Bengtsson and Holst (2003), along the line of Ledoit and Wolf (2003).

It is possible to impose many restrictions on the factor loading matrix Λ. One particular class of restrictions corresponds to a structure with global (market wide) and regional (industrial and sectoral) factors. In this case, the loading matrix Λ is of the form (assuming three regions, for example):

$$\Lambda = \begin{bmatrix} G_1 & \Gamma_1 & 0 & 0 \\ G_2 & 0 & \Gamma_2 & 0 \\ G_3 & 0 & 0 & \Gamma_3 \end{bmatrix}$$

where G_i and Γ_i are all block matrices. Swartz (2006) provides an empirical application of this model and Wang (2008) considers identification and estimation of the model.

6. BAYESIAN AND EMPIRICAL BAYE'S ESTIMATORS

The sample covariance S is an estimator based solely on data. In contrast, Bayesian methods incorporate prior information concerning Σ, be either personal beliefs, historical experience, and or some modeling theory that governs Σ. Under the APT theory, for example, Σ has a factor structural. Bayesian method allows us to take into account these considerations. Statistically, every parameter under the Bayesian method is considered a random variable. Prior information about unknown parameters is represented by distributions.

We assume the $N \times 1$ vector X_t is normally distributed with mean μ and Σ. Under Bayesian framework, μ and Σ are random variables, and therefore, μ and Σ are regarded as the conditional mean and conditional variance, respectively. We write the conditional distribution of X_t conditional on μ and Σ as

$$X_t | \, (\mu, \Sigma) \sim N(\mu, \Sigma).$$

The prior distribution for μ is usually assumed to be normally distributed, and Σ^{-1} is assumed to have a Wishart distribution, say $\Sigma^{-1} \sim W_N((\nu\Omega)^{-1}, \nu)$. where Ω and ν are hyperparameters, and are assumed known. From the Wishart distribution, $E(\Sigma^{-1}) = \Omega^{-1}$, and ν reflects the strength about the belief. A larger ν corresponds to a stronger belief about the prior mean Ω^{-1}. Under these prior distributions, the posterior distribution is given by the inverse Wishart distribution

$$\Sigma \, | \, S \sim W_N^{-1}\Big([(T-1)S + \nu\Omega]^{-1}, N + T + \nu \Big). \tag{5}$$

The mode of the posterior distribution, which is considered as a Bayesian estimator for Σ, is given by

$$\hat{\Sigma} = \frac{T-1}{T-1+\nu} S + \frac{\nu}{T-1+\nu} \Omega. \tag{6}$$

This could be viewed as the maximum likelihood estimator based on posterior distributions. Clearly, if $\nu \to \infty$, meaning the prior belief is very strong, then the Bayesian estimator in the limit collapses to Ω. We can also interpret the Bayesian estimator as a shrinkage estimator toward Ω.

In the preceding analysis, Ω and ν are called hyperparameters and are assumed to be known. In a hierarchical Bayesian analysis, the hyperparameters Ω and ν are themselves random variables. In this view, the posterior distribution in (5) is really a conditional posterior and should be written as

$$\Sigma | \, (S, \Omega, \nu).$$

Prior distributions on Ω and ν are also needed. Because Ω is an $N \times N$ matrix, it contains too many free parameters, Chen (1979) suggested to impose some structures on Ω, such as a factor structure to reduce the number of free parameters. With a given structure, we may write $\Omega = \Omega(\theta)$, with θ being a vector with a much smaller dimension. A full Bayesian analysis would require prior distributions on θ and ν, then integrate out with respect to θ and ν to obtain a genuine posterior distribution for $\Sigma | S$. This is doable via Markov Chain Monte Carlo (MCMC) once the prior distributions are specified. Alternatively, one can treat θ and ν as unknown fixed constants, and estimate them from the marginal distributions of data using the maximum likelihood method. Estimating hyperparameters based on marginal distribution of the data is the essence of the empirical Bayes procedure. In other words, the empirical Bayes method estimates the prior distribution from the same data set X. A good introduction on the topic is Gelman et al (1997). The actual computation of marginal distribution of the data (conditional on θ, and ν) can be difficult. Chen used the EM algorithm to facilitate the computation. If a factor structure on Ω is imposed such that

$$\Omega = \beta\beta' + \Delta$$

where Δ is a diagonal matrix, the corresponding parameter is $\theta = (\beta, \Delta)$. The number of parameters in θ is of order N instead of $N(N+1)/2$ (the number of elements in Ω without a structure), a considerable reduction in the number of parameters. Once θ is estimated, denoted by θ^*, the final Bayesian estimator for Σ is defined as

$$\hat{\Sigma} = \frac{T-1}{T-1+\nu^*} S + \frac{\nu^*}{T-1+\nu^*} \Omega(\theta^*) \tag{7}$$

Chen (1979) referred the above estimator as that of shrinkage to a structure. Also see Daneils and Kass (1999, 2001), Yang and Berger (1994), and Barnard et al. (2000) for related studies.

The shrinkage estimator of Ledoit and Wolf (2003) estimates Ω directly from a factor model. Computationally, the Ledoit and Wolf estimator is much easier than the EM algorithm or MCMC method. But the optimal shrinkage α^* in Ledoit and Wolf assumes Ω cannot be the same as Σ. That is, the target is a biased estimator for Σ. But the Bayesian method does not require this.

7. RANDOM MATRIX THEORY APPROACH

The sample covariance matrix S as an estimator for Σ contains a considerable amount of noise when T is much smaller than N. Random matrix theory provides a way to de-noise the matrix S. The de-noised sample covariance is used as an estimator for Σ.

We first presents some pertinent properties for the eigenvalues of the sample correlation matrix under the assumption of independent and identically distributed (iid) random variables. Let X be a $N \times T$ random matrix with elements being iid (for example, iid normal random variables), and let S be the corresponding sample covariance. Again, let $D = \text{diag}(S)$ and $C = D^{-1/2} S D^{-1/2}$, so that C is the sample correlation matrix. Suppose $T/N \to Q$. Under the random matrix assumption, the eigenvalues of C has the following density function as $N, T \to \infty$:

$$p(\lambda) = \frac{Q}{2\pi\lambda}\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{min})}, \quad \lambda_{min} < \lambda < \lambda_{max}$$

where $\lambda_{max} = (1 + \sqrt{Q^{-1}})^2$ and $\lambda_{min} = (1 - \sqrt{Q^{-1}})^2$ when $Q > 1$. For $Q < 1$, the density has a point mass of $1 - Q$ at zero. All eigenvalues are bounded by λ_{max} in the limit; see Laloux et al. (1999), Plerou et al. (1999), Z. Bai (1999), and the reference therein. For example, when $Q = 1$, $\lambda_{max} = 4$, and in this case, under random matrix assumption, most of the eigenvalues are expected to be below 4, and the largest one should not far exceed 4, under finite samples.

Existence of eigenvalues exceeding λ_{max} indicates the presence of signal rather than pure noise. For stock returns, due to strong cross-section correlation, the largest eigenvalue far exceeds the upper bound λ_{max}. In fact, If the returns are generated by a factor model, it can be shown that the maximum eigenvalue of the correlation matrix converges to infinity, as T and N going to infinity. Kapetanios (2004) and Onatski (2005) use the properties of the largest eigenvalue of a random matrix to determine the number of factors. The method below does not require a factor model.

Again consider the spectral decomposition (via the singular value decomposition, say),

$$C = \sum_{i=1}^{N} \tau_i \xi_i \xi_i'$$

where $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_N$ are the eigenvalues of C and $\{\xi_k\}$ are the corresponding eigenvectors. Suppose there are k eigenvalues larger than λ_{max}, Laloux et al. (2001) defined a cleaned correlation matrix as

$$\overline{C} = \sum_{i=1}^{k} \lambda_i \xi_i \xi_i' + a\, I_N$$

where I_N is $N \times N$ identity matrix, and a is a constant such that the trace of $\overline{C}$ is equal to that of C. This implies that

$$a = \frac{\lambda_{k+1} + \cdots \lambda_N}{N}$$

From $\xi_i'\xi_i = 1$ for all i, it is clear that $tr(\overline{C}) = \sum_{i=1}^{N} \lambda_i = tr(C) = N$. The last equality follows from the fact that the diagonal elements of C being 1.

Once given the cleaned correlation matrix, the cleaned sample covariance is constructed as

$$\overline{S} = D^{1/2}\, \overline{C}\, D^{1/2}.$$

It should be pointed that while $\overline{C}$ is positive definite, it is not a correlation matrix because the diagonal elements are not necessarily being 1. Nevertheless, $\overline{S}$ is a covariance matrix, owing to its positive definiteness. One way to make $\overline{C}$ a correlation matrix is letting the diagonal elements be 1. This implies that $\overline{C}$ can be defined as

$$\overline{C} = \sum_{i=1}^{k} \lambda_i \xi_i \xi_i' + \mathrm{diag}(a_1, ..., a_N)$$

where a_j is equal to 1 minus the jth diagonal element of $\sum_{i=1}^{k} \lambda_i \xi_i \xi_i'$. That is, $a_j = 1 - \sum_{i=1}^{k} \lambda_i \xi_{ij}^2$. However, this definition leads to a covariance $\overline{S}$ identical to the principal component estimator $\hat{\Sigma}_c$ defined earlier, provided that k is equal to the number of factors r. Under either definition, since $\sum_{i=k+1}^{N} \lambda_i \xi_i \xi_i'$ is replaced by a diagonal matrix, it is clear that the de-noised $\overline{C}$ is equivalent to shrinking the off-diagonal elements of C towards zero.

A different cleaning method is suggested by Shafiri et al. (2003). Pafka and Kondor (2003) noted that the randomness in the sample covariance S

is not as large as one might think. In particular, the large noise reported by Plerou et al (1999) is primarily due to a too small T relative to N. Also, Jagannathan and Ma (2002) show that constrained portfolio optimization (imposing nonnegative weights) based on the sample covariance performs reasonably well.

8. INVERTING HIGH DIMENSIONAL COVARIANCE MATRICES

Many algorithms and techniques are available for finding the inverse of a given matrix. Our aim here is not to discuss which algorithm or technique to use, but rather to point out some mathematical facts that simplify the inversion, regardless of which numerical method is used.

While all the covariance matrices introduced earlier are $N \times N$, no inversion of a matrix exceeding the order $m \times m$ is needed, where $m = min(N, T)$. For the covariance matrices based on factor models, inverting a fixed dimensional (the number of factors) matrix is sufficient. Consider the matrix in (2),

$$\Sigma = \beta \Omega_Z \beta' + \Omega_\epsilon$$

where Ω_Z is $k \times k$ and Ω_ϵ is $N \times N$ but diagonal, its inverse is

$$\Sigma^{-1} = \Omega_\epsilon^{-1} - \Omega_\epsilon^{-1}\beta(\Omega_Z + \beta'\Omega_\epsilon^{-1}\beta)^{-1}\beta'\Omega_\epsilon^{-1}$$

The matrix Ω_ϵ^{-1} is easy to compute since it is diagonal. In the above formula, we only need the inverse of the $k \times k$ matrix $(\Omega_Z + \beta'\Omega_\epsilon^{-1}\beta)$. For the matrix in (3), the inversion is same, taking Ω_Z as a $k \times k$ identity matrix, and taking β to be $\hat{\Lambda}$, then applying the above inversion formula.

Inverting the Bayesian estimator in (6) or (7) is not difficult. If $N < T$, we directly submit Σ^{-1} for inversion. If $N >> T$, the following method is much easier. Consider (6) for notational simplicity. Define $a = (T - 1)/(T - 1 + \nu)$ and $b = \nu/(T - 1 + \nu)$. Note that $\Omega = \beta\beta' + \Delta$, we can write (6) as

$$\hat{\Sigma} = aS + b\beta\beta' + b\Delta = \bar{\beta}\bar{\beta}' + b\Delta$$

where $\bar{\beta} = [\sqrt{a/(T-1)}(X - \bar{X}), \sqrt{b}\beta]$. This follows from $S = (X - \bar{X})(X - \bar{X})'/(T - 1)$. Note that $\bar{\beta}$ is $N \times (T + k)$ assuming β is $N \times k$. Therefore,

$$\hat{\Sigma}^{-1} = b^{-1}\Delta^{-1} - b^{-2}\Delta^{-1}\bar{\beta}[I_{T+k} + \bar{\beta}'(b\Delta)^{-1}\bar{\beta}]^{-1}\bar{\beta}'\Delta^{-1}.$$

The above formula requires the inversion of a squared matrix of $T + k$, which can be much smaller than N.

Finally, inverting the de-noised matrix $\bar{S}$ based on the random matrix theory is straightforward since $\bar{S} = D^{1/2}\bar{C}D^{1/2}$ with D being a diagonal matrix and with $\bar{C}$ having a factor structure. It follows that $\bar{S}^{-1} = D^{-1/2}\bar{C}^{-1}D^{-1/2}$ and $\bar{C}^{-1}$ is easily computed in view of its factor structure.

9. CONCLUDING REMARKS

In this paper, we review some of the recent development in the estimation of covariance matrices when the number of variables is large compared to the number of observations. For example, in finance, the number of assets can be larger than the number of observations. Several methods are discussed, including the shrinkage method, methods based on the factor models, the Bayesian approach, and the random-matrix theory approach. For each method, the construction of covariance matrices is presented. The inversion of high dimensional covariance matrices is also discussed. In addition to applications in finance, high dimension covariance matrices may be useful in GMM (generalized method of moments estimation). When the number of moment conditions is large and the optimal weighting matrix is of high dimension, the sample analog of the optimal weighting matrix has too many free parameters. The methods presented here will be useful in reducing the number of parameters, making the GMM estimation more robust. Meng et al. (2011) consider GMM estimation when the optimal weighting matrix has a factor structure. In this context, however, if the number of moments is comparable to the number of observations, reducing the number of moments is more important than correctly estimating the weighting matrix. This issue is discussed in Bai and Ng (2010).

Postscript. The first version of this article was written in 2004, when Professor Shi and I were asked to contribute an article to a book on recent developments in economics research. Professor Shi also wrote a more elegant Chinese version of this article. The book ultimately did not materialize. Sadly, Professor Shi, my co-author and advisor, died unexpectedly in 2008. The publication of this article is one form of remembrance of Professor Shi. The substance of the current version remains the same as the original version except for some minor changes. In addition, I have updated the references.

Jushan Bai

REFERENCES

Anderson, T. W., 1984. *An Introduction to Multivariate Statistical Analysis*, John Wily & Sons.

Bai, J., 2003. Inferential Theory for Factor Models of Large Dimensions. *Econometrica* **71:1**, 135–172.

Bai, J., 2004. Estimating Cross-Section Common Stochastic Trends in Non-Stationary Panel Data. *Journal of Econometrics* **122**, 137–183.

Bai, J., 2010. Principal components and factor analysis, evaluating commonly used estimation procedures, unpublished manuscript, Department of Economics, Columbia University.

Bai, J. and K. P. Li, 2010. Statistical analysis of factor models of high dimension. Unpublished manuscript, Department of Economics, Columbia University.

Bai, J. and S. Ng, 2002. Determining the Number of Factors in Approximate Factor Models. *Econometrica* **70:1**, 191–221.

Bai, J. and S. Ng, 2008. *Large Dimensional Factor Models.* NOW Publishing Inc.

Bai, J. and S. Ng, 2010. Instrumental variable estimation in a data rich environment. *Econometric Theory* **26**, 1577-1606.

Bai, J. and S. Ng, 2011. Principal components estimation and identification of factors. Unpublished manuscript, Department of Economics, Columbia University.

Bai, Z., 1999. Methodologies in spectral analysis of large dimensional random matrices, a review. *Statistica Sinica* **9**, 611-677.

Barnard, J., R. McCulloch, and X. L. Meng, 2000. Modeling Covariance Matrices In Terms Of Standard Deviations And Correlations, With Application To Shrinkage. *Statistica Sinica* **10**, 1281-1311.

Bengtsson, C. and J. Holst, 2003. On Portfolio Selection: Improved Covariance Matrix Estimation for Swedish Asset Returns. Unpublished manuscript, Department of Economics, Luud University.

Bollerslev, T., 1987. A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return. *Review of Economics and Statistics* **69**, 542-54

Campbell, J. Y., A. Lo, and A. C. MacKinlay, 1997. *The Econometrics of Financial Markets*. Princeton University Press, New Jersey.

Chamberlain, G. and M. Rothschild,1983. Arbitrage, Factor Structure and Mean-Variance Analysis in Large Asset Markets. *Econometrica* **51**, 1305–1324.

Chen, C. F., 1979. Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis. *Journal of the Royal Statistical Society, Series B* **41**, 235-248.

Chen, N., R. Roll, and S. Ross, 1986. Economic Forces and the Stock Market. *Journal of Business* **59**, 383–403.

Chib, S., F. Nardari, and N. Shephard, 2002. Analysis of High Dimensional Multivariate Stochastic Volatility Models, Working paper, Washington University in Saint Louis.

Connor, G. and R. A. Korajczyk, 1986. Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis. *Journal of Financial Economics* **15**, 373C394.

Connor, G. and R. A. Korajczyk, 1988. Risk and Return in an Equilibrium APT: Application of a New Test Methodology. *Journal of Financial Economics* **21**, 255C289.

Daniels, M. J. and R. E. Kass, 1999. Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association* **94**, 1254-1263.

Daniels, M. J. and R. E. Kass, 2001. Shrinkage Estimators for Covariance Matrices. *Biometrics* **57**, 1173-1184.

Engle, R. F., 2002. Dynamic conditional correlation - a simple class of multivariate GARCH models. *Journal of Business and Economic Statistics* **20**, 339-350.

Engle, R. F.and K. F. Kroner, 1995. Multivariate simultaneous generalized ARCH. *Econometric Theory* **11**, 122-150.

Engle, R. F., V. K. Ng, and M. Rothschild, 1990. Asset pricing with a factor ARCH covariance structure: empirical estimates for treasury bills. *Journal of Econometrics* **45**, 213-238.

Engle, R. F.and K. Sheppard, 2001. Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH, unpublished paper: UCSD.

Fama, E. F. and K. R. French, 1993. Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics* **33**, 3C56.

Fan, J., Y. Fan, and J. Lv, 2008. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* **147**, 186-187.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin, 2000. The generalized dynamic-factor model: identification and estimation. *The Review of Economics and Statistics* **82**, pp 540-554.

Gao P. J. and X. D. Huang, 2008. Aggregate Consumption-Wealth Ratio and the Cross-Section of Stock Returns: Some International Evidence. *Annals of Economics and Finance* **9**, 1-37.

Gelman, A. J. B. Clarin, H. S. Stern, and D. B. Rubin, 1997. *Bayesian Data Analysis*, London: Chapman and Hall.

Han, Y., 2003. Asset Allocation with a High Dimensional Latent Factor Model. Olin School of Business Washington University in St. Louis.

Harvey, C. R., B. Solnik, and G. Zhou, 2002. What Determines Expected International Asset Returns? *Annals of Economics and Finance* **3**, 249-298.

Jagannathan, R. and T. Ma, 2002. Risk Reduction in Large Portfolios: Why imposing the wrong constraints helps, Working Paper.

Jones, C. S., 2001. Extracting Factors from Heteroskedastic Asset Returns. *Journal of Financial Economics* **62**, 293-325.

Kapetanios, G., 2004. A new method for determining the number of factors in factor models with large datasets. Department of Economics, Queen Mary University of London, working paper 525.

Laloux, L., P. Cizeau, J. P. Bouchaud, and M. Potters, 1999. Noise Dressing of Financial Correlation Matrices. *Phys. Rev. Lett* **83**, 1467.

Laloux, L., P. Cizeau, J. P. Bouchaud, and M. Potters, 2000. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance* **3**, 391-397.

Lawley D. N. and A. E. Maxwell, 1971, *Factor Analysis as a Statistical Method*, New York: American Elsevier Publishing Company.

Ledoit, O. and M. Wolf, 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance* **10(5)**, 603-621.

Ledoit, O. and M. Wolf, 2004. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88(2)**, 365-411.

Ledoit, O., P. Santa-Clara,and M. Wolf, 2003. Flexible multivariate GARCH modeling with an application to international stock markets. *The Review of Economics and Statistics* **85**, 735-747.

Lehmann, B. and D. Modest, 1988. The empirical foundations of the arbitrage pricing theory. *Journal of Financial Economics* **21**, 213-254.

Lettau, M. and S. Ludvigson, 2001, Resurrecting the (C)CAPM: A Cross-Sectional Test When Risk Premia are Time Varying. *Journal of Political Economy* **109:6**, 1238–1287.

Lintner, J, 1965. The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *Review of Economics and Statistics* **47:1**, 1337.

Markowitz, H. M., 1952,. Portfolio Selection. *Journal of Finance* **7(1)**, 77-91.

Meng, J. G., H. Gang, and J. Bai, 2011. OLIVE: A Simple Method For Estimating Betas When Factors Are Measured With Error. *Journal of Financial Research* **34(1)**, pages 27-60.

Nardari, F. and J. T. Scruggs, 2003. Analysis of Linear Factor Models with Multivariate Stochastic Volatility for Stock and Bond Returns, Department of Finance Arizona State University.

Onatski, A., 2005. Determining the Number of Factors from Empirical Distribution of Eigenvalues. Unpublished manuscript, Department of Economics, Columbia University.

Pafka, S. and I. Kondor, 2003. Noisy Covariance Matrices and Portfolio Optimization II. *Physica A*, **319**, 487-494.

Plerou, V., P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral, and H. E. Stanley, 1999. Universal and Non-universal Properties of Cross Correlations in Financial Time Series. *Phys. Rev. Lett.* **83**, 1471.

Plerou, V., P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, and H. E. Stanley, 2000. A Random Matrix Theory Approach to Financial Cross-Correlations. *Physica A* **267**, 374-382.

Ross, S. A., 1976. The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory* **13**, 341-360.

Sharifi, S., M. Crane, A. Shamaie, and H. Ruskin, 2004. Random Matrix Theory for Portfolio Optimization: A stability Approach. *Physica A* **335**, 629-643.

Sharpe, W., 1963. A Simplified Model for Portfolio Analysis. *Management Science* **79**, 277-231.

Sharpe, W., 1964. Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk. *Journal of Finance* **19(3)**, 425-442.

Stock, J. H. and M. W. Watson, 2002. Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association* **97**, 1167-1179.

Swartz, P. 2006. Global Versus Regional Systematic Risk and International Asset Allocations in Asia. *Annals of Economics and Finance* **7**, 77-89.

Tsay, R. S., 2002. *Analysis of Financial Time Series*, Wiley New York, 303-392.

Wang, P., 2008. Large Dimensional Factor Models with a Multi-Level Factor Structure: Identification, Estimation and Inference. Unpublished manuscript, New York University.

Yang, R. and J. O. Berger, 1994. Estimation of a covariance matrix using the reference prior. *Annals of Statistics* **22**, 1195-1121.