Covariance Regression Analysis

Tao Zou [a,b], Wei Lan [c,*], Hansheng Wang [a] and Chih-Ling Tsai [d]

[a] *Peking University,* [b] *The Australian National University,* [c] *Southwestern University of Finance and Economics and* [d] *University of California at Davis*

Abstract

This article introduces covariance regression analysis for a p-dimensional response vector. The proposed method explores the regression relationship between the p-dimensional covariance matrix and auxiliary information. We study three types of estimators: maximum likelihood, ordinary least squares, and feasible generalized least squares estimators. Then, we demonstrate that these regression estimators are consistent and asymptotically normal. Furthermore, we obtain the high dimensional and large sample properties of the corresponding covariance matrix estimators. Simulation experiments are presented to demonstrate the performance of both regression and covariance matrix estimates. An example is analyzed from the

[*] Corresponding author. Tel.: +86 13980920439. Email: lanwei@swufe.edu.cn.

Tao Zou is Ph.D. Candidate, Department of Business Statistics and Econometrics, Guanghua School of Management and Center for Statistical Science, Peking University (Email: zoutao@pku.edu.cn). He will join the Research School of Finance, Actuarial Studies and Applied Statistics at the Australian National University in 2016. Wei Lan is Assistant Professor, Statistics School and Center of Statistical Research, Southwestern University of Finance and Economics (Email: lanwei@swufe.edu.cn). Hansheng Wang is Professor and Chair, Department of Business Statistics and Econometrics, Guanghua School of Management, Peking University (Email: hansheng@gsm.pku.edu.cn). Chih-Ling Tsai is Distinguished Professor and Robert W. Glock Endowed Chair in Management, Graduate School of Management, University of California at Davis, Davis, CA 95616 (Email: cltsai@ucdavis.edu). Wei Lan's research was supported by National Natural Science Foundation of China (NSFC, 11401482, 71532001). Hansheng Wang's research was supported in part by National Natural Science Foundation of China (NSFC, 11131002, 11271032, 71532001), the Business Intelligence Research Center at Peking University, and the Center for Statistical Science at Peking University. The authors are grateful to the Editor, the AE, and the anonymous referee for their insightful comments and constructive suggestions.

Chinese stock market to illustrate the usefulness of the proposed covariance regression model.

KEY WORDS: Covariance Regression; Covariance Matrix Estimation; Positive Definiteness; Portfolio Management

1 Introduction

Let $\Sigma = \mathrm{Cov}(Y_i) \in \mathbb{R}^{p \times p}$ be a p-dimensional covariance matrix for a p-dimensional response vector $Y_i = (y_{i1}, \cdots, y_{ip})^\top$ for $i = 1, \cdots, n$ observations. The estimation of Σ has played a prominent role in many fields, including but not limited to: finance and risk management (Markowitz, 1952; Jagannathan and Ma, 2003), econometrics (Chen and Conley, 2001; Fan et al., 2008), biostatistics (Tong and Wang, 2007; Friedman et al., 2008), and machine learning (Bilmes, 2000). For independent and identically distributed random observations, $Y_1, \cdots, Y_n$, the classic sample covariance matrix estimator, $\hat{\Sigma}_{\mathrm{SAM}} = n^{-1} \sum (Y_i - \overline{Y})(Y_i - \overline{Y})^\top$ with $\overline{Y} = n^{-1} \sum Y_i$, is consistent when $n \to \infty$ and p is fixed. However, $\hat{\Sigma}_{\mathrm{SAM}}$ does not perform well as $p \to \infty$ (Bai, 1999), which can yield non-negligible estimation errors (Kan and Zhou, 2007). The fundamental reason for this result is that the number of unknown parameters are too large to be accurately estimated by a limited sample size. Hence, reducing the number of parameters to be estimable in covariance matrices becomes critical.

One popular approach for bringing down the number of parameters is assuming that the covariance matrix is sparse. In the last few years, various sparsity constraints have been imposed on either Σ (Huang et al., 2006; Bickel and Levina, 2008a,b; Cai and Liu, 2011; Leng and Li, 2011), its inverse Σ^{-1} (Dempster, 1972; Friedman et al., 2008; Cai et al., 2011), or its eigenvalues (Johnstone and Lu, 2009). An alternative approach is considering a factor model (Fan et al., 2008, 2011). Although the above approaches are useful, they all require that $n \to \infty$ to assure the consistency of covariance estimators. To overcome this challenge, one can employ the commonly used structured covariance matrix models that involve one or a small number of parameters, such as compound symmetry, autoregressive (e.g., AR(1)), and moving average (e.g., MA(1)). However, neither the spare covariance approach nor the structured covariance approach can directly link the covariance estimator to the auxiliary information (e.g., explanatory variables, spatial information, and social network). This motivates us to explore a new avenue to estimate the covariance matrices.

There are many motivating examples, and we provide three here. In the area of empirical finance with responses being returns of stocks, the covariance matrix of responses plays an important role for the risk management and portfolio allocation (see, for example, Markowitz, 1952; Jagannathan and Ma, 2003; Kan and Zhou, 2007). In addition, many researchers have shown that such a covariance matrix is affected by firms' fundamentals (Roll, 1988; Chan et al., 1998, 1999). This suggests that the covariance matrix can be explained by its associated relevant explanatory variables. We next observe that, in the field of spatial data analysis, the responses are often collected from different geographical locations. It is not surprising that the responses located near each other are likely to be strongly correlated. Accordingly, spatial statistics attempts to explain the covariance structure of responses by their geographical locations (Cressie, 1991; Bivand et al., 2008; Cressie and Wikle, 2011). Finally, in the context of social networks, responses can be determined through human behaviors. Researchers also found that activities of the connected network users are likely to be correlated. This suggests that the comovement of responses is affected by the users' social networks (Glaeser et al., 1996; Akerlof, 1997; Brock and Durlauf, 2001). Hence, it is natural to estimate the covariance of responses via the social network structure.

Before proposing our covariance estimation method, we review two types of linkages between the covariance and auxiliary information (or covariates). The first type does not directly link the covariance to the auxiliary information. By using the fact that the mean vector of responses is a function of covariates, however, the resulting estimate of covariance is a function of covariates (e.g., see Anderson, 1973; Szatrowski, 1980). The second type directly links Σ to the covariates under special model structures (e.g., see Prentice, 1988; Demidenko, 2004; Hoff and Niu, 2012). It is worth noting that Anderson (1973) also modeled Σ as a linear combination of symmetric matrices, and later Szatrowski (1980) and Zwiernik et al. (2014) further studied the properties of the covariance estimates under the linear structure.

Inspired by the three motivating examples and the above methods for modeling the covariance, we integrate the similarity concept (e.g., see Johnson and Wichern, 1992), the direct linkage approach, and Anderson's (1973) linear combination method together, and then propose a covariance regression model to directly quantify the relationship between the covariance Σ and a linear combination of matrices induced by corresponding auxiliary information. We next present three types of estimators, the maximum likelihood estimator, the ordinary least squares estimator and the generalized least squares estimator of regression coefficients, and demonstrate that those estimators are asymptotically normal. It is worth noting that the maximum likelihood estimator is computationally complex and the ordinary least squares estimator is inefficient. The generalized least squares estimator not only mitigates the computational burden, but also achieves the same asymptotic efficiency as the maximum likelihood estimator. Subsequently, the convergence rates of the covariance matrix estimation errors with respect to both the spectral norm and the Frobenius norm are obtained and they indicate that the resulting covariance matrix estimators are consistent as either p or n (or both) goes to infinity. It is worth noting that we focus on $p \to \infty$ in this paper since our proposed covariance matrix is structured with a finite number of unknown parameters. For the sake of completeness, we also discuss the scenario with p fixed and $n \to \infty$ in Section 7, and the asymptotic results are similar to that of the unstructured sample covariance estimator.

The article is organized as follows. Section 2 introduces covariance regression models, and Section 3 studies the parameter space for positive definiteness of the covariance matrix. Section 4 presents theoretical results of parameter estimators, and Section 5 provides an algorithm to ensure the resulting covariance matrix estimator is positive definite. Simulation studies and an empirical example are given in Section 6, while Section 7 concludes the article with discussions. All theoretical proofs are relegated to the Appendix.

2 Covariance Regression Model

Recall that $Y_i = (y_{ij}) \in \mathbb{R}^p$ is the p-dimensional response vector collected from the i-th replication for $i = 1, \cdots, n$. For each replication, let y_j be the j-th response and denote its associated auxiliary information vector by $X_j = (X_{j1}, \cdots, X_{jK})^\top \in \mathbb{R}^K$. For example, in spatial data, consider y_j as j's housing price associated with geographical coordinates of the location captured in X_j. Furthermore, in social network analysis, y_j can be a measure of actor j's activity level and the X_j are the actor's demographic information and social characteristics. Moreover, in finance, y_j can be the j-th firm's stock return and the X_j are its financial fundamentals, including the market value, book-to-market ratio, cash flow, leverage, etc.

It is worth noting that the auxiliary information X_j in the above examples can be continuous or discrete. To gauge the closeness or distance between any two auxiliary information X_{j_1} and X_{j_2} in the K-dimensional space, we adapt the concept of pairwise comparisons (Johnson and Wichern, 1992) and consider the measures of similarity $w(X_{j_1}, X_{j_2})$ and distance $d(X_{j_1}, X_{j_2})$. For continuous variables, the commonly used Euclidean distance and Mahalanobis distance can be used as a measure for closeness. Then the similarity can be defined as a decreasing function of distance, e.g., $w(X_{j_1}, X_{j_2}) = \exp\left\{-d(X_{j_1}, X_{j_2})^2\right\}$. As for discrete variables, one can employ similarity measures as presented in Johnson and Wichern (1992). For instance, the similarity in social network studies (Scott, 1988; Wasserman, 1994) can be defined as follows:

$$w\left(X_{j_1}, X_{j_2}\right) = \begin{cases} 1 & \text{if two individuals } j_1 \text{ and } j_2 \text{ known each other;} \\ 0 & \text{otherwise.} \end{cases}$$

Although both the similarity and the distance measures are applicable for our study, we only focus on similarity in the rest of the paper.

The above examples also indicate that the responses associated with higher similarity are more likely to be correlated. Thus, it is natural to model the covariance matrix Σ as a function of the similarity measure $w\left(X_{j_1}, X_{j_2}\right)$. For the sake of illustration, we consider $n = 1$, and let $Y = Y_1$.

Then, we propose the following covariance regression model,

$$YY^\top = \beta_0 I_p + \beta_1 W(X^{(1)}) + \cdots + \beta_K W(X^{(K)}) + \mathscr{E}, \tag{2.1}$$

where the response Y is standardized to have mean zero, $X^{(k)} := (X_{1k}, \cdots, X_{pk})^\top \in \mathbb{R}^p$ for $k = 1, \cdots, K$, $W\left(X^{(k)}\right) = \left(w\left(X_{j_1 k}, X_{j_2 k}\right)\right)_{p\times p} \in \mathbb{R}^{p\times p}$ is the similarity matrix, I_p is the $p \times p$ identity matrix, the vector of regression coefficients $\beta = (\beta_0, \beta_1, \cdots, \beta_K)^\top$ describes the similarity that affects the comovement of responses, and $\mathscr{E}$ is a random matrix that satisfies $E(\mathscr{E}) = 0$. Hereafter, we denote $W_k = W(X^{(k)})$ and assume that $W_0 = I_p$, W_k are known for $k = 1, \cdots, K$, and K is fixed. Based on (2.1), we define

$$\Sigma(\beta) = \beta_0 I_p + \sum_{k=1}^{K} \beta_k W_k, \tag{2.2}$$

and it links the covariance of responses to the auxiliary information. Accordingly, the true covariance matrix of Y is $\mathbb{E}(YY^\top) = \Sigma(\beta^{(0)})$, where $\beta^{(0)} = (\beta_0^{(0)}, \cdots, \beta_K^{(0)}) \in \mathbb{R}^{K+1}$ is the true regression vector of β in (2.2). For the sake of simplicity, we denote $\Sigma_0 = \Sigma(\beta^{(0)})$.

Model (2.2) considers the covariance matrix to be a linear combination of known similarity matrices $W_0, \cdots, W_K$. This type of linear combination structure has been utilized by Anderson (1973), Szatrowski (1980) and Zwiernik et al. (2014). However, they did not directly link the covariance matrix to the similarity matrices induced by auxiliary information. Furthermore, Anderson (1973) studied the consistent estimators of the linear combination coefficients as the sample size n goes to infinity. In contrast, using the fact that the covariance regression model (2.2) is only a function of $K + 1$ parameters, we are able to obtain consistent estimators of β and Σ when $p \to \infty$ for any sample size n.

It is of interest to note that model (2.2) comprises various structured covariance matrices as special cases, such as sphericity, compound symmetry, banded, autoregressive, and moving average.

For example, the compound symmetry covariance matrix is defined by

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \varrho & \cdots & \varrho \\ \varrho & 1 & \cdots & \varrho \\ \vdots & \vdots & \vdots & \vdots \\ \varrho & \varrho & \cdots & 1 \end{pmatrix}_{p \times p}, \tag{2.3}$$

and it has the form of $\beta_0 I_p + \sum_{k=1}^{K} \beta_k W_k$ with $K = 1$, $W_1 = \sum_{|j_1 - j_2| \geq 1} E_{j_1 j_2}$, $\beta_0 = \sigma^2$ and $\beta_1 = \varrho\sigma^2$, where $E_{j_1 j_2}$ is the matrix with the (j_1, j_2)-th entry being 1 and other entries being 0. Furthermore, the banded covariance matrix is $(\sigma_{j_1 j_2} I\{|j_1 - j_2| \leq m\})_{p \times p}$, which has attracted considerable attention in the existing literature (Cai and Jiang, 2011; Qiu and Chen, 2012). When the banded covariance matrix has equal diagonal elements σ^2, it can be expressed by $\beta_0 I_p + \sum_{k=1}^{K} \beta_k W_k$, where $K = m(2p - m - 1)/2$, $\beta_0 = \sigma^2$, $\beta_1 = \sigma_{21}$, $W_1 = E_{21} + E_{21}^{\top}$, $\beta_2 = \sigma_{32}$, $W_2 = E_{32} + E_{32}^{\top}$, $\cdots$, $\beta_{p-1} = \sigma_{p,p-1}$, $W_{p-1} = E_{p,p-1} + E_{p,p-1}^{\top}$, $\beta_p = \sigma_{31}$, $W_p = E_{31} + E_{31}^{\top}$, $\cdots$, $\beta_{m(2p-m-1)/2} = \sigma_{p,p-m}$, $W_{m(2p-m-1)/2} = E_{p,p-m} + E_{p,p-m}^{\top}$. In sum, the covariance regression model not only allows us to estimate covariance by incorporating the auxiliary information, but also unifies the presentation of commonly used structured covariance matrices.

3 Parameter Space and Positive Definiteness

To ensure the covariance matrix being positive, we construct the parameter space as follows,

$$\mathcal{B} := \{\beta : \Sigma(\beta) > 0\}, \tag{3.1}$$

where $G_1 > G_2$ if the difference between any two generic matrices, $G_1 - G_2$, is positive definite. One can easily see that $\mathcal{B}$ is a nonempty set, since it contains $\{\beta : \beta_0 > 0, \beta_k \equiv 0 \text{ for any } k = 1, \cdots, K\}$ as a nontrivial subspace. To study the properties of $\mathcal{B}$ and the parameter estimators of β, we introduce the following three technical conditions. In addition, let $\lambda_j(A)$ represent the j-th largest eigenvalue of any generic symmetric matrix A and $\|G\|_2 = \{\lambda_1(G^{\top}G)\}^{1/2}$ be the spectral norm for any generic matrix G. Moreover, we use the expression "$(\text{tr}(H_{kl})$ for $k, l = 0, ..., K)_{(K+1)\times(K+1)}$" to denote a matrix that has dimension $(K+1)\times(K+1)$ and whose kl-th element is $\text{tr}(H_{kl})$ for some generic set of matrices H_{kl} with $k, l = 0, ..., K$. For the sake of simplicity, we denote it by $(\text{tr}(H_{kl}))_{(K+1)\times(K+1)}$.

(C1) For all symmetric matrices in $\{W_k \in \mathbb{R}^{p\times p} : k = 0, \cdots, K, K < \infty\}$, there exits $w > 0$ such that $\sup_{p\geq 1} \|W_k\|_2 \leq w < \infty$.
(C2) There exists $c > 0$ such that $\lambda_p(\Sigma_0) > c$ for any $p \geq 1$.
(C3) The matrix $p^{-1}(\text{tr}\left(\Sigma_0^d W_k \Sigma_0^d W_l\right))_{(K+1)\times(K+1)}$ converges to a positive definite matrix $Q_d \in \mathbb{R}^{(K+1)\times(K+1)}$, respectively, for $d = -2, -1, 0, 1$, where $\Sigma_0^0 := I_p$.

Condition (C1) assumes that the spectral norm of W_k $(k = 1, \cdots, K)$ is bounded away from infinity and is equivalent to

$$-\infty < -w < \inf_{p\geq 1} \lambda_p (W_k) \leq \sup_{p\geq 1} \lambda_1 (W_k) < w < \infty,$$

which is a reasonable condition. For example, Condition (C1) is typically satisfied when W_k is the adjacency matrix in a sparse network. Based on this condition, we obtain the following proposition.

Proposition 1. *Under Condition (C1), $\mathcal{B}$ is an open set.*

The openness of the parameter space $\mathcal{B}$ ensures that the maximum likelihood, ordinary least squares, and feasible generalized least squares estimators of regression parameters obtained in Section 4 should not lie on the boundary. Furthermore, under Condition (C2), Σ_0 is positive definite, which belongs to $\mathcal{B}$. A similar condition can be found in Bickel and Levina (2008a,b) and Wang (2009). This condition is useful in obtaining the following result.

Proposition 2. *Under Conditions (C1) and (C2), in conjunction with Lemma 2 in Appendix A, there exist two finite positive constants $\sigma_{\min}$ and $\sigma_{\max}$ and an open ball $U_\delta^{(0)} := \left\{\beta : \left\|\beta - \beta^{(0)}\right\| < \delta\right\}$ with $\delta > 0$, such that, for any $p \geq 1$,*

$$0 < \sigma_{\min} < \inf_{\beta\in U_\delta^{(0)}} \lambda_p (\Sigma(\beta)) \leq \sup_{\beta\in U_\delta^{(0)}} \lambda_1 (\Sigma(\beta)) < \sigma_{\max} < \infty, \tag{3.2}$$

ACCEPTED MANUSCRIPT

where $\|\cdot\|$ is the Euclidean norm of any arbitrary vector.

The above proposition implies that $\Sigma(\beta)$ is invertible around $U_\delta^{(0)}$. Accordingly, $\Sigma^{-1}(\beta)$ also has uniform boundedness, i.e.,

$$0 < \frac{1}{\sigma_{\max}} < \inf_{\beta \in U_\delta^{(0)}} \lambda_p \left(\Sigma^{-1}(\beta)\right) \leq \sup_{\beta \in U_\delta^{(0)}} \lambda_1 \left(\Sigma^{-1}(\beta)\right) < \frac{1}{\sigma_{\min}} < \infty. \tag{3.3}$$

In addition, it leads to, for any $p \geq 1$,

$$0 < \sigma_{\min} < \lambda_p (\Sigma_0) \leq \lambda_1 (\Sigma_0) < \sigma_{\max} < \infty. \tag{3.4}$$

The boundedness property can also be found in Bickel and Levina (2008a,b) and Wang (2009). Moreover, Condition (C3) is similar to the standard assumption imposed on the covariance matrix of regression estimators in the classical linear regression model. The above three conditions and two propositions help us to demonstrate the asymptotic properties of the proposed estimators given below.

4 Parameter Estimation and Asymptotic Properties

In this section, we assume that Y follows a multivariate normal distribution with mean 0 and covariance Σ_0, and then propose five parameter estimators of β, namely, the maximum likelihood estimator (MLE), the unconstrained and constrained ordinary least squares (OLS) estimators, and the unconstrained and constrained feasible generalized least squares (FGLS) estimators. In fact, obtaining the OLS and FGLS estimators does not require the normality assumption, however, this assumption allows us to fairly compare their efficiencies with that of MLE.

We will demonstrate that these five estimators are all $p^{1/2}$-consistent and asymptotically normal as $p \to \infty$. Hence, all the estimators fall into the small open ball $U_\delta^{(0)}$ defined in Proposition 2 with probability tending to 1 as $p \to \infty$, which is a subset of the parameter space $\mathcal{B}$ defined in (3.1). It is worth noting that the MLE and the constrained OLS and FGLS estimators are also in the

ACCEPTED MANUSCRIPT

parameter space $\mathcal{B}$ even when p is fixed. In contrast, the unconstrained OLS and FGLS estimators do not necessarily have this nice property.

4.1 Maximum Likelihood Estimator

The maximum likelihood estimator $\hat{\beta}_{p,MLE}$ can be obtained by maximizing the following log-likelihood function

$$\ell_p(\beta) = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{p}\log\left\{\lambda_j\left(\Sigma(\beta)\right)\right\} - \frac{1}{2}Y^\top\Sigma^{-1}(\beta)Y. \tag{4.1}$$

Since the log-eigenvalues of matrix $\Sigma(\beta)$ are involved in the above objective function, we have that $\hat{\beta}_{p,MLE} \in \mathcal{B}$. Before studying the asymptotic property of $\hat{\beta}_{p,MLE}$, we next present the large dimension results for the score function and the observed Fisher information of $\ell_p(\beta^{(0)})$.

Proposition 3. *Under Conditions (C1) – (C3), we have that, as $p \to \infty$,*

(i) $p^{-1/2}\dfrac{\partial\ell_p(\beta^{(0)})}{\partial\beta} \xrightarrow{d} N\left(0, \dfrac{1}{2}Q_{-1}\right)$ *and (ii)* $-\,p^{-1}\dfrac{\partial^2\ell_p(\beta^{(0)})}{\partial\beta\partial\beta^\top} \xrightarrow{p} \dfrac{1}{2}Q_{-1} > 0,$

where the matrix Q_{-1} is shorthand for Q_d, defined in Condition (C3), with $d = -1$.

The above results lead to the following conclusion.

Theorem 1. *Under Conditions (C1) – (C3), we obtain that, as $p \to \infty$,*

$$p^{1/2}\left(\hat{\beta}_{p,MLE} - \beta^{(0)}\right) \xrightarrow{d} N\left(0, 2Q_{-1}^{-1}\right).$$

Although the MLE is asymptotically efficient under the normality assumption, finding the maximal value of (4.1) becomes computationally infeasible when p gets large. To this end, we employ the conjugate gradient algorithm (Fletcher, 1987), rather than the Newton-Raphson method, to find the MLE that involves only the first-order derivative of the objective function (4.1) (see (A.4) in the Appendix). From (A.4), the computational complexity of the MLE is $O(p^4)$, which is still complicated. This motivates us to find alternative estimators that are easier to compute.

ACCEPTED MANUSCRIPT

4.2 OLS and FGLS Estimators

Motivated by the covariance regression model (2.1), the unconstrained ordinary least squares estimator can be obtained by minimizing

$$D_p(\beta) := \left\| YY^\top - \Sigma(\beta) \right\|_F^2 , \tag{4.2}$$

where $\|G\|_F = \{\mathrm{tr}\,(G^\top G)\}^{1/2}$ denotes the Frobenius norm for any generic matrix G. Specifically, solving $\partial D_p(\beta)/\partial\beta = 0$ yields the unconstrained estimator,

$$\hat{\beta}_{p,OLS} = (\mathrm{tr}\,(W_k W_l))^{-1}_{(K+1)\times(K+1)} \, (Y^\top W_k Y)_{(K+1)\times 1} , \tag{4.3}$$

whose asymptotic property is given in the following theorem.

Theorem 2. *Under Conditions (C1) – (C3), we have that, as $p \to \infty$,*

$$p^{1/2}\left(\hat{\beta}_{p,OLS} - \beta^{(0)}\right) \xrightarrow{d} N\left(0, 2Q_0^{-1}Q_1Q_0^{-1}\right).$$

Theorem 2 indicates that the unconstrained OLS estimator is $p^{1/2}$-consistent and asymptotically normal. However, it is less efficient than the MLE since its asymptotic variance, $2Q_0^{-1}Q_1Q_0^{-1}$, is greater than that of the MLE, $2Q_{-1}^{-1}$. To save the space, the proof is not presented here. It is worth noting that the calculation of the OLS estimator is simpler and faster than that of the MLE. This is because the calculation of the OLS estimator is not iterative. In addition, it only needs calculating $\mathrm{tr}(W_k W_l)$ and $Y^\top W_k Y$ in (4.3), whose computational complexity are of order $O(p^2)$ rather than $O(p^4)$ as for calculating the MLE. Based on $\hat{\beta}_{p,OLS}$, we obtain the associated covariance matrix, $\Sigma(\hat{\beta}_{p,OLS})$. Although it may not be positive definite, this problem can be fixed by imposing the constraint $\beta \in \mathcal{B}$. Accordingly, this yields the constrained OLS (OLS+) estimator $\hat{\beta}_{p,OLS+} = \arg\min_{\beta\in\mathcal{B}} D_p(\beta)$.

By Proposition 1 and Theorem 2, we know that $\mathcal{B}$ is an open set and the unconstrained OLS estimator is $p^{1/2}$-consistent, respectively. Thus, the unconstrained OLS estimator not only locates in the parameter space $\mathcal{B}$, but also falls into the small open ball $U_\delta^{(0)}$ defined in Proposition 2, as long as p is sufficiently large. This suggests that, as $p \to \infty$, the unconstrained OLS estimator is identical to the constrained OLS estimator, with probability tending to one. Consequently, $\hat{\beta}_{p,OLS}$ and $\hat{\beta}_{p,OLS+}$ have the same asymptotic distribution.

Theorems 1 and 2 yield a $p^{1/2}$-estimator $\hat{\beta}$ of $\beta^{(0)}$. Based on this result, we next derive the stochastic convergence rate of the estimated covariance matrix $\Sigma(\hat{\beta})$.

Theorem 3. *Suppose there exists an estimator $\hat{\beta}$ such that $p^{1/2}(\hat{\beta} - \beta^{(0)}) \xrightarrow{d} \mathcal{Z}$, where $\mathcal{Z}$ is a $K+1$ random vector and has a multivariate normal distribution as in Theorems 1 and 2. Then under Conditions (C1) – (C3), as $p \to \infty$, we have that*

$$\left\|\Sigma(\hat{\beta}) - \Sigma_0\right\|_2 = O_p\left(p^{-1/2}\right) \text{ and } \left\|\Sigma^{-1}(\hat{\beta}) - \Sigma_0^{-1}\right\|_2 = O_p\left(p^{-1/2}\right);$$

$$\left\|\Sigma(\hat{\beta}) - \Sigma_0\right\|_F^2 \xrightarrow{d} \mathcal{Z}^\top Q_0 \mathcal{Z} \text{ and } \left\|\Sigma^{-1}(\hat{\beta}) - \Sigma_0^{-1}\right\|_F^2 \xrightarrow{d} \mathcal{Z}^\top Q_{-2} \mathcal{Z}.$$

As a result, the orders of $p^{-1/2}\|\Sigma(\hat{\beta}) - \Sigma_0\|_F$ and $p^{-1/2}\|\Sigma^{-1}(\hat{\beta}) - \Sigma_0^{-1}\|_F$ are exactly equal to $p^{-1/2}$. Employing the known inequality,

$$p^{-1/2}\left\|\Sigma(\hat{\beta}) - \Sigma_0\right\|_F \leq \left\|\Sigma(\hat{\beta}) - \Sigma_0\right\|_2, \tag{4.4}$$

together with Theorem 3, implies that the order of $\|\Sigma(\hat{\beta}) - \Sigma_0\|_2$ and $\|\Sigma^{-1}(\hat{\beta}) - \Sigma_0^{-1}\|_2$ are also exactly equal to $p^{-1/2}$. Hence, $\Sigma(\hat{\beta})$ and $\Sigma^{-1}(\hat{\beta})$ are consistent estimators of Σ_0 and Σ_0^{-1}, respectively, in both Frobenius norm and spectral norm.

By the above theorem, we are able to obtain the feasible generalized least squares (FGLS) estimator (Wooldridge, 2012), which can improve the efficiency of the ordinary least squares estimator. Specifically, we obtain the FGLS estimator by minimizing

$$\widetilde{D}_p(\beta) = \mathrm{vec}^\top\left(YY^\top - \Sigma(\beta)\right)\left(\hat{\Sigma}^{-1} \otimes \hat{\Sigma}^{-1}\right)\mathrm{vec}\left(YY^\top - \Sigma(\beta)\right), \tag{4.5}$$

where $\hat{\Sigma}^{-1} = \Sigma^{-1}(\hat{\beta})$ is a consistent estimator of Σ_0^{-1} in Theorem 3, $\otimes$ represents the Kronecker product of two matrices, and $\mathrm{vec}(G)$ denotes the vectorization for any generic matrix G. In practice, we can consider $\hat{\Sigma}^{-1} = \Sigma^{-1}(\hat{\beta}_{p,OLS+})$, which is a positive definite matrix. After algebraic simplification, we have the unconstrained FGLS estimator,

$$\hat{\beta}_{p,FGLS} = \left(\mathrm{tr}\left(\hat{\Sigma}^{-1}W_k\hat{\Sigma}^{-1}W_l\right)\right)^{-1}_{(K+1)\times(K+1)} \left(Y^\top\hat{\Sigma}^{-1}W_k\hat{\Sigma}^{-1}Y\right)_{(K+1)\times 1}. \tag{4.6}$$

Subsequently, we demonstrate that $\hat{\beta}_{p,FGLS}$ is asymptotically as efficient as $\hat{\beta}_{p,MLE}$.

Theorem 4. *Under Conditions (C1) – (C3), as $p \to \infty$,*

$$p^{1/2}\left(\hat{\beta}_{p,FGLS} - \beta^{(0)}\right) \xrightarrow{d} N\left(0, 2Q_{-1}^{-1}\right).$$

In addition to achieving asymptotic normality and efficiency, the unconstrained FGLS estimator is computationally feasible for larger p. This is because FGLS does not require the iterative process, even though its computational complexity via (4.6) is $O(p^4)$, which is the same as that of the MLE. Since $\hat{\beta}_{p,FGLS}$ may not lie in the open set $\mathcal{B}$, we consider the constrained FGLS (FGLS+) estimator $\hat{\beta}_{p,FGLS+} = \arg\min_{\beta\in\mathcal{B}} \widetilde{D}_p(\beta)$. Employing the similar arguments to those for studying the unconstrained and constrained OLS estimators, we can show that $\hat{\beta}_{p,FGLS+}$ and $\hat{\beta}_{p,FGLS}$ have the same asymptotic distribution as $p \to \infty$.

In sum, we have obtained the asymptotic distributions of the MLE, the constrained OLS estimator, and the constrained FGLS estimators. However, the asymptotic variances of those three estimators, $2Q_{-1}^{-1}$, $2Q_0^{-1}Q_1Q_0^{-1}$, and $2Q_{-1}^{-1}$, include the unknown matrices Σ and Σ^{-1}. To resolve this problem, we can replace Σ and Σ^{-1} by their consistent estimators $\Sigma(\hat{\beta})$ and $\Sigma^{-1}(\hat{\beta})$ in Theorem 3, respectively, where $\hat{\beta}$ can be $\hat{\beta}_{p,MLE}$ (or $\hat{\beta}_{p,OLS+}$ or $\hat{\beta}_{p,FGLS+}$). Applying the same techniques as those for showing (A.10) in the proof of Theorem 4, we can demonstrate that these replacements yield the consistent estimators of the asymptotic variances mentioned above. Consequently, one can make inferences such as finding the confidence interval and testing the hypotheses for the covariance regression coefficient β_k $(k = 0, \cdots, K)$.

5 Algorithm for Constrained Estimation

In Section 4 we studied the constrained OLS and FGLS estimators. This section proposes an algorithm to ensure these two estimators exist when they are restricted to be in the parameter space $\mathcal{B}$. Let $M(\beta)$ generically represent the objective function. Accordingly, $M(\beta) = D_p(\beta)$ in (4.2) and $M(\beta) = \widetilde{D}_p(\beta)$ in (4.5). To make sure the estimated covariance matrix is positive definite, we consider the constrained parameter space $\left\{\beta : \Sigma(\beta) \geq \epsilon I_p\right\}$ for some arbitrarily small $\epsilon > 0$, where $G_1 \geq G_2$ denotes that the difference between any two generic matrices, $G_1 - G_2$, is positive semidefinite. Following the suggestion of Xue et al. (2012), we set $\epsilon = 10^{-5}$. As a result, the constrained estimator can be defined by

$$\hat{\beta}_+ = \arg\min_{\Sigma(\beta)\geq\epsilon I_p} M(\beta).$$

We next employ the alternating direction method to directly solve the optimization problem (see Fortin and Glowinski, 1983, Boyd et al., 2011 and Xue et al., 2012). To this end, an augmented parameter $\Theta \in \mathbb{R}^{p\times p}$ is introduced, which results in an equivalent optimization problem,

$$\left(\hat{\beta}_+, \hat{\Theta}_+\right) = \arg\min_{\beta,\Theta} \left\{M(\beta) : \Sigma(\beta) = \Theta, \Theta \geq \epsilon I_p\right\}. \tag{5.1}$$

To obtain constrained estimators, we then minimize the augmented Lagrangian function with a given penalty parameter $\mu > 0$,

$$L(\beta, \Theta; \Lambda) = M(\beta) - \langle\Lambda, \Theta - \Sigma(\beta)\rangle + \frac{1}{2\mu} \|\Theta - \Sigma(\beta)\|_F^2 , \tag{5.2}$$

where $\Lambda \in \mathbb{R}^{p\times p}$ is the Lagrange multiplier, and the inner product for two symmetric matrices is

$$\langle\Lambda, \Theta - \Sigma(\beta)\rangle = \text{vec}^\top(\Lambda)\,\text{vec}(\Theta - \Sigma(\beta)) = \text{tr}\left(\Lambda\left\{\Theta - \Sigma(\beta)\right\}\right).$$

To find the minimum of (5.2), we propose a three-step iteration procedure as follows:

Θ Step: $\Theta^{i+1} = \arg\min_{\Theta\geq\epsilon I_p} L(\beta^i, \Theta; \Lambda^i)$;

β Step: $\beta^{i+1} = \arg\min_{\beta} L(\beta, \Theta^{i+1}; \Lambda^i)$;

Λ Step: $\Lambda^{i+1} = \Lambda^i - \frac{1}{\mu}\left\{\Theta^{i+1} - \Sigma(\beta^{i+1})\right\}$;

where $i = 0, 1, \cdots$, and the iteration stops until the sequence $\{\beta^i\}$ meets certain stopping criterion.

In fact, the first two steps can be simplified to get explicit solutions for each iteration. In the Θ step, we have that

$$\begin{aligned}\Theta^{i+1} &= \arg\min_{\Theta \geq \epsilon I_p} \left\{ -\left\langle \Lambda^i, \Theta - \Sigma(\beta^i) \right\rangle + \frac{1}{2\mu} \left\| \Theta - \Sigma(\beta^i) \right\|_F^2 \right\} \\ &= \arg\min_{\Theta \geq \epsilon I_p} \left\| \Theta - \left\{ \Sigma(\beta^i) + \mu\Lambda^i \right\} \right\|_F^2 = \left(\Sigma(\beta^i) + \mu\Lambda^i \right)_+ ,\end{aligned}$$

where $(\Sigma(\beta^i) + \mu\Lambda^i)_+ = \sum_{j=1}^p \max\{\lambda_j(\Sigma(\beta^i) + \mu\Lambda^i), \epsilon\} v_j v_j^\top$, and v_j is the eigenvector corresponding to the j-th largest eigenvalue $\lambda_j(\Sigma(\beta^i) + \mu\Lambda^i)$. In the β step, we obtain that

$$\beta^{i+1} = \arg\min_{\beta} \left\{ M(\beta) + \frac{1}{2\mu} \left\| \Theta^{i+1} - \left\{ \Sigma(\beta) + \mu\Lambda^i \right\} \right\|_F^2 \right\},$$

which can be solved precisely. Specifically, let $M(\beta) = D_p(\beta)$ in (4.2). Then, the resulting $(i+1)$-th iteration OLS+ estimator is

$$\beta^{i+1} = \frac{2\mu}{2\mu+1}\hat{\beta}_{p,OLS} + \frac{1}{2\mu+1} \left(\mathrm{tr}\,(W_k W_l)\right)^{-1}_{(K+1)\times(K+1)} \left(\mathrm{tr}\left(W_k \left(\Theta^{i+1} - \mu\Lambda^i \right) \right) \right)_{(K+1)\times 1}. \quad (5.3)$$

In contrast, let $M(\beta) = \widetilde{D}_p(\beta)$ in (4.5), which leads to the $(i+1)$-th iteration FGLS+ estimator

$$\beta^{i+1} = \left(\mathrm{tr}\left(2\mu\hat{\Sigma}^{-1} W_k \hat{\Sigma}^{-1} W_l + W_k W_l \right) \right)^{-1} \left(2\mu Y^\top \hat{\Sigma}^{-1} W_k \hat{\Sigma}^{-1} Y + \mathrm{tr}\left(W_k \left(\Theta^{i+1} - \mu\Lambda^i \right) \right) \right). \quad (5.4)$$

From the above discussion, we refine the three-step procedure and introduce the following algorithm.

Algorithm 1. *Obtaining the constrained estimators $\hat{\beta}_{p,OLS+}$ and $\hat{\beta}_{p,FGLS+}$.*

1. *Input: μ, ϵ, β^0 and Λ^0.*

2. *The three-step procedure involved in the* $(i+1)$*-th iteration,*

   Θ *Step:* $\Theta^{i+1} = \left(\Sigma(\beta^i) + \mu\Lambda^i\right)_+$;

   β *Step:* β^{i+1} *is given in (5.3) for the OLS+ estimator, while* β^{i+1} *is given in (5.4) for the FGLS+ estimator;*

   Λ *Step:* $\Lambda^{i+1} = \Lambda^i - \mu^{-1}\left\{\Theta^{i+1} - \Sigma(\beta^{i+1})\right\}$.

3. *Repeat Step 2 till the stopping criterion* $\|\beta^{i+1} - \beta^i\| < \xi$ *is met for some sufficiently small* $\xi > 0$.

Remark 1. *For the OLS+ estimator, it can be shown that if* $\Sigma(\hat{\beta}_{p,OLS}) \geq \epsilon I_p$*, then* $\hat{\beta}_{p,OLS+} = \hat{\beta}_{p,OLS}$*. Accordingly, we do not need to employ Algorithm 1 to get the constrained OLS estimator. This can mitigate the computational burden, especially when* p *is large. As for the FGLS+ estimator, we can make the following transformations*

$$\widetilde{Y} = \hat{\Sigma}^{-1/2}Y, \quad \widetilde{W}_k = \hat{\Sigma}^{-1/2}W_k\hat{\Sigma}^{-1/2}, \text{ and } \widetilde{\Sigma}(\beta) = \sum_{k=0}^{K}\widetilde{W}_k\beta_k.$$

It can be demonstrated that the unconstrained FGLS estimator, $\hat{\beta}_{p,FGLS}$*, obtained from (4.5) via* $\{Y, \Sigma(\beta)\}$ *is the same as the unconstrained OLS estimator,* $\tilde{\beta}_{p,OLS}$*, obtained from (4.2) via the transformed* $\{\widetilde{Y}, \widetilde{\Sigma}(\beta)\}$*. As discussed above, we do not need to employ Algorithm 1 to get the constrained FGLS estimator if* $\widetilde{\Sigma}(\tilde{\beta}_{p,OLS}) \geq \epsilon I_p$*, which can reduce computational complexity for large* p*.*

6 Numerical Studies

6.1 Simulation Studies

To evaluate the performance of the proposed method, we conduct Monte Carlo studies in various settings. Specifically, the response vector Y is simulated by $Y = \Sigma_0^{1/2}Z$, where the components of the vector Z are independently and identically generated from the standard normal distribution

$N(0,1)$. In addition, $\Sigma_0 = \beta_0^{(0)} I_p + \beta_1^{(0)} W_1 + \beta_2^{(0)} W_2$, where $\beta_0^{(0)} = 5, \beta_1^{(0)} = 0.5$, and $\beta_2^{(0)} = 0.2$, the diagonals of the similarity matrices $W_k = (w_{j_1 j_2}^{(k)})_{p \times p}$ are set to be zeros for $k = 1, 2$, and the off-diagonals of W_k are independently and identically generated from Bernoulli distributions with probability $5p^{-1}$ for $k = 1$ and probability $10p^{-1}$ for $k = 2$, respectively. The above model settings satisfy Conditions (C1) – (C3). We consider the sample size $n = 1$ and four different dimensions of responses, $p = 50$, 100, 200 and 500. For each of the above settings, a total of 1,000 realizations are conducted.

Although we are mainly interested in the estimations of MLE, OLS+, and FGLS+, we still present the results of OLS and FGLS. This allows us to examine the difference between the unconstrained and constrained estimates (i.e., OLS and FGLS vs OLS+ and FGLS+) numerically and check whether these two type estimates are almost identical as p becomes large as stated in Section 4.2. To obtain the OLS+ and FGLS+ estimates, we employ Algorithm 1 with the recommended values, $\mu = 0.05$, $\epsilon = 10^{-5}$ (suggested by Xue et al., 2012), $\xi = 10^{-8}$, $\beta^0 = \hat{\beta}_{p,OLS}$, and $\Lambda^0 = 0$. However, this does not exclude other possible value settings. We also follow the approach in Remark 1 to reduce the computations. For the FGLS and FGLS+ estimates, we consider $\hat{\Sigma}^{-1} = \Sigma^{-1}(\hat{\beta}_{p,OLS+})$ in the objective function (4.5).

Since our focus is on the covariance matrix, we begin by demonstrating the performance of the estimated covariance matrices and then show their associated regression parameter estimates. Table 1 reports the two types of averaged estimation errors, Spectral-Error and Frobenius-Error, of covariance matrices measured under the spectral norm and the Frobenius norm (i.e., $||\Sigma(\hat{\beta}) - \Sigma_0||_2$ and $p^{-1/2}||\Sigma(\hat{\beta}) - \Sigma_0||_F$), across the five proposed covariance matrix estimates $\Sigma(\hat{\beta}_{p,MLE})$, $\Sigma(\hat{\beta}_{p,OLS})$, $\Sigma(\hat{\beta}_{p,OLS+})$, $\Sigma(\hat{\beta}_{p,FGLS})$, and $\Sigma(\hat{\beta}_{p,FGLS+})$. In addition, the execution times with programming in C by using an Intel (R) Xeon (R) CPU (2.40 GHz) are also reported to assess the computational complexity. Overall, there are four important findings via simulation studies. The first is that the averaged Spectral-Errors and Frobenius-Errors evaluated at five different types of estimates via 1,000 realizations decrease as p gets large. This is consistent with the theoretical result of Theorem

3. The second is that the OLS and OLS+ (or the FGLS and FGLS+) estimates are almost 100% identical as p becomes moderate or large. The third is that, as $p = 500$, OLS+ (or OLS) takes less than 0.235 ($\times$ 1,000) seconds to finish the computation, while the MLE takes more than 87,417 ($\times$ 1,000) seconds to complete the computation. Since the MLE's finite dimension performance corroborates the theoretical finding and requires a large amount of computation time, we do not conduct the MLE computation for $p = 500$ in the remaining simulation studies. In addition, the computation of the FGLS+ (or FGLS) takes more than 1,785 ($\times$ 1,000) seconds to complete when $p = 500$. Although the execution time is much less than that of the MLE, it is significantly larger than that of OLS+ (or OLS). In sum, the execution times are consistent with the computational complexities of the proposed estimators discussed in Section 4.

The last finding is to examine the finite dimension performance of the five proposed covariance matrix estimates. Both Spectral-Error and Frobenius-Error show that MLE performs the best across $p = 50$, 100, and 200. This finding is not surprising since the MLE regression estimates are efficient (see Theorem 1). It is of interest to note that OLS and OLS+ are slightly better than FGLS and FGLS+ when $p = 50$ and 100. This is because the FGLS and FGLS+ estimates rely on $\Sigma^{-1}(\hat{\beta}_{p,OLS+})$, which is approximately consistent only when p gets large. Accordingly, as p increases to 200 and 500, FGLS and FGLS+ are superior to OLS and OLS+. It is also of interest to note that FGLS+ (or FGLS) is slightly superior to MLE when $p = 500$. This finding is not surprising since both types of estimators are consistent and efficient. Thus, their finite dimension performance can slightly vary. Based on the above findings, we conclude that MLE should be used only for small p, FGLS+ can be considered for moderate p, and OLS+ is favorable for large p even losing some efficiency. One can also use FGLS and OLS when p is moderate and large, respectively.

It is worth noting that the performance of the covariance matrix estimate depends on its regression parameter estimates. To assess the five proposed regression estimates, $\hat{\beta}_{p,MLE}$, $\hat{\beta}_{p,OLS}$, $\hat{\beta}_{p,OLS+}$, $\hat{\beta}_{p,FGLS}$, and $\hat{\beta}_{p,FGLS+}$, Table 2 presents their averaged biases (BIAS), standard deviations (SD), and root mean squared errors (RMSE) via 1,000 realizations. It shows three interesting findings. First, BIAS and SD generally become smaller when p gets larger. Hence, it is not surprising that the RMSEs reveal the same pattern. Second, the OLS and OLS+ yield smaller RMSEs than those of FGLS and FGLS+, respectively, when $p = 50$ and 100. This is due to the fact that FGLS and FGLS+ estimates rely on $\Sigma^{-1}(\hat{\beta}_{p,OLS+})$ and $\hat{\beta}_{p,OLS+}$ only being consistent in a large p. Third, MLE generally performs the best for $p = 50$ and 100, while FGLS and FGLS+ are usually superior to OLS and OLS+, respectively, as $p = 200$ and 500. Accordingly, the above findings are consistent to these recommendations obtained from Table 1 for estimating covariance matrices.

To assess the robustness of the covariance matrix estimates against the normality assumption of responses, we follow the same settings as the previous Monte Carlo study, except assuming that the components of Z are independently and identically generated from the standardized exponential distribution and the mixture normal distribution $0.9N(0, 5/9)+0.1N(0, 5)$, respectively. Under both distributions, the Spectral-Errors and Frobenius-errors calculated from the five covariance matrix estimates, $\Sigma(\hat{\beta}_{p,MLE})$, $\Sigma(\hat{\beta}_{p,OLS})$, $\Sigma(\hat{\beta}_{p,OLS+})$, $\Sigma(\hat{\beta}_{p,FGLS})$, and $\Sigma(\hat{\beta}_{p,FGLS+})$, show similar findings to those of Table 1, although their performances are slightly inferior to that of the normal distribution, as we expected. To save space, we only presented the case of the mixture normal distribution in Table 3. Consequently, our proposed estimate can be considered for non-normal responses. Finally, we conduct a Monte Carlo study with the same simulation settings as for the normal distribution by changing the similarity matrices to $\overline{W}_k = (\overline{w}^{(k)}_{j_1 j_2})_{p\times p}$, where the diagonals of the similarity matrices $\overline{W}_k$ are set to be zeros for $k = 1, 2$, the off-diagonals $\overline{w}^{(k)}_{j_1 j_2} = \exp\{-\overline{d}^{(k)}_{j_1 j_2}{}^2\}$, and $\overline{d}^{(k)}_{j_1 j_2}$ are independently and identically generated from uniform distributions $U(p^{-1/2}, p^{1/2})$ and $U(p^{-1/3}, p^{1/3})$, respectively, for $k = 1$ and $k = 2$. The results are in Table 4. In sum, our proposed covariance matrix estimates support theoretical findings and can be used for empirical applications.

6.2 Case Study

In this study, we employ our proposed covariance regression model to analyze the quarterly returns of $p = 660$ stocks in Shanghai Stock Exchange and Shenzhen Stock Exchange of China from 2007 to 2011, where the data were collected from the China Stock Market and Accounting Research (CSMAR) database. For each given quarter, the response variable Y is the corresponding returns (in percentages) of the 660 stocks, standardized by subtracting the sample mean. There are $T = 20$ quarters in total. In empirical finance, the covariance matrix of a large pool of stock returns measures the stock return comovement or synchronicity. As indicated by Roll (1988), stocks' comovement depends on the relative amounts of firm and market level information capitalized into stock prices, which is also directly related to the theory of market efficiency (Fama, 1970). Since the pioneering work of Roll (1988), considerable effort has been devoted to exploring the relationship between the stock return comovement (or synchronicity) and firms' fundamentals, which motivates us to employ our proposed method to estimate the covariance of stock returns via some relevant information of firms' fundamentals.

In practice, common experience suggests that the returns of the two stocks in the same industry are more highly correlated than those of two stocks in different industries, which was confirmed by Chan et al. (1999). In addition, Chan et al. (1998, 1999) found that the cash flow, stock size, and book-to-market ratio can help to explain the covariation in returns. Furthermore, Gul et al. (2010) employed leverage, size, and book-to-market ratio as control variables that are known to affect the stock return synchronicity. According to the above and extant literature, we consider the following $K = 5$ covariates to represent firms' fundamentals in this study: IND (industry classified by China Securities Regulatory Commission with 14 categories); LEV (leverage computed by liability-to-asset ratio); CF (cash flow of the firm); SIZE (measured by the logarithm of market value); and BM (book-to-market ratio). We label them as covariates $X^{(k)} = (X_{1k}, \cdots, X_{pk})^\top \in \mathbb{R}^p$ for $k = 1, \cdots, 5$, respectively. For the variable IND, let the off-diagonal element in the associated similarity matrix be 1 if two stocks belong to the same industry, and 0 otherwise, keeping this setting across all 20 quarters. For each given quarter, we next standardize the rest of the four variables via $p = 660$ observations so that they have zero mean and unit variance. Subsequently, we set the off-diagonal elements of the similarity matrices to be $\exp\left\{-(X_{j_1 k} - X_{j_2 k})^2\right\}$ for stocks $j_1 \neq j_2$ and covariates $k = 2, \cdots 5$, and let the diagonal elements be zeros.

The goal of this study is to assess the performance of portfolio by solving the Markowitz optimization problem (Markowitz, 1952). To this end, we adopt the commonly used rolling window procedure (see Chapter 9 of Zivot and Wang, 2007; Xue et al., 2012; Fan et al., 2014) with the window length $n = 1$ to construct and assess portfolio returns. Suppose that the t-th quarter data is $(Y_t, \boldsymbol{X}_t)$, where $Y_t \in \mathbb{R}^{p\times 1}$, $\boldsymbol{X}_t = (X_t^{(1)}, \cdots, X_t^{(K)}) \in \mathbb{R}^{p\times K}$ and $t = 1, \cdots, T$. Since the covariance matrix is time varying, we utilize each single period data at time t to fit the proposed covariance regression model and then estimate the covariance matrix $\Sigma_t = \mathrm{Cov}(Y_t)$. Hence, the estimation is based on the sample size $n = 1$ and $p = 660$. Following our simulation experience, we employ the FGLS+ approach to estimate the covariance matrix for each given t. Based on the estimated matrix $\Sigma(\hat{\beta}^{(t)}_{p,FGLS+})$, we obtain the optimal portfolio weights by minimizing the variance of the portfolio,

$$\omega_t^* = \arg\min_{\omega\in\mathbb{R}^p} \omega^\top \Sigma_t \omega, \tag{6.1}$$

such that $\omega^\top l = 1$ and $l = (1, \cdots, 1)^\top \in \mathbb{R}^p$, for $t = 1, \cdots, T$. In general, there are two types of optimal portfolio weights: one is obtained by imposing the no-shortsale constraint $\omega \geq 0$ in (6.1) (e.g., see Jagannathan and Ma, 2003; Xue et al., 2012), and the other one is calculated without imposing the no-shortsale constraint (e.g., see DeMiguel et al., 2009). Replacing Σ_t in (6.1) with its estimate $\Sigma(\hat{\beta}^{(t)}_{p,FGLS+})$ yields an empirical version of the portfolio weight $\hat{\omega}_t^*$. This allows us to compute the next period portfolio return $\hat{\omega}_t^{*\top} Y_{t+1}$, namely the out-of-sample portfolio return. Denote the resulting constrained portfolio return with $\omega \geq 0$ and unconstrained portfolio return without $\omega \geq 0$, at $t + 1$, by r^c_{t+1} and r^u_{t+1}, respectively. To make a comparison, we use the market portfolio return, r^m_{t+1}, as a benchmark, which is the average of all the stock returns with weights proportional to their market capitalization.

To examine the portfolio performance via r^c_{t+1} r^u_{t+1}, and r^m_{t+1} ($t = 1, \cdots, T$), we next consider the five commonly used measures (e.g., see Bodie et al., 2009): Mean (the average return of investment portfolios); SD (the risk of investment portfolios); Sharpe ratio (the Sharpe ratio is the excess return of the investment portfolio over the risk-free rate by adjusting SD); Alpha (the alpha coefficient is a risk-adjusted excess return of the investment portfolio over the benchmark); and Beta (the beta coefficient close to 1 indicates the out-of-sample portfolio has almost the same volatility as the benchmark). Table 5 presents five measures across the constrained portfolio, the unconstrained portfolio, and the market portfolio. It indicates that both constrained and unconstrained portfolios perform similarly, and they are able to earn around 4.8% ($\{1 + (3.57 - 2.38)\%\}^4 - 1$) and 8.1% ($\{1 + (4.34 - 2.38)\%\}^4 - 1$) annualized excess returns, respectively, over the market portfolio. In addition, their risks measured by SD and Beta are very close to that of the market portfolio. Hence, both portfolios can earn higher returns than the market portfolio at similar risk levels. As a result, it is not surprising that the Sharpe ratios of the two portfolios are around twice as large as that of the market portfolio (i.e., 0.13 vs 0.07 and 0.16 vs 0.07, respectively). Moreover, the Alpha coefficients of the two portfolios are greater than 0 so that they indicate some arbitrage opportunities for investing these two portfolios, although the Alpha coefficients are not significant at the 5% level.

The above study demonstrates the performance of the out-of-sample quarterly portfolio returns by changing the portfolio weight $\hat{\omega}^*_t$ at each quarter t. In practice, investors are also concerned about the fluctuation of the portfolio return within each quarter for risk control reasons. Hence, we further examine the robustness of the performance of the constrained and unconstrained portfolio returns over a shorter period of time, which also gives a more precise assessment of the portfolio risks (e.g., see Fan et al., 2014). Specifically, at quarter $t + 1$, we have collected $p = 660$ daily stock returns $Y^{(t+1)}_s \in \mathbb{R}^p$ ($s = 1, \cdots, T_{t+1}$) from T_{t+1} trading days. Then, the daily constrained (or unconstrained) portfolio returns at quarter $t + 1$ are $\hat{\omega}^*_t Y^{(t+1)}_s$. We then assess the daily returns calculated from the constrained portfolio, unconstrained portfolio, and market portfolio procedures (1,219 trading days from quarter 2 to quarter 21). Table 6 shows more profound results by comparing the constrained and unconstrained portfolio returns with the market portfolio return. In particular, the constrained and unconstrained portfolio procedures enable earning around 12.6% $(\{1+(0.072-0.025)\%\}^{252}-1)$ and 13.4% $(\{1+(0.075-0.025)\%\}^{252}-1)$ in annualized excess returns, respectively, over the market portfolio. Additionally, the Alpha coefficients are significantly positive.

To further explore the performance of out-of-sample portfolio returns calculated from our proposed covariance estimate, we compare them with the portfolio returns constructed by using the shrinkage estimate of the covariance matrix (Ledoit and Wolf, 2004) with $n = 1$. According to their approach, Σ_t in (6.1) can be estimated by $\bar{\Sigma}_t = \rho\mathrm{tr}(\bar{S}_t)I_p/p + (1-\rho)\bar{S}_t$, where $\bar{S}_t = Y_tY_t^\top$, $0 \leq \rho \leq 1$ is chosen to minimize the expected quadratic loss of the estimate $\bar{\Sigma}_t$, and $t = 1, \cdots, T$. However, the estimation of ρ calculated via Lemma 3.4 of Ledoit and Wolf (2004) is 0 for $n = 1$, which leads to $\bar{\Sigma}_t$ being singular. Therefore, we consider all possible values of ρ ranging from 0.1 to 0.9 in 0.1 increments. For the sake of comparison, we label the constrained and unconstrained portfolio returns constructed by $\Sigma(\hat{\beta}^{(t)}_{p,FGLS+})$ as Constrained I and Unconstrained I, respectively, and we label the constrained and unconstrained portfolio returns constructed by $\bar{\Sigma}_t$ as Constrained II and Unconstrained II, correspondingly. We also use the market portfolio return as a benchmark. Figure 1 depicts the quarterly and daily Sharpe ratios and Alpha values of the above five portfolio returns. It indicates that Constrained I and Unconstrained I are superior to Constrained II and Unconstrained II, respectively, and they all outperform the market return. In conclusion, our proposed covariance regression and covariance matrix estimates obtained from the firms' fundamentals can be effectively used for portfolio analysis.

ACCEPTED MANUSCRIPT

7 Concluding Remarks

In this paper, we utilize auxiliary information and employ a covariance regression approach to estimate the covariance matrix. Three estimation methods (MLE, OLS, and FGLS) have been proposed and their theoretical properties for both regression and covariance estimators are obtained. Simulation results demonstrate their performance, which supports theoretical findings. We also provide recommendations for using these three type estimators. An application for analyzing portfolio returns shows our proposed method performs well.

As addressed in the paper, the theoretical results of Theorems 1 through 4 are obtained by letting p go to infinity with $n = 1$. To improve the accuracy of estimating the covariance and its associated coefficients, we extend our results to $n \geq 1$. To this end, let $Y_1, \cdots, Y_n$ be the n independent and identically distributed samples as stated in Section 1. Adapting equations (4.1), (4.2), and (4.5), the corresponding objective functions for obtaining the MLE, OLS and FGLS estimators via the n observations, are as follows:

$$\ell_{np}(\beta) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\sum_{j=1}^{p}\log\left\{\lambda_j\left(\Sigma(\beta)\right)\right\} - \frac{1}{2}\sum_{i=1}^{n} Y_i^\top \Sigma^{-1}(\beta) Y_i, \tag{7.1}$$

$$D_{np}(\beta) = \sum_{i=1}^{n}\left\|Y_i Y_i^\top - \Sigma(\beta)\right\|_F^2, \text{ and} \tag{7.2}$$

$$\widetilde{D}_{np}(\beta) = \sum_{i=1}^{n}\mathrm{vec}^\top\left(Y_i Y_i^\top - \Sigma(\beta)\right)\left(\hat{\Sigma}^{-1}\otimes\hat{\Sigma}^{-1}\right)\mathrm{vec}\left(Y_i Y_i^\top - \Sigma(\beta)\right). \tag{7.3}$$

The resulting parameter estimators can be found in the supplementary material.

As suggested by an anonymous reviewer, we further discuss the asymptotic results of these three estimators under the following scenarios: (i) n is fixed and $p \to \infty$; (ii) $n \to \infty$ and $p \to \infty$; and (iii) $n \to \infty$ and p is fixed. Table 7 summarizes our findings. It is worth noting that, although the order of estimated regression coefficients is denoted by $O_p(n^{-1/2}p^{-1/2})$ as long as the data size $np \to \infty$, it is equivalent to the order of $O_p(p^{-1/2})$ when n is fixed and $O_p(n^{-1/2})$ when p is fixed. Accordingly, in scenarios (i) and (ii), under Conditions (C1) – (C3), the estimated regression coefficients are $\sqrt{p}$-consistent and $\sqrt{np}$-consistent, respectively. In scenario (iii), under Conditions (C2) and (C3′), the $\sqrt{n}$-consistency holds, where Condition (C3′) in Table 7 is slightly modified from Condition (C3). In this scenario, Condition (C1) is not required since it is automatically satisfied for fixed p. Furthermore, using the result of Schott (1997, p. 404), we can show that the asymptotic property of $\Sigma(\hat{\beta})$ is similar to that of the unstructured sample covariance estimator. The proofs of theoretical results presented in Table 7 are given in the supplementary material. Finally, the OLS and FGLS constrained estimators can be obtained by letting the objective function $M(\beta)$ in Section 5 be (7.2) and (7.3), respectively.

To expand the application of covariance regression analysis, we propose the following four major future research avenues. First, generalize the linear regression setting to the nonparametric or semiparametric setting, e.g., (i) change $W(X^{(k)})$ in (2.1) to $g_k(W(X^{(k)}))$, where g_k are smooth functions that can possibly be estimated via the kernel-based covariance approach (e.g., see Yin et al., 2010); (ii) motivated by an anonymous reviewer's comment, allow β_k in (2.1) to be a function of time (or a covariate) for $k = 0, \cdots, K$, which leads to a dynamic (or varying-coefficient) covariance regression model. Second, develop the model selection criterion for choosing the relevant variables when the dimension of auxiliary information K is large. Third, extend the multivariate normal distribution model to a multivariate generalized linear model (e.g., see the matrix-logarithmic covariance model proposed by Chiu et al., 1996 and the correlated binary regression model considered by Prentice, 1988). Fourth, the model used by Anderson (1973) and Szatrowski (1980) as well as the model introduced by Hoff and Niu (2012) have considered both mean regression function and covariance matrix. This, together with an anonymous reviewer's suggestion, motivates us to propose using the covariance regression approach for future application to longitudinal data analysis (Diggle et al., 2002). Specifically, this approach allows one to simultaneously model the conditional mean response with covariates and the covariance with a similarity matrix, under the scenarios of fixed p or $p \to \infty$. We believe all of these extensions would further demonstrate the usefulness of employing the regression approach to estimate the covariance matrix when the sample size is small.

Appendix A: Proofs of Lemmas and Propositions

Lemma 1. *For a real matrix* $\mathcal{W} \in \mathbb{R}^{p\times p}$*, let* $\sigma_1(\mathcal{W}) = \{\lambda_1(\mathcal{W}^\top \mathcal{W})\}^{1/2} = \|\mathcal{W}\|_2$ *be the largest singular values of* $\mathcal{W}$*. Then, for any two real matrices* $\mathcal{W}_1$ *and* $\mathcal{W}_2 \in \mathbb{R}^{p\times p}$*, we have (i)* $\sigma_1(\mathcal{W}_1 + \mathcal{W}_2) \le \sigma_1(\mathcal{W}_1) + \sigma_1(\mathcal{W}_2)$ *and (ii)* $\sigma_1(\mathcal{W}_1\mathcal{W}_2) \le \sigma_1(\mathcal{W}_1)\sigma_1(\mathcal{W}_2)$*.*

Proof: The results (i) and (ii) are exacted, respectively, from the result of 6.72 on page 118 and the result of 6.80 on page 120 of Seber (2008).

Lemma 2. *For any symmetric matrix* $\mathcal{W} \in \mathbb{R}^{p\times p}$*, we have that (i)* $\sigma_1(\mathcal{W}) = \max\{|\lambda_1(W)|, |\lambda_p(W)|\}$*. In addition, for symmetric matrices* $\mathcal{W}_k \in \mathbb{R}^{p\times p}, k = 1, \cdots, \bar{K}$*, if there exits* $w_{\max} > 0$ *such that*

$$-\infty < -w_{\max} \le \inf_{p\ge1} \lambda_p(\mathcal{W}_k) \le \sup_{p\ge1} \lambda_1(\mathcal{W}_k) \le w_{\max} < \infty, \text{ for all } k = 1, \cdots, \bar{K},$$

then (ii) the absolute eigenvalues of $\eta_1\mathcal{W}_1 + \cdots \eta_{\bar{K}}\mathcal{W}_{\bar{K}}$ *are all bounded away from infinity for any* $p \ge 1$ *and any constants* $\eta_1, \cdots, \eta_{\bar{K}} \in \mathbb{R}$*. Moreover, (iii)* $p^{-1}\mathrm{tr}\,(\eta_1\mathcal{W}_1 + \cdots \eta_{\bar{K}}\mathcal{W}_{\bar{K}}) = O(1)$*. Finally, for any two symmetric matrices* $\mathcal{W}_1$ *and* $\mathcal{W}_2$*, we have that (iv)* $p^{-1}\mathrm{tr}\,(\mathcal{W}_1\mathcal{W}_2) = O(1)$ *and (v) the absolute eigenvalues of* $\mathcal{W}_1\mathcal{W}_2\mathcal{W}_1$ *are bounded away from infinity for any* $p \ge 1$*.*

Proof: By the definition of the largest singular value in Lemma 1, we obtain (i). The results of (ii) and (iii) follow directly from Lemma 1 (i). Then, employing the inequality results of trace and eigenvalues in 6.77 of Seber (2008, p. 120), we obtain (iv). Lastly, Lemma 1 (ii) implies (v), which completes the proof.

Lemma 3. *Let*

$$S_p = \begin{pmatrix} vec^\top(\Upsilon_0) \\ vec^\top(\Upsilon_1) \\ \vdots \\ vec^\top(\Upsilon_K) \end{pmatrix} vec(ZZ^\top - I_p),$$

where $\Upsilon_k \in \mathbb{R}^{p\times p}$ are symmetric matrices whose absolute eigenvalues are all bounded away from infinity for $k = 0, \cdots, K, K < \infty$, and $p \geq 1$, and $Z \sim N\left(0, I_p\right)$. Then we have that $p^{-1/2-\epsilon} S_p \xrightarrow{L_2} 0$, for any $\epsilon > 0$. In addition, if $p^{-1}Cov(S_p) \to \mathcal{S} > 0$, then $p^{-1/2} S_p \xrightarrow{d} N(0, \mathcal{S})$.

Proof: By Theorem 9.19 of Schott (1997, p. 392), we have that

$$\mathrm{Cov}\left(\mathrm{vec}(ZZ^\top - I_p)\right) = I_{p^2} + K_{pp} =: 2N_p, \tag{A.1}$$

where $K_{pp} = \sum_{j_1,j_2=1}^{p}(E_{j_1j_2} \otimes E_{j_1j_2}^\top)$ is a $p^2 \times p^2$ matrix and $E_{j_1j_2}$ is defined below (2.3). Theorem 7.34 of Schott (1997, p. 282) shows that $N_p\mathrm{vec}(\Upsilon_k) = \mathrm{vec}\left(\Upsilon_k + \Upsilon_k^\top\right)/2$. Hence, for symmetric matrices Υ_k, $k = 0, \cdots, K$,

$$\mathrm{Cov}\left(\mathrm{vec}^\top(\Upsilon_k)\mathrm{vec}(ZZ^\top - I_p), \mathrm{vec}^\top(\Upsilon_l)\mathrm{vec}(ZZ^\top - I_p)\right) = 2\mathrm{vec}^\top(\Upsilon_k)N_p\mathrm{vec}(\Upsilon_l) = 2\mathrm{tr}\left(\Upsilon_k\Upsilon_l\right).$$

Then we obtain $\mathrm{Cov}(S_p) = 2(\mathrm{tr}\left(\Upsilon_k\Upsilon_l\right))_{(K+1)\times(K+1)}$. This, together with Lemma 2, implies that $\mathrm{Cov}\left(p^{-1/2-\epsilon}S_p\right) \to 0$, which immediately leads to $p^{-1/2-\epsilon}S_p \xrightarrow{L_2} 0$ since $\mathbb{E}(S_p) = 0$. The proof of the first part is complete.

We next show the second part of Lemma 3. By the Cramér-Wold device, it suffices to establish the asymptotic normality of $p^{-1/2}\theta^\top S_p$ for any arbitrary vector $\theta = (\theta_0, \cdots, \theta_K)^\top \in \mathbb{R}^{K+1}$. After algebraic simplification, we have

$$p^{-1/2}\theta^\top S_p = p^{-1/2}\left\{Z^\top\Upsilon(\theta)Z - \mathrm{tr}\left(\Upsilon(\theta)\right)\right\} = p^{-1/2}\sum_{j=1}^{p}\left\{\lambda_j(\theta)\overline{z}_j^2 - \lambda_j(\theta)\right\},$$

where $\lambda_1(\theta), \cdots, \lambda_p(\theta)$ are the eigenvalues of $\Upsilon(\theta) = \sum_{k=0}^{K}\theta_k\Upsilon_k$ that are all bounded via Lemma 2, and $\overline{z}_1, \cdots, \overline{z}_p \overset{iid}{\sim} N(0,1)$.

ACCEPTED MANUSCRIPT

Using the condition $p^{-1}\mathrm{Cov}(S_p) \to \mathcal{S}$, we further have that

$$\text{(a)} \quad \sum_{j=1}^{p} p^{-1}\mathbb{E}\left\{\lambda_j(\theta)\overline{z}_j^2 - \lambda_j(\theta)\right\}^2 = 2p^{-1}\sum_{j=1}^{p}\lambda_j^2(\theta) = 2p^{-1}\theta^\top \left(\mathrm{tr}\left(\Upsilon_k\Upsilon_l\right)\right)_{(K+1)\times(K+1)}\theta \to \theta^\top\mathcal{S}\theta$$

and, for any $\epsilon > 0$,

$$\begin{aligned} \text{(b)} \quad & \sum_{j=1}^{p} p^{-1}\mathbb{E}\left\{\lambda_j(\theta)\overline{z}_j^2 - \lambda_j(\theta)\right\}^2 I\left\{p^{-1/2}\left|\lambda_j(\theta)\overline{z}_j^2 - \lambda_j(\theta)\right| > \epsilon\right\} \\ \leq \quad & \frac{1}{p^2\epsilon^2}\sum_{j=1}^{p}\mathbb{E}\left\{\lambda_j(\theta)\overline{z}_j^2 - \lambda_j(\theta)\right\}^4 = \frac{60}{p^2\epsilon^2}\sum_{j=1}^{p}\lambda_j^4(\theta) = O(p^{-1}) \to 0, \end{aligned}$$

where the last equality is due to the boundedness of $\lambda_1(\theta), \cdots, \lambda_p(\theta)$. It is worth noting that (b) satisfies the Lindeberg-Feller condition. From (a) and (b), we then employ the central limit theorem and obtain $p^{-1/2}\theta^\top S_p \xrightarrow{d} N(0, \theta^\top\mathcal{S}\theta)$. The proof of the second part is complete.

Proof of Proposition 1: It is sufficient to show that, for any $\beta^* \in \mathcal{B}$ and $u \in \mathbb{R}^{K+1}$, there exists a positive constant δ_p such that, for any $\beta = (\beta^* + u) \in \{\beta : \|\beta - \beta^*\| < \delta_p\}$ (i.e., $\|u\| < \delta_p$), we obtain $\beta \in \mathcal{B}$. Based on (2.2), we have that $\Sigma\left(\beta^* + u\right) = \Sigma(\beta^*) + \Sigma(u) =: \Sigma^* + \Sigma(u)$. From (3.1), $\beta^* \in \mathcal{B}$ implies that there exists $\epsilon_p > 0$ such that $\lambda_p(\Sigma^*) \geq \epsilon_p$. Let $\|\cdot\|_1$ denote the L_1 norm. According to Condition (C1), we have that $\lambda_p\left(\Sigma(u)\right) \geq -w\|u\|_1$ and $w > 0$. In addition, choosing $\delta_p > 0$ small enough so that, for any $\|u\| < \delta_p$, we have $-w\|u\|_1 > -\epsilon_p$. The above results, together with the result of 6.70 in Seber (2008, p. 116), yield

$$\lambda_p\left(\Sigma\left(\beta^* + u\right)\right) \geq \lambda_p(\Sigma^*) + \lambda_p\left(\Sigma(u)\right) \geq \epsilon_p - w\|u\|_1 > 0. \tag{A.2}$$

Accordingly, there exists $\delta_p > 0$ such that for any $\beta^* + u$ satisfying $\|u\| < \delta_p$, $(\beta^* + u) \in \mathcal{B}$, which completes the proof.

Proof of Proposition 2: To show this proposition, it suffices to find positive constants $\delta, \sigma_{\min}$, and $\sigma_{\max}$ such that, for any β and $u \in \mathbb{R}^{K+1}$ satisfying $\beta = (\beta^{(0)} + u) \in U_\delta^{(0)}$ (i.e., $\|u\| < \delta$), the inequality (3.2) holds. Using the fact that $\Sigma\left(\beta^{(0)} + u\right) = \Sigma_0 + \Sigma(u)$ and the result of 6.70 in Seber (2008, p.

116), we have

$$\lambda_p(\Sigma_0) + \lambda_p\left(\Sigma(u)\right) \le \lambda_p\left(\Sigma\left(\beta^{(0)} + u\right)\right) \le \lambda_1\left(\Sigma\left(\beta^{(0)} + u\right)\right) \le \lambda_1(\Sigma_0) + \lambda_1\left(\Sigma(u)\right). \tag{A.3}$$

According to Condition (C2), there exist $\sigma_{\min} > 0$ and $\epsilon > 0$ such that $\lambda_p(\Sigma_0) \ge \sigma_{\min} + \epsilon$ for any p. Then, employing similar techniques to those used in the proof of Proposition 1, we obtain that

$$\lambda_p(\Sigma_0) + \lambda_p\left(\Sigma(u)\right) \ge \sigma_{\min} + \epsilon - w\,\|u\|_1 > \sigma_{\min}.$$

The last inequality given above is due to the fact that there exists $\delta > 0$, which satisfies $\|u\| < \delta$ and then yields $-w\,\|u\|_1 > -\epsilon$. Furthermore, Condition (C1) and Lemma 2 imply that there exists $\sigma_{\max} > 0$ such that $\lambda_1(\Sigma_0) < \sigma_{\max} - \epsilon$. Combing the above results, we have completed the proof of (3.2).

Proof of Proposition 3: From equation (4.1), we get the first two derivatives of the log likelihood function:

$$\frac{\partial \ell_p(\beta)}{\partial \beta_k} = \frac{1}{2}\text{vec}^\top\left(\Sigma^{-1}(\beta) W_k \Sigma^{-1}(\beta)\right)\text{vec}\left(YY^\top - \Sigma(\beta)\right) \text{ and} \tag{A.4}$$

$$\frac{\partial^2 \ell_p(\beta)}{\partial \beta_k \partial \beta_l} = -\frac{1}{2}\text{vec}^\top\left(\Sigma^{-1}(\beta) W_k \Sigma^{-1}(\beta) W_l \Sigma^{-1}(\beta) + \Sigma^{-1}(\beta) W_l \Sigma^{-1}(\beta) W_k \Sigma^{-1}(\beta)\right)\text{vec}\left(YY^\top - \Sigma(\beta)\right)$$

$$- \frac{1}{2}\text{tr}\left(\Sigma^{-1}(\beta) W_k \Sigma^{-1}(\beta) W_l\right),$$

where $k, l = 1, \cdots, p$. Define $Z := \Sigma_0^{-1/2} Y \sim N(0, I_p)$. Then, the first order derivative can be rewritten as

$$\frac{\partial \ell_p\left(\beta^{(0)}\right)}{\partial \beta} = \frac{1}{2}\begin{pmatrix} \text{vec}^\top\left(\Sigma_0^{-1}\right) \\ \text{vec}^\top\left(\Sigma_0^{-1/2} W_1 \Sigma_0^{-1/2}\right) \\ \vdots \\ \text{vec}^\top\left(\Sigma_0^{-1/2} W_K \Sigma_0^{-1/2}\right) \end{pmatrix}\text{vec}(ZZ^\top - I_p).$$

According to (A.1) and Condition (C3), we have

$$p^{-1}\text{Cov}\left(\frac{\partial \ell_p\left(\beta^{(0)}\right)}{\partial \beta}\right) = \frac{1}{2}p^{-1}\left(\text{tr}\left\{\Sigma_0^{-1}W_k\Sigma_0^{-1}W_l\right\}\right)_{(K+1)\times(K+1)} \to \frac{1}{2}Q_{-1} > 0.$$

This, in conjunction with Conditions (C1) and (C2), Lemmas 2 and 3, and (3.4), implies (i). The proof of the first part is complete.

We next verify (ii). According to Lemma 1, we obtain

$$\sigma_1\left(\Sigma_0^{-1}W_k\Sigma_0^{-1}W_l\Sigma_0^{-1} + \Sigma_0^{-1}W_l\Sigma_0^{-1}W_k\Sigma_0^{-1}\right) \le 2\left\{\sigma_1\left(\Sigma_0^{-1}\right)\right\}^3 \sigma_1(W_k)\,\sigma_1(W_l)\,.$$

Then, by Conditions (C1) and (C2), Lemma 2, and Proposition 2, we have that the eigenvalues of the symmetric matrix $\Sigma_0^{-1}W_k\Sigma_0^{-1}W_l\Sigma_0^{-1} + \Sigma_0^{-1}W_l\Sigma_0^{-1}W_k\Sigma_0^{-1}$ are bounded for any p. Furthermore, using the fact that $\text{vec}\,(YY^\top - \Sigma_0) = (\Sigma_0^{1/2}\otimes\Sigma_0^{1/2})\text{vec}\left(ZZ^\top - I_p\right)$ and employing Lemmas 2 and 3, we get

$$p^{-1}\text{vec}^\top\left(\Sigma_0^{-1}W_k\Sigma_0^{-1}W_l\Sigma_0^{-1} + \Sigma_0^{-1}W_l\Sigma_0^{-1}W_k\Sigma_0^{-1}\right)\text{vec}\,(YY^\top - \Sigma_0) = o_p(1).$$

This, together with Condition (C3), leads to

$$-p^{-1}\frac{\partial^2 \ell_p(\beta^{(0)})}{\partial\beta\partial\beta^\top} \xrightarrow{p} \frac{1}{2}Q_{-1} > 0,$$

which completes the second part of the proof.

Appendix B: Proofs of Theorems

Proof of Theorem 1:

Part I. We first prove the $p^{1/2}$-consistency of the estimator $\hat{\beta}_{p,MLE}$. Following the approach of Fan and Li (2001), it suffices to show that, for any given $\epsilon > 0$, there exist $M_\epsilon > 0$ and $u \in \mathbb{R}^{K+1}$ such that,

$$\Pr\left\{\sup_{||u||=M_\epsilon} \ell_p\left(\beta^{(0)} + p^{-1/2}u\right) < \ell_p\left(\beta^{(0)}\right)\right\} \ge 1 - \epsilon, \tag{A.5}$$

ACCEPTED MANUSCRIPT

when p is sufficiently large enough. Accordingly, $\beta^{(0)} + p^{-1/2}u$ can be in the small open ball $U_{\delta}^{(0)}$ defined in Proposition 2 (i.e., $\|p^{-1/2}u\| < \delta$) when $\|u\| = M_{\epsilon}$ and p is sufficiently large. Employing the Taylor series expansion, we then obtain that

$$\begin{aligned} & \sup_{\|u\|=M_{\epsilon}} \left\{\ell_p\left(\beta^{(0)} + p^{-1/2}u\right) - \ell_p\left(\beta^{(0)}\right)\right\} \\ = & \sup_{\|u\|=M_{\epsilon}} \left\{p^{-1/2}\frac{\partial \ell_p\left(\beta^{(0)}\right)}{\partial \beta^{\top}}u - \frac{1}{2}p^{-1}u^{\top}\left\{-\frac{\partial^2 \ell_p\left(\beta^{(0)}\right)}{\partial\beta\partial\beta^{\top}}\right\}u + R_p(u)\right\} \\ \leq & M_{\epsilon}O_p(1) - \frac{1}{2}\lambda_{K+1}\left(\frac{1}{2}Q_{-1}\right)M_{\epsilon}^2 + o_p(1), \end{aligned} \tag{A.6}$$

where $\lambda_{K+1}(\cdot)$ is the smallest eigenvalue of $Q_{-1}/2$ and

$$R_p(u) = \frac{1}{6}p^{-3/2}u^{\top}\left(I_{K+1} \otimes u^{\top}\right)\frac{\partial}{\partial\beta^{\top}}\mathrm{vec}\left(\frac{\partial^2 \ell_p\left(\beta^{(0)} + p^{-1/2}\bar{c}u\right)}{\partial\beta\partial\beta^{\top}}\right)u \text{ with } 0 \leq \bar{c} \leq 1.$$

The inequality (A.6) is obtained from Proposition 3 and the result of $R_p(u) = O_p(p^{-1/2})$, which can be shown by applying Proposition 2 as well as employing similar techniques to those used in the proofs of Lemma 3 and Proposition 3. Note that $M_{\epsilon}O_p(1) - \lambda_{K+1}(Q_{-1}/2)M_{\epsilon}^2/2$ is a quadratic function of M_{ϵ} and Condition (C3) implies its quadratic coefficient $-\lambda_{K+1}(Q_{-1}/2)/2 < 0$. Hence, as long as M_{ϵ} is sufficient large, we have

$$\sup_{\|u\|=M_{\epsilon}} \left\{\ell_p\left(\beta^{(0)} + p^{-1/2}u\right) - \ell_p\left(\beta^{(0)}\right)\right\} < 0, \tag{A.7}$$

with probability tending to 1, which demonstrates (A.5). As a result of (A.7), there exists a local maximizer $\hat{\beta}_{p,MLE}$ such that $\|\hat{\beta}_{p,MLE} - \beta^{(0)}\| \leq p^{-1/2}M_{\epsilon}$ when p is large enough. Since $\ell_p(\beta)$ is a concave function, the local maximizer is also the global maximizer. Based on the above results, we have

$$\Pr\left(\left\|\hat{\beta}_{p,MLE} - \beta^{(0)}\right\| \leq p^{-1/2}M_{\epsilon}\right) \geq \Pr\left\{\sup_{\|u\|=M_{\epsilon}} \ell_p\left(\beta^{(0)} + p^{-1/2}u\right) < \ell_p\left(\beta^{(0)}\right)\right\} \geq 1 - \epsilon,$$

which implies that $p^{1/2}\|\hat{\beta}_{p,MLE} - \beta^{(0)}\| = O_p(1)$. This completes the proof of the first part.

PART II. We next show that $\hat{\beta}_{p,MLE}$ is asymptotically normal. Using the result of Part I and applying the Taylor series expansion, we have that $0 = \partial \ell_p \left(\hat{\beta}_{p,MLE}\right)/\partial\beta = \partial \ell_p(\beta^{(0)})/\partial\beta + \partial^2 \ell_p(\beta^{(0)})/\partial\beta\partial\beta^\top(\hat{\beta}_{p,MLE} - \beta^{(0)}) + \bar{R}_p,$ where

$$\bar{R}_p = \frac{1}{2}\left\{I_{K+1} \otimes \left(\hat{\beta}_{p,MLE} - \beta^{(0)}\right)^\top\right\} \frac{\partial}{\partial\beta^\top}\mathrm{vec}\left(\frac{\partial^2 \ell_p\left(\bar{\beta}\right)}{\partial\beta\partial\beta^\top}\right)\left(\hat{\beta}_{p,MLE} - \beta^{(0)}\right)$$

and $\bar{\beta}$ lies between $\hat{\beta}_{p,MLE}$ and $\beta^{(0)}$. It is worth noting that $\bar{\beta}$ also lies in the open ball $U_\delta^{(0)}$ as $p \to \infty$. Applying Proposition 2 as well as employing similar techniques to those used in the proofs of Lemma 3 and Proposition 3, we obtain $p^{-1}\partial\, \mathrm{vec}(\partial^2 \ell_p\left(\bar{\beta}\right)/\partial\beta\partial\beta^\top)/\partial\beta^\top = O_p(1)$. Consequently,

$$p^{1/2}\left(\hat{\beta}_{p,MLE} - \beta^{(0)}\right) = p^{-1/2} 2 Q_{-1}^{-1} \frac{\partial \ell_p\left(\beta^{(0)}\right)}{\partial\beta} + o_p(1) \xrightarrow{d} N\left(0, 2Q_{-1}^{-1}\right),$$

which completes the entire proof.

Proof of Theorem 2: By (4.3), $\hat{\beta}_{p,OLS} - \beta^{(0)} = (\mathrm{tr}\,(W_k W_l))^{-1}_{(K+1)\times(K+1)}\, S_{p,OLS}(\beta^{(0)}),$ where

$$S_{p,OLS}\left(\beta^{(0)}\right) = \begin{pmatrix} \mathrm{vec}^\top(\Sigma_0) \\ \mathrm{vec}^\top\left(\Sigma_0^{1/2} W_1 \Sigma_0^{1/2}\right) \\ \vdots \\ \mathrm{vec}^\top\left(\Sigma_0^{1/2} W_K \Sigma_0^{1/2}\right) \end{pmatrix} \mathrm{vec}(ZZ^\top - I_p).$$

According to (A.1) in Lemma 3 and Condition (C3), we have that $p^{-1}\mathrm{Cov}\left(S_{p,OLS}(\beta^{(0)})\right) = 2p^{-1}$ $(\mathrm{tr}\,(\Sigma_0 W_k \Sigma_0 W_l))_{(K+1)\times(K+1)} \to 2Q_1 > 0$. Then, applying Conditions (C1) and (C2), Lemmas 2 and 3, and (3.4), we have $p^{-1/2} S_{p,OLS}(\beta^{(0)}) \xrightarrow{d} N(0, 2Q_1)$. In addition, Condition (C3) implies that $p^{-1}\,(\mathrm{tr}\,(W_k W_l))_{(K+1)\times(K+1)} \to Q_0 > 0$. Accordingly, by Slutsky's theorem, we complete the proof.

Proof of Theorem 3: It can be seen that $\Sigma(\hat{\beta}) - \Sigma_0 = \sum_{k=0}^K (\hat{\beta}_k - \beta_k^{(0)}) W_k$. Then, under Condition (C1) and applying the similar techniques to those used in the proof of the second inequality of (A.2), we obtain that

$$\|\Sigma(\hat{\beta}) - \Sigma_0\|_2 \le w \left\|\hat{\beta} - \beta^{(0)}\right\|_1 = O_p(p^{-1/2}). \tag{A.8}$$

We next study the asymptotic property of $\Sigma^{-1}(\hat{\beta})$. Note that $\hat{\beta}$ is a $p^{1/2}$-consistent estimator and hence it lies in the open ball $U_\delta^{(0)}$ (defined in Proposition 2) as $p \to \infty$. This allows us to apply the Taylor series expansion and obtain $\Sigma^{-1}(\hat{\beta}) = \Sigma_0^{-1} - \sum_{k=0}^K (\hat{\beta}_k - \beta_k^{(0)})\Sigma^{-1}(\bar{\bar{\beta}})W_k\Sigma^{-1}(\bar{\bar{\beta}})$, where $\bar{\bar{\beta}}$ lies between $\hat{\beta}$ and $\beta^{(0)}$. Applying Proposition 2 and then employing similar techniques to those used in the proof of Lemma 2, there exists a constant $c_{\max} > 0$ such that, for all $k = 0, \cdots, K$ and any $p \geq 1$,

$$\sup_{\beta \in U_\delta^{(0)}} \max_{j \in \{1,\cdots,p\}} \left| \lambda_j \left(\Sigma^{-1}(\beta) W_k \Sigma^{-1}(\beta) \right) \right| \leq c_{\max}.$$

Subsequently, using the technique from the proof of (A.8), we obtain that $\|\Sigma^{-1}(\hat{\beta}) - \Sigma_0^{-1}\|_2 = O_p(p^{-1/2})$, which completes the entire proof under the spectral norm.

Finally, under the Frobenius norm, we show the asymptotic properties of the covariance matrix estimator and its inverse. By the theorem's premise and Condition (C3), we have

$$\|\Sigma(\hat{\beta}) - \Sigma_0\|_F^2 = p^{1/2} \left(\hat{\beta} - \beta^{(0)} \right)^\top \left(p^{-1} \mathrm{tr}\left(W_k W_l \right) \right)_{(K+1)\times(K+1)} p^{1/2} \left(\hat{\beta} - \beta^{(0)} \right) \xrightarrow{d} \mathcal{Z}^\top \mathcal{Q}_0 \mathcal{Z}.$$

Using the fact that $\|\Sigma^{-1}(\hat{\beta}) - \Sigma_0^{-1}\|_F^2 = \mathrm{vec}^\top(\Sigma^{-1}(\hat{\beta}) - \Sigma_0^{-1})\mathrm{vec}(\Sigma^{-1}(\hat{\beta}) - \Sigma_0^{-1})$, we then employ the Taylor series expansion and obtain that

$$\left\|\Sigma^{-1}(\hat{\beta}) - \Sigma_0^{-1}\right\|_F^2 = p^{1/2} \left(\hat{\beta} - \beta^{(0)} \right)^\top \left(p^{-1} \mathrm{tr}\left(\Sigma_0^{-2} W_k \Sigma_0^{-2} W_l \right) \right)_{(K+1)\times(K+1)} p^{1/2} \left(\hat{\beta} - \beta^{(0)} \right) + \tilde{R}_p, \tag{A.9}$$

where

$$\tilde{R}_p = \frac{1}{6} \left(\hat{\beta} - \beta^{(0)} \right)^\top \left(I_{K+1} \otimes \left(\hat{\beta} - \beta^{(0)} \right)^\top \right) \frac{\partial}{\partial \beta^\top} \mathrm{vec} \left(\frac{\partial^2 \left\| \Sigma^{-1}(\tilde{\tilde{\beta}}) - \Sigma^{-1} \right\|_F^2}{\partial \beta \partial \beta^\top} \right) \left(\hat{\beta} - \beta^{(0)} \right)$$

and $\tilde{\tilde{\beta}}$ lies between $\hat{\beta}$ and $\beta^{(0)}$. Applying Proposition 2 and employing similar techniques to those used in the proofs of Lemma 3 and Proposition 3, we can show that $\tilde{R}_p = O_p(p^{-1/2})$. This, together with the theorem's premise, Condition (C3), equation (A.9), and the Slutsky's theorem, implies that $\|\Sigma^{-1}(\hat{\beta}) - \Sigma_0^{-1}\|_F^2 \xrightarrow{d} \mathcal{Z}^\top \mathcal{Q}_{-2} \mathcal{Z}$, which completes the entire proof.

Proof of Theorem 4: From (4.6), $\hat{\beta}_{p,FGLS} - \beta^{(0)} = (\mathrm{tr}(\hat{\Sigma}^{-1}W_k\hat{\Sigma}^{-1}W_l))^{-1}S_{p,FGLS}(\hat{\beta}),$ where

$$S_{p,FGLS}(\hat{\beta}) = \begin{pmatrix} \mathrm{vec}^\top\left(\Sigma_0^{1/2}\hat{\Sigma}^{-2}\Sigma_0^{1/2}\right) \\ \mathrm{vec}^\top\left(\Sigma_0^{1/2}\hat{\Sigma}^{-1}W_1\hat{\Sigma}^{-1}\Sigma_0^{1/2}\right) \\ \vdots \\ \mathrm{vec}^\top\left(\Sigma_0^{1/2}\hat{\Sigma}^{-1}W_K\hat{\Sigma}^{-1}\Sigma_0^{1/2}\right) \end{pmatrix} \mathrm{vec}(ZZ^\top - I_p),$$

and $k, l = 0 \cdots, K$. For simplicity, define $a_{kl}(\hat{\beta}) = \mathrm{tr}(\hat{\Sigma}^{-1}W_k\hat{\Sigma}^{-1}W_l) = \mathrm{vec}^\top(\hat{\Sigma}^{-1}W_k\hat{\Sigma}^{-1})\mathrm{vec}\,(W_l)$. Since $\hat{\beta}$ is $p^{1/2}$-consistent, it lies in the open ball $U_\delta^{(0)}$ (defined in Proposition 2) as $p \to \infty$. Then, applying the Taylor series expansion of a_{kl} around $\beta^{(0)}$, we obtain that $a_{kl}(\hat{\beta}) = \mathrm{tr}\left(\Sigma_0^{-1}W_k\Sigma_0^{-1}W_l\right) + \partial a_{kl}(\tilde{\beta})/\partial\beta^\top(\hat{\beta} - \beta^{(0)})$, where $\tilde{\beta}$ is between $\hat{\beta}$ and $\beta^{(0)}$. Using Proposition 2 and employing similar techniques to those used in the proofs of Lemma 3 and Proposition 3, we can demonstrate that $p^{-1}\partial a_{kl}(\tilde{\beta})/\partial\beta = O_p(1)$. This, together with Condition (C3), leads to

$$p^{-1}\left(\mathrm{tr}\left(\hat{\Sigma}^{-1}W_k\hat{\Sigma}^{-1}W_l\right)\right)_{(K+1)\times(K+1)} \xrightarrow{p} Q_{-1} > 0. \tag{A.10}$$

We next take the Taylor series expansion of $S_{p,FGLS}(\hat{\beta})$ about $\beta^{(0)}$ and obtain that

$$\begin{aligned} S_{p,FGLS}(\hat{\beta}) &= 2\frac{\partial\ell_p\left(\beta^{(0)}\right)}{\partial\beta} + \frac{\partial S_{p,FGLS}\left(\beta^{(0)}\right)}{\partial\beta^\top}\left(\hat{\beta} - \beta^{(0)}\right) \\ &\quad + \frac{1}{2}\left\{I_{K+1} \otimes \left(\hat{\beta} - \beta^{(0)}\right)^\top\right\}\frac{\partial}{\partial\beta^\top}\mathrm{vec}\left(\frac{\partial S_{p,FGLS}\left(\breve{\beta}\right)}{\partial\beta^\top}\right)\left(\hat{\beta} - \beta^{(0)}\right), \end{aligned}$$

where $\breve{\beta}$ lies between $\hat{\beta}$ and $\beta^{(0)}$. By Lemmas 2 and 3, we can verify that $p^{-1}\partial S_{p,FGLS}(\beta^{(0)})/\partial\beta^\top = o_p(1)$. Furthermore, applying Proposition 2 and employing similar techniques to those used in the proofs of Lemma 3 and Proposition 3, we have $p^{-1}\partial\,\mathrm{vec}(\partial S_{p,FGLS}\left(\breve{\beta}\right)/\partial\beta^\top)/\partial\beta^\top = O_p(1)$. Accordingly, $p^{-1/2}S_{p,FGLS}(\hat{\beta}) = 2p^{-1/2}\partial\ell_p(\beta^{(0)})/\partial\beta + o_p(1)$. This, together with Proposition 3, Slutsky's theorem, and equation (A.10), completes the proof.

References

Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *The Annals of Statistics*, 135–141.

Akerlof, G. A. (1997). Social distance and social decisions. *Econometrica*, 65, 1005–1027.

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

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, 36, 2577–2604.

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics,* 36, 199–227.

Bilmes, J. A. (2000). Factored sparse inverse covariance matrices. *2000 IEEE International Conference*, 2, 1009–1012.

Bivand, R. S., Pebesma, E. J. and Gómez-Rubio, V. (2008). *Applied spatial data analysis with R*. Springer.

Bodie, Z., Kane, A. and Marcus, A. J. (2009). *Investments*. McGraw-Hill.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3, 1–122.

Brock, W. A. and Durlauf, S. N. (2001). Discrete choice with social interactions. *The Review of Economic Studies*, 68, 235–260.

Cai, T. T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106, 672–684.

Cai, T. T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106, 594–607.

Cai, T. T. and Jiang, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *The Annals of Statistics*, 39, 1496–1525.

Chan, L. K., Karceski, J. and Lakonishok, J. (1998). The risk and return from factors. *Journal of Financial and Quantitative Analysis*, 33, 159–188.

Chan, L. K., Karceski, J. and Lakonishok, J. (1999). On portfolio optimization: Forecasting covariances and choosing the risk model. *Review of Financial Studies*, 12, 937–974.

Chen, X. and Conley, T. G. (2001). A new semiparametric spatial model for panel time series. *Journal of Econometrics*, 105, 59–83.

Chiu, T. Y. M., Leonard, T. and Tsui, K.-W. (1996). The matrix-logarithmic covariance model. *Journal of the American Statistical Association*, 91, 198–210.

Cressie, N. (1991). *Statistics for Spatial Data*. Wiley Series in Probability and Statistics.

Cressie, N. and Wikle, C. K. (2011). *Statistics for spatio-temporal data*. John Wiley & Sons, Inc.

DeMiguel, V., Garlappi, L., Nogales, F. J., and Uppal, R. (2009). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55, 798–812.

Demidenko, E. (2004). *Mixed Models – Theory and Applications*. John Wiley & Sons, Inc.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, 157–175.

Diggle, P. J., Heagerty, P. J., Liang, K.-Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd Edition. Oxford University Press.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25, 383–417.

Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147, 186–197.

Fan, J., Han, F. and Liu, H. (2014). PAGE: Robust pattern guided estimation of large covariance matrix. *Technical Report*. Princeton University.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J., Liao, Y. and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39, 3320–3356.

Fletcher, R. (1987). *Practical Methods of Optimization*. John Wiley & Sons, Inc.

Fortin, M. and Glowinski, R. (1983). *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*. Amsterdam: North-Holland.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9, 432–441.

Glaeser, E. L., Sacerdote, B. and Scheinkman, J. A. (1996). Crime and social interactions. *The Quarterly Journal of Economics*, 507–548.

Gul, F. A., Kim, J. B. and Qiu, A. A. (2010). Ownership concentration, foreign shareholding, audit quality, and stock price synchronicity: Evidence from China. *Journal of Financial Economics*, 95, 425–442.

Hoff, P. D. and Niu, X. (2012). A covariance regression model. *Statistica Sinica*, 22, 729–753.

Huang, J. Z., Liu, N., Pourahmadi, M. and Liu, L. (2006). Covariance matrix selection and estimation via penalized normal likelihood. *Biometrika*, 93, 85–98.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, 58, 1651–1684.

Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis*. Englewood Cliffs, NJ: Prentice Hall.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104, 682–693.

Kan, R. and Zhou, G. (2007). Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis*, 42, 621–656.

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

Leng, C. and Li, B. (2011). Forward adaptive banding for estimating large covariance matrices. *Biometrika*, 98, 821–830.

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

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44, 1033–1048.

Qiu, Y. and Chen, S. X. (2012). Test for bandedness of high-dimensional covariance matrices and bandwidth estimation. *The Annals of Statistics*, 40, 1285–1314.

Roll, R. (1988). R^2. *Journal of Finance*, 43, 541–566.

Schott, J. R. (1997). *Matrix Analysis for Statistics*. John Wiley & Sons, Inc.

Scott, J. (1988). Social network analysis. *Sociology*, 22, 109–127.

Seber, G. A. (2008). *A Matrix Handbook for Statisticians*. John Wiley & Sons, Inc.

Szatrowski, T. H. (1980). Necessary and sufficient conditions for explicit solutions in the multivariate normal estimation problem for patterned means and covariances. *The Annals of Statistics*, 8, 802–810.

Tong, T. and Wang, Y. (2007). Optimal shrinkage estimation of variances with applications to microarray data analysis. *Journal of the American Statistical Association*, 102, 113–122.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104, 1512–1524.

Wasserman, S. (1994). *Social network analysis: Methods and applications*. Cambridge University Press.

Wooldridge, J. (2012). *Introductory Econometrics: A Modern Approach*. Cengage Learning.

Xue, L., Ma, S. and Zou, H. (2012). Positive-definite ℓ_1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, 107, 1480–1491.

Yin, J., Geng, Z., Li, R. and Wang, H. (2010). Nonparametric covariance model. *Statistica Sinica*, 20, 469–479.

Zivot, E. and Wang, J. (2007). *Modeling Financial Time Series with S-Plus*. Springer.

Zwiernik, P., Uhler, C. and Richards, D. (2014). Maximum likelihood estimation for linear gaussian covariance models. *arXiv preprint*, 1408.5604.

Table 1: Comparison of MLE, OLS, OLS+, FGLS, and FGLS+ covariance matrix estimates. Four measures are considered: the averaged execution time (Time, in seconds), the averaged spectral norm and Frobenius norm estimation errors (Spectral-Error and Frobenius-Error), and the percentage of the unconstrained covariance estimate being identical to its associated constrained estimate (Percentage). The response variable follows the normal distribution and the similarity matrices are W_k.

		$p = 50$	$p = 100$	$p = 200$	$p = 500$
MLE	Time	10.4524	121.0258	1,648.1558	87,417.6100
	Spectral-Error	4.2083	3.0879	2.1557	1.3545
	Frobenius-Error	1.5945	1.0921	0.7761	0.4849
OLS	Time	0.0001	0.0004	0.0014	0.0205
	Spectral-Error	4.4538	3.2327	2.3444	1.4735
	Frobenius-Error	1.6677	1.1293	0.8326	0.5179
	Percentage	91.8%	96.4%	99.6%	99.9%
OLS+	Time	0.0143	0.0429	0.0482	0.2345
	Spectral-Error	4.3420	3.1944	2.3384	1.4728
	Frobenius-Error	1.6348	1.1186	0.8312	0.5178
FGLS	Time	0.1297	1.7018	25.3935	1,785.5373
	Spectral-Error	4.9085	3.3923	2.2311	1.3357
	Frobenius-Error	1.8065	1.1873	0.7989	0.4790
	Percentage	94.6%	97.8%	99.5%	100.0%
FGLS+	Time	0.2532	2.0311	25.9578	1,785.5373
	Spectral-Error	5.1565	3.6692	2.2561	1.3357
	Frobenius-Error	1.9421	1.3093	0.8139	0.4790

Table 2: Comparison of MLE, OLS, OLS+, FGLS, and FGLS+ regression parameter estimates. Three measures are considered: the averaged bias of the estimate (BIAS), the standard deviation of the estimate (SD), and the root mean squared error of the estimate (RMSE).

		$p = 50$			$p = 100$			$p = 200$			$p = 500$		
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
MLE	BIAS	0.0419	-0.0746	0.0065	-0.0104	-0.0327	-0.0051	-0.0050	-0.0142	0.0029	0.0005	0.0050	0.0007
	SD	1.0660	0.4278	0.3476	0.7377	0.2900	0.2270	0.5327	0.2069	0.1556	0.3330	0.1327	0.0935
	RMSE	1.0668	0.4343	0.3477	0.7378	0.2918	0.2271	0.5327	0.2074	0.1557	0.3330	0.1328	0.0935
OLS	BIAS	-0.0251	-0.0606	-0.0362	-0.0270	-0.0269	-0.0225	-0.0092	-0.0160	0.0038	-0.0032	-0.0004	-0.0007
	SD	1.0876	0.4889	0.3487	0.7472	0.3157	0.2250	0.5339	0.2358	0.1680	0.3292	0.1467	0.1032
	RMSE	1.0879	0.4927	0.3506	0.7477	0.3169	0.2261	0.5340	0.2363	0.1680	0.3292	0.1467	0.1032
OLS+	BIAS	-0.0013	-0.0770	-0.0388	-0.0205	-0.0321	-0.0240	-0.0082	-0.0167	0.0035	-0.0031	-0.0005	-0.0008
	SD	1.1123	0.4512	0.3359	0.7524	0.3031	0.2217	0.5354	0.2337	0.1671	0.3292	0.1466	0.1030
	RMSE	1.1123	0.4578	0.3382	0.7527	0.3048	0.2230	0.5355	0.2343	0.1671	0.3292	0.1466	0.1030
FGLS	BIAS	-0.1400	-0.1488	-0.0663	-0.0688	-0.0753	-0.0365	-0.0369	-0.0475	-0.0094	-0.0108	-0.0080	-0.0037
	SD	1.1747	0.4808	0.4221	0.7662	0.3106	0.2324	0.5363	0.2128	0.1534	0.3291	0.1300	0.0935
	RMSE	1.1830	0.5033	0.4273	0.7693	0.3196	0.2352	0.5376	0.2180	0.1537	0.3293	0.1303	0.0936
FGLS+	BIAS	-0.0395	-0.1357	-0.0546	-0.0454	-0.0733	-0.0330	-0.0361	-0.0485	-0.0088	-0.0108	-0.0080	-0.0037
	SD	1.2248	0.4525	0.3781	0.7963	0.3069	0.2262	0.5363	0.2149	0.1545	0.3291	0.1300	0.0935
	RMSE	1.2255	0.4724	0.3820	0.7976	0.3155	0.2286	0.5375	0.2203	0.1547	0.3293	0.1303	0.0936

Table 3: Comparison of QMLE, OLS, OLS+, FGLS, and FGLS+ covariance matrix estimates. Four measures are considered: the averaged execution time (Time, in seconds), the averaged spectral norm and Frobenius norm estimation errors (Spectral-Error and Frobenius-Error), and the percentage of the unconstrained covariance estimate being identical to its associated constrained estimate (Percentage). The response variable follows the mixture normal distribution and the similarity matrices are W_k. Dashes indicate procedures that were not executed due to prohibitively time-intensity.

		$p = 50$	$p = 100$	$p = 200$	$p = 500$
QMLE	Time	13.9023	141.1893	2,041.2088	-
	Spectral-Error	4.8963	3.6173	2.6006	-
	Frobenius-Error	2.0376	1.4781	1.0503	-
OLS	Time	0.0003	0.0006	0.0025	0.0259
	Spectral-Error	5.3133	3.7689	2.7510	1.7866
	Frobenius-Error	2.1704	1.5279	1.0900	0.6973
	Percentage	89.7%	95.8%	99.4%	100.0%
OLS+	Time	0.0404	0.0741	0.0666	0.0259
	Spectral-Error	5.1475	3.7239	2.7490	1.7866
	Frobenius-Error	2.1238	1.5159	1.0895	0.6973
FGLS	Time	0.3057	2.6710	40.6825	1,949.7568
	Spectral-Error	5.5171	3.9104	2.6628	1.6208
	Frobenius-Error	2.2401	1.5625	1.0727	0.6563
	Percentage	95.5%	98.3%	99.5%	100.0%
FGLS+	Time	0.5002	3.0632	41.4874	1,949.7568
	Spectral-Error	5.5521	4.1598	2.6578	1.6208
	Frobenius-Error	2.2821	1.6675	1.0813	0.6563

Table 4: Comparison of MLE, OLS, OLS+, FGLS, and FGLS+ covariance matrix estimates. Four measures are considered: the averaged execution time (Time, in seconds), the averaged spectral norm and Frobenius norm estimation errors (Spectral-Error and Frobenius-Error), and the percentage of the unconstrained covariance estimate being identical to its associated constrained estimate (Percentage). The response variable follows the normal distribution and the similarity matrices are $\overline{W}_k$. Dashes indicate procedures that were not executed due to prohibitively time-intensity.

		$p = 50$	$p = 100$	$p = 200$	$p = 500$
MLE	Time (s)	6.1404	123.9995	2,158.6198	-
	Spectral-Error	4.7748	4.3375	3.3905	-
	Frobenius-Error	1.6075	1.1541	0.7830	-
OLS	Time (s)	0.0001	0.0004	0.0016	0.0211
	Spectral-Error	5.1860	4.9735	4.7913	4.1025
	Frobenius-Error	1.7042	1.2624	0.9628	0.6093
	Percentage	91.9%	95.4%	93.4%	96.0%
OLS+	Time (s)	0.0141	0.0595	0.7594	13.8428
	Spectral-Error	5.0633	4.8603	4.5276	4.0160
	Frobenius-Error	1.6760	1.2430	0.9316	0.6019
FGLS	Time (s)	0.1293	1.6756	26.1394	1,806.5509
	Spectral-Error	5.9917	4.9837	4.2002	2.8152
	Frobenius-Error	1.9239	1.2799	0.9087	0.5049
	Percentage	95.7%	98.4%	99.1%	99.5%
FGLS+	Time (s)	0.2185	1.8370	27.0289	1,821.6573
	Spectral-Error	5.8848	4.7254	4.0558	2.7448
	Frobenius-Error	1.9738	1.2423	0.8976	0.4949

Table 5: The quarterly Mean, standard deviation (SD), Sharpe ratio, Alpha, and Beta calculated from the 20 quarters' returns (%) of the market portfolio, constrained portfolio, and unconstrained portfolio, respectively, from 2007 to 2011 on the Chinese stock market. The numbers inside the parentheses are the standard errors of the Alpha and Beta coefficients, respectively.

	Mean	SD	Sharpe Ratio	Alpha	Beta
Market Portfolio Return	2.38	21.76	0.07	0	1
Constrained Portfolio Return	3.57	21.56	0.13	1.26 (1.31)	0.96 (0.06)
Unconstrained Portfolio Return	4.34	22.61	0.16	2.05 (2.19)	0.94 (0.10)

Table 6: The daily Mean, standard deviation (SD), Sharpe ratio, Alpha, and Beta calculated from 1,219 trading days of returns (%) in the market portfolio, constrained portfolio, and unconstrained portfolio, respectively, from 2007 to 2011 on the Chinese stock market. The numbers inside the parentheses are the standard errors of the Alpha and Beta coefficients, respectively.

	Mean	SD	Sharpe Ratio	Alpha	Beta
Market Portfolio Return	0.025	2.165	0.008	0	1
Constrained Portfolio Return	0.072	2.395	0.027	0.046 (0.020)	1.058 (0.009)
Unconstrained Portfolio Return	0.075	2.466	0.027	0.049 (0.024)	1.069 (0.011)

Table 7: Asymptotic results of the MLE, OLS, and FGLS estimators based on (7.1), (7.2) and (7.3), respectively. Note that Condition (C3′) is slightly modified from Condition (C3), to be "For fixed $p \geq 1$, the matrix $\mathcal{Q}_{p,d} := p^{-1}(\mathrm{tr}\left(\Sigma_0^d W_k \Sigma_0^d W_l\right))_{(K+1)\times(K+1)}$ is positive definite when $d = -2, -1, 0, 1$, and $\Sigma_0^0 := I_p$." The proofs of asymptotic results can be found in the supplementary material. It is worth noting that $O_p(n^{-1/2}p^{-1/2}) = O_p(p^{-1/2})$ when n is fixed and $p \to \infty$ and $O_p(n^{-1/2}p^{-1/2}) = O_p(n^{-1/2})$ when $n \to \infty$ and p is fixed.

Estimators \ Conditions	fixed n and $p \to \infty$ (or $n \to \infty$ and $p \to \infty$) Conditions (C1) – (C3)	fixed p and $n \to \infty$ Conditions (C2) and (C3′)
$\hat{\beta}_{np,MLE}$	$\sqrt{np}(\hat{\beta}_{np,MLE} - \beta^{(0)}) \xrightarrow{d} N(0, 2\mathcal{Q}_{-1}^{-1})$	$\sqrt{np}(\hat{\beta}_{np,MLE} - \beta^{(0)}) \xrightarrow{d} N(0, 2\mathcal{Q}_{p,-1}^{-1})$
$\hat{\beta}_{np,OLS}$	$\sqrt{np}(\hat{\beta}_{np,OLS} - \beta^{(0)}) \xrightarrow{d} N(0, 2\mathcal{Q}_0^{-1}\mathcal{Q}_1\mathcal{Q}_0^{-1})$	$\sqrt{np}(\hat{\beta}_{np,OLS} - \beta^{(0)}) \xrightarrow{d} N(0, 2\mathcal{Q}_{p,0}^{-1}\mathcal{Q}_{p,1}\mathcal{Q}_{p,0}^{-1})$
$\hat{\beta}_{np,FGLS}$	$\sqrt{np}(\hat{\beta}_{np,FGLS} - \beta^{(0)}) \xrightarrow{d} N(0, 2\mathcal{Q}_{-1}^{-1})$	$\sqrt{np}(\hat{\beta}_{np,FGLS} - \beta^{(0)}) \xrightarrow{d} N(0, 2\mathcal{Q}_{p,-1}^{-1})$
$\Sigma(\hat{\beta})$	$\Vert\Sigma(\hat{\beta}) - \Sigma_0\Vert_2 = O_p(n^{-1/2}p^{-1/2})$ $n\Vert\Sigma(\hat{\beta}) - \Sigma_0\Vert_F^2 \xrightarrow{d} \mathcal{Z}^\top \mathcal{Q}_0 \mathcal{Z}$ and $\mathcal{Z}$ is some random vector asymptotically converged by $\sqrt{np}(\hat{\beta} - \beta^{(0)})$	$\Vert\Sigma(\hat{\beta}) - \Sigma_0\Vert_2 = O_p(n^{-1/2}p^{-1/2})$ $n\Vert\Sigma(\hat{\beta}) - \Sigma_0\Vert_F^2 \xrightarrow{d} \mathcal{Z}^\top \mathcal{Q}_{p,0} \mathcal{Z}$
$\Sigma^{-1}(\hat{\beta})$	$\Vert\Sigma^{-1}(\hat{\beta}) - \Sigma_0^{-1}\Vert_2 = O_p(n^{-1/2}p^{-1/2})$ $n\Vert\Sigma^{-1}(\hat{\beta}) - \Sigma_0^{-1}\Vert_F^2 \xrightarrow{d} \mathcal{Z}^\top \mathcal{Q}_{-2} \mathcal{Z}$	$\Vert\Sigma^{-1}(\hat{\beta}) - \Sigma_0^{-1}\Vert_2 = O_p(n^{-1/2}p^{-1/2})$ $n\Vert\Sigma^{-1}(\hat{\beta}) - \Sigma_0^{-1}\Vert_F^2 \xrightarrow{d} \mathcal{Z}^\top \mathcal{Q}_{p,-2} \mathcal{Z}$

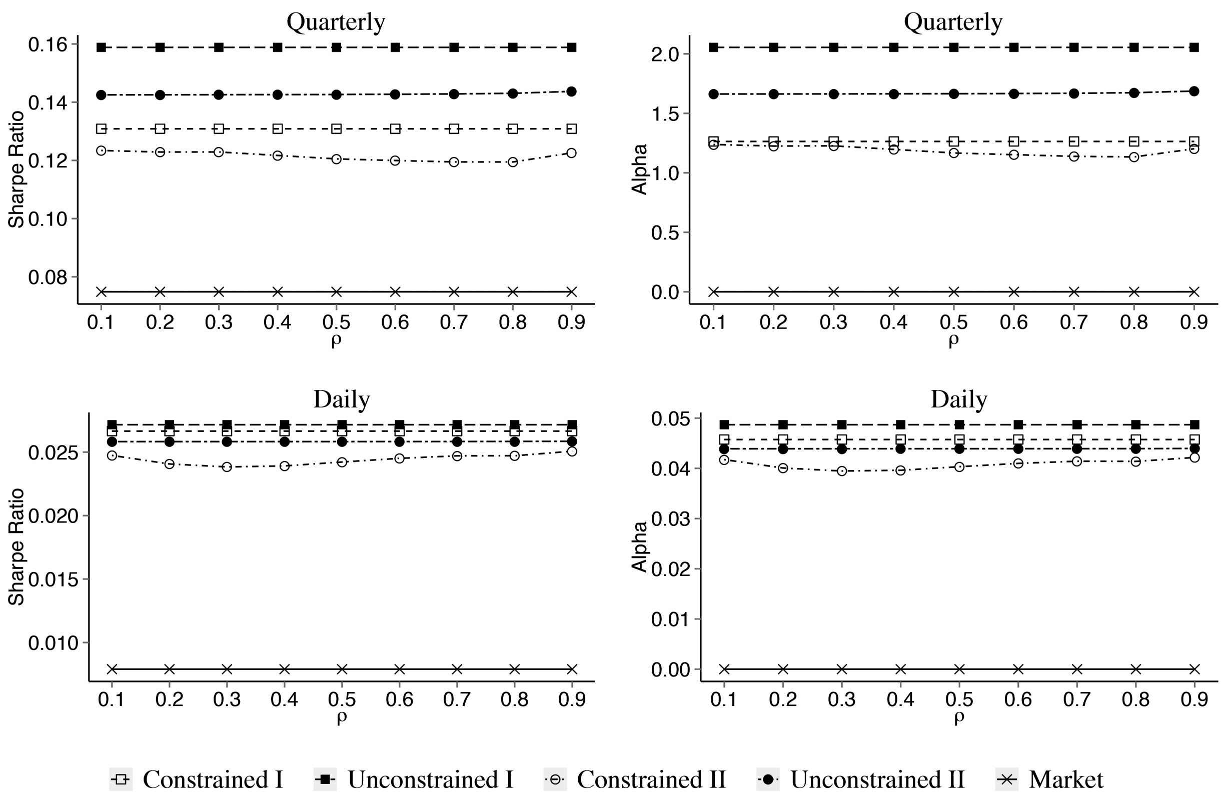

Figure 1: The quarterly and daily Sharpe ratios and Alpha values of the five portfolio returns. The results of Constrained I and Unconstrained I are obtained, correspondingly, from the constrained and unconstrained portfolio returns constructed via our proposed covariance estimator. The results of Constrained II and Unconstrained II are obtained, respectively, from the constrained and unconstrained portfolio returns constructed by the shrinkage covariance estimator $\overline{\Sigma}_t = \rho \mathrm{tr}(\overline{S}_t) I_p / p + (1-\rho)\overline{S}_t$ for $\rho = 0.1, 0.2, \cdots, 0.9$. Market is the market portfolio return. Note that the calculations of Constrained I, Unconstrained I, and Market are independent of ρ.