Covariance Model with General Linear Structure and Divergent Parameters

Xinyan Fan\(^a,\) Wei Lan\(^b,\) Tao Zou\(^c,\) and Chih-Ling Tsai\(^d\)

\(^a\)Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing, China; \(^b\)School of Statistics and Center of Statistical Research, Southwestern University of Finance and Economics, Chengdu, China; \(^c\)Research School of Finance, Actuarial Studies & Statistics, Australian National University, Canberra, Australia; \(^d\)Graduate School of Management, University of California, Davis, Davis, CA

\section{Introduction}

The estimation of a large covariance matrix has played a prominent role in modern multivariate analysis and its applications across various fields such as risk management, portfolio allocation, biostatistics, social networks and health sciences (see, e.g., Kan and Zhou 2007; Friedman, Hastie, and Tibshirani 2008; Yuan and Lin 2007; Fan, Han, and Liu 2014). When the dimension of the covariance matrix is large, the classical sample covariance matrix estimator does not perform well and yields a non-negligible estimation error (see, e.g., Marchenko and Pastur 1967; Battey 2017). To tackle this issue, one possible approach is to directly impose the sparsity assumption on the covariance matrix (see, e.g., Bickel and Levina 2008a, 2008b; Cai and Liu 2011; Goes, Lerman, and Nadler 2020) or its inverse (see, e.g., Yuan and Lin 2007; Friedman, Hastie, and Tibshirani 2008; Cai, Liu, and Luo 2011). Another approach is to consider the dimension reduction of a covariance matrix, such as the low rank or factor structure (see, e.g., Fan, Fan, and Lv 2008; Fan, Liao, and Mincheva 2011; Chang, Bin, and Yao 2015; Fan, Liu, and Wang 2018). Other possible approaches, such as shrinking eigenvalues, can be found in Ledoit and Wolf (2012) and a seminal book of Pourahmadi (2013).

It is worth noting that the aforementioned approaches require the number of observations to tend toward infinity to ensure a reliable estimator. In practice, this requirement is not necessarily valid. Hence, various covariance structures have been proposed to estimate the covariance matrix with limited sample size; see, for example the compound symmetry (Driessen, Maenhout, and Vilkov 2009), Toeplitz (Fuhrmann and Miller 1988), banded, autoregressive, and moving average structures. Note that the above structures can be considered as a special case of a linear combination of pre-specified known matrices such as weight and design matrices; see Zou et al. (2017) and Zheng et al. (2019). We name this the linear covariance model, and related references can be found in Anderson (1973), Pourahmadi (2013), Banerjee et al. (2015), Lan et al. (2018), and Niu and Hoff (2019).

In addition to linear covariance models, some nonlinear covariance models were proposed to build up a nonlinear relationship between the covariance matrix and a linear combination of known matrices. For example, Chiu, Leonard, and Tsui (1996), Pourahmadi (2011), and Battey (2017) modeled the logarithm of a covariance matrix as a linear function of known matrices to assure the positive definiteness of the covariance matrix. In addition, modeling the inverse covariance matrix can be found in Pourahmadi (2011). It is worth noting that several types of multivariate models, such as the spatial autoregressive models (see, e.g., LeSage and Pace 2010) and structured vector autoregressive models (see, e.g., Zhu et al. 2017), can be special cases of the covariance matrix model with a nonlinear structure. The above linear and nonlinear covariance models motivate us to establish a unified framework to study the structured covariance matrix by linking the covariance matrix of response...
variables to the linear combination of weight matrices with the general link function. We name it the covariance model with general linear structure (CMGL). As a result, the linear, logarithm, inverse covariance models and covariance models induced by spatial autoregressive models can be viewed as special cases of CMGL.

The above articles collectively established the relationship between the covariance and the linear combination of known matrices and studied parameter estimations, statistical inference and empirical applications. Those papers usually assume that the responses follow some distribution function (e.g., the normal distribution) or that the number of known matrices is fixed. In practice, however, the data may not satisfy the normality assumption and the number of known (weight or design) matrices can be large. A typical example is modeling the covariance of stock returns in portfolio management. First, the stock returns can be large. A typical example is modeling the covariance of stock returns in portfolio management. First, the stock returns are unlikely to be normally distributed (see, e.g., Karoglou 2010; Kan and Zhou 2017). Second, the weight matrices induced by the firms’ fundamentals can be large (see, e.g., 447 covariates were considered for modeling stock returns in Hou, Xue, and Zhang 2020). These motivate us to explore the following four topics to broaden the usefulness of the covariance model. First, relax the distribution assumption imposed on the continuous response variable and consider employing the quasi-likelihood function (or quadratic loss function) to obtain parameter estimators; second, allow the number of known matrices to be divergent; third, select relevant weight matrices from a large number of candidate matrices; fourth, test the adequacy of link functions.

To accomplish these four tasks, we face three challenges. (i) Obtain the theoretical properties of estimated coefficients associated with the weight matrices when the number of coefficients diverges. Note that the score function and the Hessian matrix of the quasi-likelihood function (or quadratic loss function) used in parameter estimation are divergent. As a result, the traditional pointwise convergence approach is not applicable. To overcome this challenge, we propose using the vector and matrix norms to evaluate the orders of the score function and the Hessian matrix, and then achieve our goal. (ii) Develop the covariance model selection criterion with a diverging number of weight matrices. Based on our knowledge, there is no existing method that can be applied directly. For example, Chen and Chen’s (2008, 2012) selection methods are designed for mean regression models such as linear and generalized linear regression models in high dimensional data. However, the quasi-likelihood function of regression parameters in our setting is associated with the covariance matrices rather than the covariates in mean regression models. Accordingly, the quadratic forms of the score function obtained from the quasi-likelihood function are correlated, which means they are not independent as required in Chen and Chen (2008, 2012). To this end, we propose a novel approach by introducing the tail probability of the maximum of quadratic forms to solve this problem. (iii) Test the adequacy of nonnested link functions. There are various methods being proposed to test nonnested mean regression models. For example, Vuong (1989) and Clarke (2007) obtained tests by assuming the responses are independent. However, the responses of covariance models usually do not satisfy this assumption. In addition, there is no other available method that can be adapted. Hence, we propose a test statistic to assess nonnested link functions, and we obtain its asymptotic property.

The aim of this article is to develop parameter estimation, model selection, and link function testing for covariance models with general linear structure (CMGL) with a diverging number of parameters without assuming a distribution function imposed on responses. Specifically, we investigate the asymptotic properties of the quasi-maximum likelihood estimator (QMILE). Under a special consideration with the linear covariance model, the ordinary least squares estimator (OLS) with a closed form is obtained. The asymptotic properties of the above two types estimator are established when both the covariance dimension $p$ and the number of weight matrices $K$ tend to infinity. Subsequently, we propose an extended Bayesian information criteria (EBIC) for model selections when the number of weight matrices diverges. In addition, the consistency of EBIC is obtained. Finally, we introduce a nonnested test statistic to examine the adequacy of the link function in CMGL.

The remainder of this article is organized as follows. Section 2 introduces covariance models with general linear structure and obtains the quasi-maximum likelihood estimator. Then, that section proposes an extended Bayesian information criteria for covariance model selections. Section 3 presents the ordinary least squares estimator and model selections under linear covariance models. Section 4 provides a test for assessing nonnested link functions. Simulation studies and an empirical example are given in Section 5. Section 6 concludes the article with short discussions. All theoretical proofs are relegated to the Appendix and supplementary materials.

2. Model Estimation and Selection

2.1. Model Estimation

Let $Y = (Y_1, \ldots, Y_p)^\top \in \mathbb{R}^p$ be the $p$-dimensional response vector, $x_j = (x_{j1}, \ldots, x_{jk})^\top \in \mathbb{R}^K$ be the $K$-dimensional covariate vector for $j = 1, \ldots, p$, and $X_k = (x_{k1}, \ldots, x_{kp})^\top \in \mathbb{R}^p$ for $k = 1, \ldots, K$. In addition, assume that $K$ diverges along with $p \to \infty$. We then adapt the approach of Johnson and Wichern (1992) and Zou et al. (2017) to construct the weight matrix $W_k$ induced by covariate $X_k$ for $k = 1, \ldots, K$. Specifically, for continuous $X_k$, the $(j_1, j_2)$th element of $W_k = (w_{k,j_1,j_2})$ is defined as $w_{k,j_1,j_2} = \exp \{- (x_{j_1} - x_{j_2})^2 \}$, while the diagonal elements of $W_k$ are set to be 0. As for discrete $X_k$, define $w_{k,j_1,j_2} = 1$ if $x_{j_1,k}$ and $x_{j_2,k}$ belong to same group and $w_{k,j_1,j_2} = 0$ otherwise.

To link the covariance matrix of $Y$ to the linear combination of weight matrices, we propose the covariance model with general linear structure (CMGL) given below,

$$
\text{cov}(Y) = \Sigma(\beta) = G(\beta_0^0 + W_1 \beta_1 + \cdots + W_K \beta_K) =: G(\beta),
$$

(1)

where the response $Y$ is standardized to have mean zero, $I_p$ is the $p \times p$ identity matrix, $\beta = (\beta_0, \ldots, \beta_K) \in \mathbb{R}^{K+1}$, and $G(\cdot)$ is a known function mapping from $\mathbb{R}^{p \times p}$ to $\mathbb{R}^{p \times p}$, such that the output is a positive definite matrix. The detailed illustrations of $G$ and its derivatives with respect to $\beta$ are given in Section S.1 of the supplementary materials. Denote $\beta_0^0$ the true vector of parameters and $\Sigma_0 := \Sigma(\beta_0^0)$. Since $Y$ is standardized to have
mean zero, the true covariance matrix of $Y$ is $E(YY^\top) = \Sigma_0 = \Sigma(\beta_0^*)$.

The model (1) comprises various structured covariance matrices as its special cases; see the following three examples.

(i.) $G(B) = I(B) = B$, where $I$ is an identity mapping function and it also is called the linear function. Accordingly, $\text{cov}(Y) = I_p \beta_0 + W_1 \beta_1 + \ldots + W_K \beta_K$, which is the linear covariance model. By appropriately choosing the weights, the linear covariance models comprise the compound symmetry, Toeplitz, banded, autoregressive (AR) and moving average (MA) structures as its special cases (see, e.g., Zou et al. 2017; Zheng et al. 2019). (ii.) $G(B) = \exp(B)$. Then, model (1) links the logarithm of the covariance matrix with a linear combination of weight matrices (see, e.g., Chiu, Leonard, and Tsui 1996; Pourrahmadi 2000; Battey 2017). (iii.) $G(B) = B^{-1}(B^{-1})^\top$. The covariance functions for the spatial autoregressive models is proportion to $(I_p - \beta_1 W_1)^{-1}(I_p - \beta_2 W_1^\top)^{-1}$, where $W_1$ is the spatial weight matrix and $\beta_1$ is the spatial autocorrelation parameter (see, e.g., LeSage and Pace 2010).

In this article, we do not make any distribution assumption for responses. Hence, we employ the quasi-maximum likelihood approach (see, e.g., Lee 2004; Tsay 2014) to estimate unknown parameters. Based on (1), the quasi-loglikelihood function of $\beta$ is

$$
\ell_Q(\beta) = \frac{-p}{2} \log(2\pi) - \frac{1}{2} \sum_{j=1}^p \sum_{j=1}^p \log |\lambda_j(\Sigma(\beta))| - \frac{1}{2} Y^\top \Sigma^{-1}(\beta) Y,
$$

(2)

where $\lambda_j(\Sigma(\beta))$ represents the $j$th largest eigenvalue of $\Sigma(\beta)$. Let $B = \{\beta : \Sigma(\beta) > 0\}$ be the parameter space such that $\Sigma(\beta)$ is a positive definite matrix. Then the quasi-maximum likelihood estimator (QMLE), $\hat{\beta}_Q$, can be obtained by maximizing (2). That is $\hat{\beta}_Q = \arg\max_{\beta \in B} \ell_Q(\beta)$, which can be calculated via the Newton-type algorithm; see, for example, Section 4.2 of Dennis and Schnabel (1996). The Newton-type algorithm is implemented via the function "nlm" in R, and the initial parameter $\beta(0) = (\beta_1(0), \beta_2(0), \ldots, \beta_d(0))^\top$ is set to be $\beta_j(0) = 0$ for $j = 1, \ldots, d$ and $\beta_0(0) = \arg\min_{\beta_0} (\|YY^\top - G(I_p \beta_0)\|^2_F)$.

After simple calculations, $\beta_0(0) = p^{-1} \sum_{i=1}^p Y_i^2$ for the linear link function and $\beta_0(0) = \log(p^{-1} \sum_{i=1}^p Y_i^2)$ for the exponential link function.

2.2. Asymptotic Properties of QMLE

To study the asymptotic property of QMLE, we first define some notation. For any generic matrix $D$, let $\lambda_j(D), \|D\|_2 = (\lambda_1(D^\top D))^1/2$, $\|D\|_F = (\text{tr}(D^\top D))^{1/2}$, and $\|D\|_1 = \max_j \|D_j\|$ denote the $j$th largest eigenvalue, the spectral norm, Frobenius norm, and $\ell_1$ norm of $D$, respectively. In addition, for any matrix $H_{kl}$ ($k,l = 0, \ldots, K$), let $(\text{tr}(H_{kl}))_{(K+1)\times(K+1)}$ denote a $(K+1) \times (K+1)$ matrix, whose $(k,l)$th element is $\text{tr}(H_{kl})$. Moreover, for any matrices $A = (a_{ij})$ and $B = (b_{ij})$ with the same dimension, $A \otimes B = (a_{ij} b_{ij})$ denotes the Hadamard product. We next introduce five technical conditions.

**Condition 1.** The dimension of $K$ satisfies $K = O(p^\kappa)$ for some $\kappa < 1/4$ as $p \to \infty$.

**Condition 2.** Assume $Y = \Sigma_0^{-1/2}Z$, where $Z = (Z_1, \ldots, Z_p)^\top$ satisfies:

(i) $E(Z_0^n) = \mu(0)$ with $\mu(0) = 0$ and $\mu(2) = 1$;
(ii) $E(Z_{1i}^n \mid F_{i-1}) = E(Z_{1i}^n)$ for $n = 1, \ldots, 4$, where $F_i$ is the $\sigma$-field generated by $\{Z_{ij} : j = 1, \ldots, i\}$;
(iii) $E(Z_{1i}^n Z_{1j}^m Z_{1k}^l Z_{1l}^s \mid F_{i-1}) = E(Z_{1i}^n)E(Z_{1j}^m)E(Z_{1k}^l)E(Z_{1l}^s)$ for any $i_1 \neq i_2 \neq i_3 \neq i_4$ and $v_1 + v_2 + v_3 + v_4 \leq 8$;
(iv) there exists a finite positive constant $c$, such that $\text{Pr}(|\sigma(\Sigma(\beta) - E(\Sigma(\beta)))| > t) \leq 2 \exp(-t^2/2)$ for any $1$-Lipschitz convex function $\sigma : \mathbb{R}^d \to \mathbb{R}$ and $t > 0$.

**Condition 3.** For any symmetric matrices in $\{\Sigma_1(\beta)\}$, $k = 0, \ldots, K$, $\{\Sigma_2(\beta)\}$, $j, k = 0, \ldots, K$ and $\{\Sigma_3(\beta)\}$, $j, k, l = 0, \ldots, K$, there exists $\tau_0 > 0$ and an open ball $U_\beta = \{\beta : \|\beta - \beta_0\| < \delta\}$ with $\delta > 0$, such that

$$
\max_{j,k,\beta \in U_\beta} \left\| \frac{\partial \Sigma(\beta)}{\partial \beta_j} \right\|, \sup_{j,k,l,\beta \in U_\beta} \left( \left\| \frac{\partial^2 \Sigma(\beta)}{\partial \beta_j \partial \beta_k} \right\| \right) \leq \tau_0 < \infty.
$$

**Condition 4.** There exist three finite positive constants $\tau_1, \tau_2$, and $\tau_3$ such that, for any $p \geq 1, 0 < \tau_1 < \inf_{\beta \in U_\beta} \lambda_{1}(\Sigma(\beta)) \leq \sup_{\beta \in U_\beta} \lambda_{1}(\Sigma(\beta)) < \tau_2 < \infty$ and $\sup_{\beta \in U_\beta} (\|\Sigma^{-1/2}(\beta)\|_1 + \|\Sigma^{-1/2}(\beta)\|_1) < \tau_3$, where $U_\beta$ is defined in Condition 3.

**Condition 5.** Assume that

(i) $p^{-1}(\text{tr}(\Sigma_0^{-1} \frac{\partial \Sigma_0(\beta_0)}{\partial \beta_j} \Sigma_0^{-1} \frac{\partial \Sigma_0(\beta_0)}{\partial \beta_l}))_{(K+1)\times(K+1)}$ converges to a finite positive definite matrix $Q$ in the Frobenius norm, that is,

$$
p^{-1} \left( \text{tr}(\Sigma_0^{-1} \frac{\partial \Sigma_0(\beta_0)}{\partial \beta_j} \Sigma_0^{-1} \frac{\partial \Sigma_0(\beta_0)}{\partial \beta_l}) \right)_{(K+1)\times(K+1)} - Q_F = o(1),
$$

where $\lambda_{K+1}(Q) > \phi_0 > 0$ for a positive constant $\phi_0$;

(ii) $p^{-1}(\text{tr}(\Sigma_0^{-1/2} \frac{\partial \Sigma_0(\beta_0)}{\partial \beta_j} \Sigma_0^{-1/2} \frac{\partial \Sigma_0(\beta_0)}{\partial \beta_l}))_{(K+1)\times(K+1)}$ converges to a matrix $\Delta$ in the Frobenius norm, and $0 < \eta_1 < \lambda_{K+1}(Q + \mu(4) - 3) \Delta_1 = O(K)$, where $\eta_1$ is a finite constant and $\mu(4) = E(Z_i^4)$.

The above conditions are sensible and mild. **Condition 1** assumes that the number of parameters $K$ tends to infinity at a slow rate of $p$. **Conditions 2(i) to 2(iii)** are moment conditions. **Condition 2(iv)** is a sub-Gaussian type condition that requires $Z$ to have the convex concentration property (see Adamczak 2015), and it is weaker than the commonly used normal condition. **Condition 3** assumes the boundedness of the $\ell_1$ norm for the first, second and third derivatives of $\Sigma(\beta)$. **Condition 4** guarantees the invertibility of $\Sigma(\beta)$ around $U_\beta$. **Condition 5** ensures the convergence and positive definiteness of the Hessian matrix at the true value $\beta_0$, and it is critical for evaluating the asymptotic covariance of $\hat{\beta}_Q$. Based on the above conditions, we obtain the asymptotic property of $\hat{\beta}_Q$ below.
Theorem 1. Under Conditions 1–5, we obtain that, as \( p \to \infty \),
\[
\sqrt{p/K}AQ(\hat{\beta}_0 - \beta^0) \xrightarrow{d} N(0, I_Q),
\]
where \( A \) is an arbitrary \( M \times (K + 1) \) matrix satisfying \( M < \infty \), \( \|A\|_2 \to \infty \) and \( K^{-1}A[2Q + (\mu(4) - 3)\Delta]A^\top \to I_Q \), and \( I_Q \) is a \( M \times M \) nonnegative symmetric matrix.

The nonrandom matrix \( A \) is introduced in order to project the infinite dimension of covariance regression coefficients into a finite dimension. Accordingly, the asymptotic distribution of regression estimators can be obtained by employing the properties of the finite-dimensional random variables based on the Cramér-Wold device. Similar techniques can be found in the existing literature; see, for example, Fan and Peng (2004) and Gupta and Robinson (2018). The above theorem indicates that \( \hat{\beta}_0 \) is the \((p/K)^{1/2}\)-consistent estimator and is asymptotically normal. Based on this finding, we can estimate the covariance matrix by \( \Sigma(\hat{\beta}_0) \), whose asymptotic properties are given below.

Theorem 2. Under Conditions 1–5, we obtain that, as \( p \to \infty \),
\[
\|\Sigma(\hat{\beta}_0) - \Sigma_0\|_2 = O_p(Kp^{-1/2}) \quad \text{and} \quad \|\Sigma^{-1}(\hat{\beta}_0) - \Sigma_0^{-1}\|_2 = O_p(Kp^{-1/2}).
\]

The above theorem demonstrates that \( \Sigma(\hat{\beta}_0) \) and \( \Sigma^{-1}(\hat{\beta}_0) \) are consistent estimators of \( \Sigma_0 \) and \( \Sigma_0^{-1} \), respectively. Using the fact that \( p^{-1/2}\|\Sigma(\hat{\beta}_0) - \Sigma_0\|_F \) and \( p^{-1/2}\|\Sigma^{-1}(\hat{\beta}_0) - \Sigma_0^{-1}\|_F \) are correspondingly less than \( \|\Sigma(\hat{\beta}_0) - \Sigma_0\|_2 \) and \( \|\Sigma^{-1}(\hat{\beta}_0) - \Sigma_0^{-1}\|_2 \), their orders are not larger than \( O_p(Kp^{-1/2}) \). As a result, \( \Sigma(\hat{\beta}_0) \) and \( \Sigma^{-1}(\hat{\beta}_0) \) are consistent in Frobenius norm. Note that the asymptotic covariance of \( \hat{\beta}_0 \) are related to \( \mu(6) \), \( Q \), and \( \Delta \), which can be consistently estimated by \( \hat{\mu}^{(4)} = p^{-1} \sum_{i=1}^p \hat{Z}_i^4 \), \( \hat{Q} = p^{-1}(\text{tr}(\Sigma^{-1}(\hat{\beta}_0) - \Sigma_0^{-1}(\hat{\beta}_0)))(\hat{\Sigma}^{-1}(\hat{\beta}_0) - \Sigma_0^{-1}(\hat{\beta}_0))(\hat{\Sigma}^{-1}(\hat{\beta}_0) - \Sigma_0^{-1}(\hat{\beta}_0))^{\top} \), and \( \hat{\Delta} = p^{-1}(\text{tr}(\Sigma^{-1}(\hat{\beta}_0) - \Sigma_0^{-1}(\hat{\beta}_0))\Sigma^{-1/2}(\hat{\beta}_0))\sigma(\Sigma^{-1}(\hat{\beta}_0) - \Sigma_0^{-1}(\hat{\beta}_0))(\hat{\Sigma}^{-1/2}(\hat{\beta}_0))^\top \), where \( \hat{Z} = (\hat{Z}_1, \ldots, \hat{Z}_p)^\top = \Sigma^{-1/2}(\hat{\beta}_0)Y \).

Remark 1. It is worth noting that although the theoretical results of Theorems 1 and 2 are obtained by letting \( p \) go to infinity and \( n \) be fixed, these results can be extended to allow \( np \to \infty \). This extension includes the following two scenarios as special cases: (i) \( n \to \infty \) and \( p \to \infty \); and (ii) \( n \to \infty \) and \( p \) is fixed. To save space, the asymptotic properties of parameter estimators with \( np \to \infty \) are presented in Section S.7 of the supplementary materials.

### 2.3. Extended Bayesian Information Criteria

In this article, the number of weight matrices \( K \) is allowed to tend to infinity. It is natural to consider selecting the relevant weight matrices from a family of candidate models. As noted by Wang, Li, B. and Leng (2009) and Chen and Chen (2008, 2012), the traditional Bayesian information criterion is inconsistent and tends to select too many features when \( K \to \infty \) in linear regression models. To address this issue, Chen and Chen (2008, 2012) introduced the extended Bayesian information criterion (EBIC) for the linear regression model and generalized linear model, respectively. This motivates us to develop an extended Bayesian information criteria for selecting weight matrices.

Let \( s = \{0, \ldots, K\} \) be the full model that contains all weight matrices. In addition, let \( s_0 \subset s \) be the set of weight matrices that are associated with nonzero elements in \( \beta^0 \), and we name this the true model. For any candidate model \( s \subset s_0 \), let \( \beta(s) \) be the vector of the components in \( \beta \) that are associated with weight matrices in \( s \). Denote \( \hat{\beta}_Q(s) \) the QMLE of \( \beta^0(s) \). Then, adapting the approach of Chen and Chen (2008, 2012), we propose EBIC for the weight matrix selection as follows.

\[
\text{EBIC}_Q(s) = -2\ell_Q(\hat{\beta}_Q(s)) + v(s) \log p + 2v(s)\gamma \log K,
\]
where \( \ell_Q(\cdot) \) is defined in (2), \( v(s) \) is the number of weight matrices in \( s \), and \( \gamma > 0 \) is a constant. It is worth noting that \( \ell_Q(\hat{\beta}_Q(s)) \) obtained from (2) is not a sum of iid components, which complicates the proof of consistency in comparison with that of mean regression model selections (see, e.g., Chen and Chen 2008, 2012).

We next present theoretical properties of the proposed EBIC. Before that, we introduce the following notation and two conditions. Analogous to the definition in Chen and Chen (2008), define
\[
A_0(q) = \{s : s_0 \subset s, v(s) \leq q\} \quad \text{and} \quad A_1(q) = \{s : s_0 \not\subset s, v(s) \leq q\}.
\]

For the sake of convenience, denote \( A_0(q) \) and \( A_1(q) \) by \( A_0 \) and \( A_1 \), respectively. Then, we consider the following conditions.

Condition 6. Define \( H_Q(\beta(s)) = \Delta + \partial_2(\ell_Q(\beta(s)))/\partial_2(\beta(s))\partial_2(\beta(s))^\top \). Assume there exists a finite positive constant \( q \) that is larger than \( v(s_0) \). Then, for all \( s \) such that \( v(s) \leq q \), assume that \( p^{-1}H_Q(\beta(s)) - p^{-1}H_Q(\beta(s_0)) \) converges to \( 0 \) uniformly for some positive matrix \( H_Q(\beta(s)) \). In addition, assume that \( 0 < \varphi_1 < \inf_{s \in U_1} \lambda_1(p^{-1}H_Q(\beta(s))) \leq \sup_{s \in U_1} \lambda_1(p^{-1}H_Q(\beta(s))) < \varphi_2 < \infty \) with probability approaching 1, where \( \varphi_1 \) and \( \varphi_2 \) are two finite constants, and \( U_1 \) is defined in Condition 3.

Condition 7. Assume \( \sqrt{p/\log \min_{s \in A_0} \beta^0(s)} \to \infty \) as \( p \to \infty \).

Condition 6 is similar to Condition 5, while it bounds the eigenvalues of the Hessian matrix around \( U_1 \). A similar condition can be found in Chen and Chen (2012). Condition 7 assumes that the order of nonzero coefficients is larger than \( \log p/p \), and similar conditions can be found for linear regression model selections (see, e.g., Wang, Li, B. and Leng 2009).

The above two conditions, together with the previous five conditions, yield the following result.

Theorem 3. Under Conditions 1–7, we have, as \( p \to \infty \),
\[
\text{(i) } \Pr\left\{ \min_{s \in A_0, s \neq s_0} \text{EBIC}_Q(s) \leq \text{EBIC}_Q(s_0) \right\} \to 0 \quad \text{and} \\
\text{(ii) } \Pr\left\{ \min_{s \in A_1} \text{EBIC}_Q(s) \leq \text{EBIC}_Q(s_0) \right\} \to 0,
\]
for \( \gamma > \varphi_4/(8C_1\varphi_1 - 1/2\kappa) \), where \( \kappa \) is defined in Condition 1, \( q \) and \( \varphi_1 \) are defined in Condition 6, and \( c_1 \) and \( \varphi_4 \) are defined in Lemmas 3 and 5, respectively, in the supplementary materials.
The above theorem indicates that EBIC$_Q$ can identify the true model consistently as long as $p$ tends to infinity. The lower bound $\gamma > \varphi_4/(8\varsigma\varphi_1) - 1/(2\kappa)$ is required to avoid possible overfitting. In practice, the magnitude of $\gamma$ is affected by many factors, such as the divergence rate of $K$, which is controlled by $\kappa$; the upper bound of the number of weight matrices in candidate models, which is $g$; and the bounds of eigenvalues of the second derivative matrices of the quasi-loglikelihood function. To implement EBIC$_Q$, we apply the backward elimination method (see, e.g., Zhang and Wang 2011; Schellendorf, Meier, and Bühlmann 2014), which reduces the computational complexity from $2^K$ to $O(K^2)$. Thus, EBIC$_Q$ is computable when $K$ is divergent.

3. Linear Covariance Models

In the context of the generalized covariance model, we have so far studied the quasi-likelihood estimation and selection. However, QMLE does not have a closed form, which increases its computational complexity for large $p$. As noted by Zou et al. (2017), under the linear covariance structure with finite weight matrices, the ordinary least squares estimator (OLS) exists and has a closed form. Hence, its computational complexity is reduced compared to QMLE. This motivates us to investigate OLS for the linear covariance structure when the number of weight matrices $K$ is divergent.

Consider $G$ in (1) being an identity mapping function $I$. We then obtain the linear covariance model, which has the form:

$$
cov(Y) = I_p \beta_0 + W_1 \beta_1 + \cdots + W_K \beta_K. \tag{3}
$$

For the sake of simplicity, we denote $W \tilde{\Sigma} \beta = I_p \beta_0 + W_1 \beta_1 + \cdots + W_K \beta_K$, where $W = (I_p, W_1, \ldots, W_K)$ with $W_i = I_p$. By the definition of $\Sigma(\beta)$, we have $\Sigma(\beta) = W \tilde{\Sigma} \beta$ and $\Sigma_0 = W \tilde{\Sigma} \beta_0$.

For the linear covariance model, we are able to obtain the constrained OLS estimator by minimizing $\|YY^T - \Sigma(\beta)\|_F^2$ with the constraint that $\beta \in B$ (i.e., $\Sigma(\beta)$ is positive definite). To ease calculation, we further consider the OLS estimator without the constraint. The resulting estimator has the closed form given as follows:

$$
\hat{\beta}_{\text{OLS}} = \left(\text{tr}(W_k W_l)\right)^{-1}_{(K+1) \times (K+1)}(Y^TW_k W_l)(Y^TW_k W_l)_{(K+1) \times (K+1)}.
$$

For finite $K$, Zou et al. (2017) demonstrated that, as $p \to \infty$, the unconstrained OLS estimator is identical to the constrained OLS estimator with probability tending to 1. One can show that it is also valid when $K$ is divergent. Accordingly, we refer to the unconstrained ordinary least squares estimator as the OLS estimator. It is worth noting that the calculation of $\hat{\beta}_{\text{OLS}}$ is much simpler than that of $\hat{\beta}_Q$.

To show the theoretical property of OLS, we introduce an additional condition given below.

**Condition 8.** Assume that (i) the matrix $p^{-1}\text{tr}(\Sigma_0^2 W_k^2 \Sigma_0^2 W_l^2)_{(K+1) \times (K+1)}$ converges to a positive definite matrix $Q_\beta$ in the Frobenius norm, that is, $\|p^{-1}\text{tr}(\Sigma_0^2 W_k^2 \Sigma_0^2 W_l^2)_{(K+1) \times (K+1)} - Q_\beta\|_F = o(1)$. In addition, $\lambda_2(Q_\beta) > \varphi_0 > 0$ for $d = 0, 1$, where $\varphi_0$ is a finite constant and $\Sigma_0^2 := I_p$; (ii) the matrix $p^{-1}\text{tr}(\Sigma_0^{1/2} W_k \Sigma_0^{1/2})\text{tr}(\Sigma_0^{1/2} W_l \Sigma_0^{1/2})_{(K+1) \times (K+1)}$ converges to a matrix $\Delta_1$ in the Frobenius norm and $0 < \delta_1 < \lambda_2(2Q_1 + (\mu^{(4)})-3\Delta_1) \leq \lambda_1(2Q_1 + (\mu^{(4)})-3\Delta_1) = O(K)$, where $\delta_1$ is a finite constant and $\mu^{(4)}$ is defined in **Condition 5**.

The above condition is similar to **Condition 5**. It, in conjunction with the first four conditions, yields the following result.

**Theorem 4.** Under **Conditions 1–4** and **Condition 8**, we obtain that, as $p \to \infty$,

$$
\sqrt{p/K}(\hat{\beta}_{\text{OLS}} - \beta^0) \xrightarrow{d} N(0, I_0),
$$

where $A$ is an arbitrary $M \times (K+1)$ matrix for any finite $M$ with $\|A\|_2 < \infty$, $K^{-1}A(2Q_1 + (\mu^{(4)})-3\Delta_1)A^\top \to I_0$, and $I_0$ is a $M \times M$ nonnegative symmetric matrix.

**Theorem 4** indicates that $\hat{\beta}_{\text{OLS}}$ is $(p/K)^{1/2}$ consistent and asymptotically normal. Accordingly, the covariance matrix can be estimated by $\hat{\Sigma}(\hat{\beta}_{\text{OLS}})$. Analogous to **Theorem 2**, we are able to show that $\|\hat{\Sigma}(\hat{\beta}_{\text{OLS}}) - \hat{\Sigma}_0\|_2 = O_p(Kp^{-1/2})$ and $\|\hat{\Sigma}(\hat{\beta}_{\text{OLS}}) - \hat{\Sigma}_0\|_2 = O_p(Kp^{-1/2})$, as $p \to \infty$, under **Conditions 1–4** and **Condition 8**. It is also of interest to note that QMLE is asymptotically more efficient than the OLS estimator when the link function is linear and $\mu^{(4)} = 3$.

Based on OLS, we propose EBIC for linear covariance model selections as follows:

$$
\text{EBIC}_{\text{OLS}}(s) = \log(\hat{\sigma}_s^2) + \hat{\gamma}(s)\log p/p^2 + 2\nu(s)\gamma\log K/p^2,
$$

where $\hat{\gamma}_s^2 = \|YY^T - W_s \tilde{\Sigma}\hat{\beta}_{\text{OLS}}(s)|/\|\Sigma(\hat{\beta}_{\text{OLS}}(s))\|_F^2$, $\hat{\beta}_{\text{OLS}}(s)$ is the OLS estimator of $\beta^0(s)$, and $W_s = (W_k)_{k \in s}$. The theoretical properties of $\text{EBIC}_{\text{OLS}}(s)$ are given below.

**Theorem 5.** Under **Conditions 1–4** and **Conditions 7–8**, we have, as $p \to \infty$,

(i) $\Pr\left\{\min_{s \in A_0, s \neq s_0} \text{EBIC}_{\text{OLS}}(s) \leq \text{EBIC}_{\text{OLS}}(s_0)\right\} \to 0$ and

(ii) $\Pr\left\{\min_{s \in A_1} \text{EBIC}_{\text{OLS}}(s) \leq \text{EBIC}_{\text{OLS}}(s_0)\right\} \to 0$,

for $\gamma > \tau_q^2 \alpha_0 q/(2c_1 a_1 \xi_1) - 1/(2\kappa)$, where $\kappa$ and $\tau_2$ are defined in **Conditions 1 and 4**, respectively, $q$ is a finite constant larger than $\nu(s_0)$ defined in **Section 2.3**, $c_1$ is defined in Lemma 3, $a_0$ and $a_1$ are defined in Lemma 7, $\xi_1$ is defined in Lemma 10, and Lemmas 3, 7, and 10 are presented in the supplementary materials.

The above theorem shows that EBIC$_{\text{OLS}}$ is a consistent selection criterion.

4. Link Function Test

In **Section 3**, we consider the linear covariance model by assuming that the identity link function is adequate. In practice, however, one may not know whether this assumption is valid or not a priori. This motivates us to develop a testing procedure to assess the suitableness of link functions. To this end, we require the replicates of $Y$ to estimate the asymptotic variance of our test presented in **Theorem 6**. It is worth noting that the replicates of
Let $Y_i \in \mathbb{R}^p$ be the $p$-dimensional response vector collected from the $i$th replication for $i = 1, \ldots, n$. In addition, let $G_0$, $G_1$, and $G_2$ be three specific link functions, and denote $\Sigma_G(\beta_G) = G(\mu_{p}\beta_{0G} + W_1\beta_{1G} + \cdots + W_K\beta_{KG})$, the covariance matrix associated with the link function $G \in \{G_0, G_1, G_2\}$ and a set of unknown coefficients $\beta_G = (\beta_{0G}, \beta_{1G}, \cdots, \beta_{KG})'$. Under the link function $G$, $Y_i$ follows a distribution with mean 0 and covariance matrix $\Sigma_G = \Sigma_G(\beta_G)$. The resulting quasi-density function of $Y_i$ is denoted by $h(y, \Sigma) = h(y, \Sigma_G(\beta_G))$, and its corresponding quasi-loglikelihood function is obtained by substituting $\Sigma$ in (2.2) with $\Sigma_G$. Under $G_0$, $G_1$, and $G_2$, we denote their quasi-density functions by $h(y, \Sigma_0) = h(y, \Sigma_{G_0}(\beta_{G_0})), h(y, \Sigma_{G_1}(\beta_{G_1})), \text{ and } h(y, \Sigma_{G_2}(\beta_{G_2}))$, respectively, where $G_0$ denotes the true link function and $\beta_{G_0} = \beta_0$ defined in Section 2.1.

For any two specific link functions $G_1$ and $G_2$, they are not necessarily nested to each other. Hence, we adapt the approach of Vuong (1989) and Clarke (2007) to assess whether $G_1$ and $G_2$ are equally close to $G_0$. Specifically, we employ the Kullback-Leibler information criteria to measure the closeness of two link functions. To this end, define the Kullback-Leibler distance between the link function $G \in \{G_1, G_2\}$ and $G_0$ as $KLIC := E\left[ \log h(Y_i, \Sigma_0) - E \log h(Y_i, \Sigma_G(\beta_G)) \right]$, where $E\left[ \log h(Y_i, \Sigma_G(\beta_G)) \right]$ is evaluated under the true model and

$$\beta^*_G = \text{argmax}_{\beta_G} E \left[ \log h(Y_i, \Sigma_G(\beta_G)) \right]$$

$$= \text{argmax}_{\beta_G} \int_{\mathbb{R}^p} \log h(y, \Sigma_G(\beta_G)) h(y, \Sigma_0) dy.$$  

By (2.2), the quasi-loglikelihood function of $\beta_G$ with the link function $G$ is $\ell_Q(\beta_G) = \sum_{i=1}^n \log \left[ h(Y_i, \Sigma_G(\beta_G)) \right]$, where

$$\log \left[ h(Y_i, \Sigma_G(\beta_G)) \right] = -\frac{p}{2} \log(2\pi) - \frac{1}{2} \sum_{j=1}^p \log \lambda_j(\Sigma_G(\beta_G)) - \frac{1}{2} Y_i' \Sigma^{-1}_G(\beta_G) Y_i.$$  

Then the quasi-maximum likelihood estimator of $\beta_G$ is $\hat{\beta}_G = \text{argmax}_{\beta_G} \ell_Q(\beta_G)$ for $G \in \{G_1, G_2\}$, and it is the empirical version of $\beta^*_G$. According to the results of Lemmas 8 and 9 in Appendix A, $\hat{\beta}_G$ is also a consistent estimator of $\beta^*_G$.

5. Numerical Studies

5.1. Simulations

To assess the finite sample performance of the proposed methods, we conduct simulation studies in two parts. Part I evaluates the asymptotic properties of $T_{LR}$. We introduce the following conditions:

**Condition 9.** For $G \in \{G_1, G_2\}$, assume that, as $p \to \infty$,

(i). $p^{-1}(\text{tr}(\bar{\Sigma}_G^r \bar{\Sigma}_G^{-1}))_{(K+1) \times (K+1)}$ converges to a finite positive definite matrix $Q_G$ in the Frobenius norm, that is,

$$\|p^{-1}(\text{tr}(\bar{\Sigma}_G^r \bar{\Sigma}_G^{-1}))_{(K+1) \times (K+1)} - Q_G\|^2_F = o(1),$$

and $\lambda_p(Q_G) > \psi_{G,0} > 0$ for a positive constant $\psi_{G,0}$;  

(ii). $p^{-1}(\text{tr}(\bar{\Sigma}_G^r \bar{\Sigma}_G^{-1})_{(K+1) \times (K+1)})$ converges to a matrix $\Delta_G$ in the Frobenius norm, and $0 < \eta_{G,1} < \psi_{G,0} > 0$, where $\eta_{G,1}$ is a finite constant and $\mu(\cdot)$ is defined in Condition 5.

**Condition 10.** For $G \in \{G_1, G_2\}$, assume that, as $p \to \infty$,

$$p^{-1}\left( \text{tr}(\bar{\Sigma}_G^r \bar{\Sigma}_G^{-1})_{(K+1) \times (K+1)} \right) \to \Lambda_G$$

in the Frobenius norm, where $\Lambda_G$ is a finite positive definite matrix.

**Condition 11.** Assume that, as $p \to \infty$, $p^{-1}\left[ 2\text{tr}(\Sigma_{G_0}^{-1} - \Sigma_{G_1}^{-1}) \Sigma_0 (\Sigma_{G_0}^{-1} - \Sigma_{G_1}^{-1}) \Sigma_0' + (\mu(\cdot) - 3)\text{tr}(\Sigma_0^{-1}(\Sigma_{G_0}^{-1} - \Sigma_{G_1}^{-1}) \Sigma_0' \Sigma_0^{-1} \Sigma_{G_0}^{-1} \Sigma_0') \right] \to \sigma_{G_1, G_2}^2$ for a finite positive constant $\sigma_{G_1, G_2}^2$.

The above conditions are applications of the law of large numbers. Based on those conditions, we have the following result.

**Theorem 6.** Assume that $K$ satisfies Condition 1 and $\Sigma_G(\beta_G)$ satisfies Conditions 3 and 4 for $G \in \{G_1, G_2\}$. Assume Conditions 1-4 and Conditions 9-11 hold. Under the null hypothesis of $H_0$, we have

$$2(np)^{-1/2} \left[ \ell_Q(G_1(\hat{\beta}_G)) - \ell_Q(G_2(\hat{\beta}_G)) \right] \to_d N(0, \sigma_{G_1, G_2}^2),$$

as $\min(n,p) \to \infty$.

**Theorem 6** indicates that, under the null hypothesis of $H_0$, $2(np)^{-1/2}T_{LR}$ asymptotically follows a normal distribution with mean zero and variance $\sigma_{G_1, G_2}^2$. In practice, $\sigma_{G_1, G_2}^2$ is unknown, and it can be estimated by the sample variance of $p^{-1/2}Y_i'\left(\Sigma_{G_0}^{-1}(\hat{\beta}_G) - \Sigma_{G_2}^{-1}(\hat{\beta}_G)\right)Y_i$. As a result, for a given significance level $\alpha$, we reject $H_0$ and choose link function $G_1$ if $2(np)^{-1/2}T_{LR}/\hat{\sigma}_{G_1, G_2} > z_{\alpha}$, where $z_{\alpha}$ stands for the $\alpha$th upper quantile of the standard normal distribution. In contrast, we reject $H_0$ and choose link function $G_2$ if $2(np)^{-1/2}T_{LR}/\hat{\sigma}_{G_1, G_2} < -z_{\alpha}$. Otherwise, $G_1$ and $G_2$ are equivalent.
model estimation and selection, while Part II examines the quasi-likelihood ratio statistic for testing the link function.

**Part I:** Consider \( p = 400 \) and \( 600 \), \( K = 10 \) and 15, and \( K_0 = \nu(s_0) = 3 \). The random variables \( Z_j \) for \( j = 1, \ldots, p \) are independent and identically generated from the three distributions: (i) the standard normal distribution \( N(0, 1) \); (ii) the mixture distribution \( 0.9N(0, 5/9) + 0.1N(0, 5) \); and (iii) the standardized exponential distribution \( \exp(1) = 1 \). The response \( Y \) is generated by \( \Sigma_{1/2} Z \). There are two types of link functions used to generate \( \Sigma_0 \), that is, the identity and exponential link functions. Under the identity link function, the resulting covariance model is \( \Sigma_0 = \beta_0^0 I_p + \beta_0^1 W_1 + \cdots + \beta_0^K W_K \), where \( \beta_0^0 = 10, \beta_j^0 = (-1)^{j-1} \) for \( j = 1, \ldots, K_0 \), and \( \beta_j^0 = 0 \) for \( j > K_0 \). Under the exponential link function, the resulting covariance model is \( \Sigma_0 = \exp(\beta_0^0 I_p + \beta_0^1 W_1 + \cdots + \beta_0^K W_K) \), where \( \beta_0^0 = 0.3, \beta_j^0 = 0.15, -0.15, -0.15 \) for \( j = 1, \ldots, K_0 \), and \( \beta_j^0 = 0 \) for any \( j > K_0 \). The weight matrices \( W_k = (w_{jk,j'k}) \) are generated as follows. The diagonal elements of the \( W_k \)'s are \( 1 \), and the off-diagonal elements of the \( W_k \)'s are \( \frac{\beta_{jk,j'k}}{\sqrt{\gamma_j^k \gamma_{j'k}^k}} \), where \( \gamma_j^k = \frac{1}{\delta_j^k} + \frac{1}{\delta_{j'k}} \) for any \( j \neq j' \).

Tables S.1–S.12. Tables 1 and 4 indicate that the difference between SD and ESD are small for any \( p \). This finding is robust across two scenarios of weight matrices, \( (a) \) and \( (b) \), and two different sizes of weight matrices, \( K = 10 \) and 15. Tables 2 and 5 indicate that the values of EE, SE and FE decrease as \( p \) gets larger. Thus, simulation results demonstrate the consistent property of the QMLE and OLS estimators. Because \( K_0 = 3 \), it is sensible to find that the values of EE, SE and FE under \( K = 10 \) are not larger than those under \( K = 15 \). In addition, Table 2 shows that the OLS estimates yield larger values of EE, SE, and FE compared to the QMLE estimates. **Table 1.** Two performance measures (SD and ESD) of parameter estimates based on the linear covariance models with the normal distribution in Part I.

| \( K \) | \( p = 400 \) | \( p = 600 \) |
|---|---|---|
| \( 10 \) | SD | 0.741 | 0.215 | 0.215 | 0.215 | 0.218 | 0.218 | 0.226 | 0.226 | 0.603 | 0.192 | 0.194 | 0.194 | 0.196 |
| | ESD | 0.756 | 0.278 | 0.274 | 0.280 | 0.272 | 0.272 | 0.601 | 0.213 | 0.211 | 0.224 | 0.201 |
| \( 15 \) | SD | 0.744 | 0.186 | 0.185 | 0.188 | 0.189 | 0.189 | 0.604 | 0.181 | 0.183 | 0.184 | 0.184 |
| | ESD | 0.754 | 0.285 | 0.284 | 0.300 | 0.299 | 0.291 | 0.605 | 0.222 | 0.221 | 0.225 | 0.212 |

NOTE: To save space, we only present the results of \( \beta_k \) for \( k = 0, \ldots, 4 \) since the results for other coefficients are similar.
with QMLE. This finding is not surprising since the QMLE is usually more efficient than the OLS estimator, although the OLS estimate is computationally efficient.

To evaluate the performance of EBIC, Tables 3 and 6 present the results of covariance matrix selections. The values of TPR and CT approach to 1 and the values of FDR decrease to 0 as \( p \) increases for both linear and exponential covariance models. These findings are consistent with the theoretical results of EBIC in Theorems 3 and 5. Since QMLE is usually more efficient than OLS, it is not surprising that the performance of EBIC evaluated at QMLE is superior to that evaluated at OLS, when \( p \) is large. In addition, the exponential covariance yields a stronger signal than the linear covariance. Hence, Table 6 shows a better performance of EBIC in comparison with those in Table 3. Finally, the values of CT indicate that EBIC performs better under \( K = 10 \) than \( K = 15 \). This is because the larger value

Table 2. Three performance measures (EE, SE, and FE) of parameter estimates based on the linear covariance models with the normal distribution in Part I.

| \( K \) | \( p \) | \( \beta_0 \) | \( \beta_1 \) | \( \beta_2 \) | \( \beta_3 \) | \( \beta_4 \) | \( \beta_0 \) | \( \beta_1 \) | \( \beta_2 \) | \( \beta_3 \) | \( \beta_4 \) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| \( K \) | \( p \) | \( \beta_0 \) | \( \beta_1 \) | \( \beta_2 \) | \( \beta_3 \) | \( \beta_4 \) | \( \beta_0 \) | \( \beta_1 \) | \( \beta_2 \) | \( \beta_3 \) | \( \beta_4 \) |
| 10 | 400 | 1.402 | 5.462 | 4.477 | | | | | | | |
| | | (0.966) | (2.162) | (2.219) | | | | | | | |
| | 600 | 0.918 | 4.173 | 2.719 | | | | | | | |
| | | (0.659) | (1.438) | (1.297) | | | | | | | |
| 15 | 400 | 1.881 | 6.619 | 6.663 | | | | | | | |
| | | (1.159) | (2.399) | (2.991) | | | | | | | |
| | 600 | 1.142 | 5.236 | 3.918 | | | | | | | |
| | | (0.656) | (2.028) | (1.655) | | | | | | | |

Table 3. Three performance measures (TPR, FDR, and CT) of EBIC (\( \gamma = 0.5 \)) based on the linear covariance models with the normal distribution in Part I.

| \( K \) | \( p \) | \( \beta_0 \) | \( \beta_1 \) | \( \beta_2 \) | \( \beta_3 \) | \( \beta_4 \) | \( \beta_0 \) | \( \beta_1 \) | \( \beta_2 \) | \( \beta_3 \) | \( \beta_4 \) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| \( K \) | \( p \) | \( \beta_0 \) | \( \beta_1 \) | \( \beta_2 \) | \( \beta_3 \) | \( \beta_4 \) | \( \beta_0 \) | \( \beta_1 \) | \( \beta_2 \) | \( \beta_3 \) | \( \beta_4 \) |
| 10 | 400 | 0.830 | 0.904 | 0.960 | 0.976 | | | | | | |
| | | (0.199) | (0.143) | (0.102) | (0.074) | | | | | | |
| | 600 | 0.912 | 0.941 | 0.923 | 0.976 | | | | | | |
| | | (0.054) | (0.120) | (0.035) | (0.096) | | | | | | |
| 15 | 400 | 0.807 | 0.904 | 0.959 | 0.975 | | | | | | |
| | | (0.204) | (0.139) | (0.101) | (0.075) | | | | | | |
| | 600 | 0.942 | 0.909 | 0.959 | 0.975 | | | | | | |
| | | (0.055) | (0.113) | (0.035) | (0.094) | | | | | | |

Table 4. Two performance measures (SD and ESD) of parameter estimates based on the exponential covariance models with the normal distribution in Part I.

| \( K \) | \( p \) | \( \beta_0 \) | \( \beta_1 \) | \( \beta_2 \) | \( \beta_3 \) | \( \beta_4 \) | \( \beta_0 \) | \( \beta_1 \) | \( \beta_2 \) | \( \beta_3 \) | \( \beta_4 \) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| \( K \) | \( p \) | \( \beta_0 \) | \( \beta_1 \) | \( \beta_2 \) | \( \beta_3 \) | \( \beta_4 \) | \( \beta_0 \) | \( \beta_1 \) | \( \beta_2 \) | \( \beta_3 \) | \( \beta_4 \) |
| 10 | SD | 0.070 | 0.031 | 0.031 | 0.031 | 0.032 | 0.058 | 0.025 | 0.025 | 0.025 | 0.025 | 0.025 |
| | ESD | 0.071 | 0.032 | 0.032 | 0.032 | 0.032 | 0.056 | 0.025 | 0.026 | 0.026 | 0.026 | 0.026 |
| 15 | SD | 0.071 | 0.031 | 0.031 | 0.031 | 0.031 | 0.032 | 0.058 | 0.025 | 0.025 | 0.025 | 0.025 | 0.025 |
| | ESD | 0.073 | 0.034 | 0.031 | 0.031 | 0.031 | 0.033 | 0.061 | 0.025 | 0.028 | 0.027 | 0.027 | 0.027 |

NOTE: The mean and standard error of the three measures are calculated based on 500 realizations.

NOTE: The mean and standard error of the TPR and FDR values are calculated from 500 realizations.

NOTE: To save space, we only present the results of \( \beta_k \) for \( k = 0, \ldots, 4 \) since the results for other coefficients are similar.
Three performance measures (EE, SE, and FE) of parameter estimates based on the exponential covariance models with the normal distribution in Part I.

| p  | K = 10  | K = 15  |
|----|---------|---------|
|    | EE      | SE      | FE    | EE      | SE      | FE    |
| 400| 0.016   | 1.225   | 0.164 | 0.023   | 1.512   | 0.244 |
|    | (0.011) | (0.422) | (0.089)| (0.012) | (0.473) | (0.108)|
| 600| 0.010   | 0.993   | 0.106 | 0.015   | 1.240   | 0.162 |
|    | (0.006) | (0.323) | (0.051)| (0.007) | (0.381) | (0.068)|

Table 6. Three performance measures (TPR, FDR, and CT) of EBIC (β = 0.5) based on the exponential covariance models with the normal distribution in Part I.

| p  | K = 10  | K = 15  |
|----|---------|---------|
|    | Scenario (a) | Scenario (b) |
|    | TPR      | FDR     | CT      | TPR      | FDR     | CT      |
| 400| 0.978    | 0.006   | 0.884   | 0.979    | 0.006   | 0.892   |
|    | (0.074) | (0.035) | (0.062) | (0.071) | (0.035) | (0.062) |
| 600| 0.977    | 0.004   | 0.884   | 0.979    | 0.004   | 0.892   |
|    | (0.082) | (0.034) | (0.062) | (0.072) | (0.034) | (0.062) |

Note: The mean and standard error of the TPR and FDR values are calculated from 500 realizations.

Table 7. The empirical performance of the quasi-likelihood ratio test for Part II when the random variables $Z_i (i = 1, \ldots, p)$ defined in Section 5.1 are generated from a standard normal distribution; the values reported in the table are the percentages of nonrejection/rejection, respectively, calculated from 500 realizations.

| Scenario | n  | p  | B vs exp(B) | $B^2$ vs exp(B) | $B^{-1}$ vs exp(B) |
|----------|----|----|-------------|-----------------|-------------------|
| (a)      | 25 | 100| 19.6/80.4   | 72.6/27.4       | 17.4/82.6         |
|          | 300| 0/100| 26.4/73.6 | 0.2/99.8       | 0/100             |
|          | 75 | 100| 0/2/99.9   | 31.2/68.8       | 0/100             |
|          | 300| 0/100| 0.8/99.2  | 30.2/69.8       | 0/100             |
| (b)      | 25 | 100| 33.6/66.4  | 81.2/18.8       | 0/100             |
|          | 300| 0/100| 31.8/68.2 | 0/100           | 1/99              |
|          | 75 | 100| 0/100      | 36.2/63.8       | 0/100             |
|          | 300| 0.6/99.4| 0/100      |                |                   |

of $K$ leads to more overfitting, so the estimates under $K = 10$ are more reliable than those under $K = 15$.

Part II: This part evaluates the performance of the proposed quasi-likelihood ratio statistic for assessing the adequacy of the link function. We consider $p = 100$ and 300, $n = 25$ and 75, $K = 15$, and $K_0 = 3$. The responses $Y_i$, for $i = 1, \ldots, n$, and the weight matrices $W_k$, for $k = 1, \ldots, K$, are generated from the same process as those in Part I. In addition, let the matrix $B = \beta_0 + \beta_1 W_1 + \cdots + \beta_K W_K$, where $\beta_0^0 = 0.3$, $\beta_1^0 = 0.15$, $-0.15$, $-0.15$ for $j = 1, \ldots, K_0$, and $\beta_j^0 = 0$ for any $j > K_0$. In this study, we assume that the true link function is exponential. Then, three pairs of link functions are compared as follows: $G_1(B) = B$ versus $G_2(B) = \exp(B)$, $G_1(B) = B^2$ versus $G_2(B) = \exp(B)$, and $G_1(B) = B^{-1}$ versus $G_2(B) = \exp(B)$. For each simulation setting, there are 200 replications. The simulation results for the normal distribution are presented in Table 7.

According to Table 7, we find that the percentage of the correct link function increases as either $n$ or $p$ gets larger. This finding is as expected and supports our theoretical result in Theorem 6. It is of interest to note that, as both $n$ and $p$ are small, our proposed test is more likely to differentiate between the exponential link and the inverse (or linear) link than between the exponential link and the quadratic link. This finding is sensible since the exponential function can be approximated by a summation of the polynomial functions of $B$ via the Taylor series expansion. Finally, the simulation results for testing the true exponential link function under the mixture normal and the standardized exponential distributions show similar findings quantitatively; see Tables S.13 and S.14 in the supplemental materials.

5.2. Real Data Analysis

The mean-variance portfolio theory of Markowitz (1952) plays a fundamental role in modern finance theory. The major assumption made in the theory is that investment decisions are solely based on the mean and covariance of the investment returns. Accordingly, the optimal portfolio can be constructed by generating a maximum return based on a given level of risk or by minimizing risk for a given level of expected return. Among various portfolio optimization models, we consider the minimum variance approach since accurately estimating mean returns can be difficult in practice (see, e.g., Jagannathan and Ma 2003; DeMiguel et al. 2009). To construct the minimum variance portfolios, we employ our proposed methods to accurately estimate covariance matrices.
literature (see, e.g., Chan, Karceski, and Lakonishok 1998; Hou, the weight matrices, and they are commonly used in the monthly returns, and there are 31 months in total. Table 8 Xue, and Zhang 2020). Denote the vector of stock returns and where the data were downloaded from Wharton Research Data Services. The response variable consists of stocks in the U.S. stock market from January 2016 to July 2018, and the Sharpe ratio (SR) of the excess return of the investment portfolio over the risk-free rate by adjusting SD, where the risk-free rate is proxied by the return of 1-month treasury rate.

Based on the 22 weighted matrices and 30 samples, we fit the data with the linear covariance model and obtain both QMLE and the OLS estimates. In addition, we fit the exponential covariance model and compute its QMLE. To assess the adequacy of the identity link function versus the exponential link function, the quasi-likelihood ratio test is applied. The $p$-value of the test statistics is 0.009, which indicates that the exponential link function is better than linear function. Furthermore, we employ the model selection criterion EBIC with $\gamma = 0.5$ to select most relevant variables. In sum, we have fit the data with four different models. Table 7 presents four models and their corresponding sample mean (Mean), standard deviation (SD), and Sharpe ratio (SR). Both measures of SD and SR indicate that the exponential covariance submodel performs the best. Although the linear covariance model with QMLE has the largest Mean, it has a high SD and a low SR. The results are consistent with our quasi-likelihood ratio test, which indicates that the exponential link function is better than the linear function. It is also worth noting that the Sharpe ratio of the exponential covariance submodel is 32.2% higher than that of the full exponential covariance model. Consequently, the above results demonstrate the usefulness of the generalized covariance model along with its estimation, testing, and selection.

### 6. Concluding Remarks

In this article, we propose a unified framework to study the structured covariance matrix. Specifically, we introduce the general link function to connect the covariance matrix of responses to the linear combination of weight matrices. The quasi-maximum likelihood estimator and the ordinary least squares estimator are obtained, as well as their corresponding asymptotic properties, without imposing any specific distribution on the response variable and with allowing the number of weight matrices to diverge. An extended Bayesian information criteria (EBIC) for weight matrix selection is proposed and its consistent property is established. To assess the adequacy of the link function, we further consider the quasi-likelihood ratio test and obtain its limiting distribution.

To broaden the applications of generalized covariance models, we identify four avenues for future research. The first is extending our approach to study covariance matrices for matrix-variate regression models (Ding and Cook 2018), reduced rank

| Abbreviation | Description |
|--------------|-------------|
| LAS | The logarithm of the value of total assets |
| LCF | The logarithm of the cash flow value |
| SIZE | The logarithm of the market value |
| bm | Book value/market value |
| pe-op | Price/operating earnings |
| ps | Price/gross sales |
| pcf | Price/cash flow |
| npm | Net profit margin |
| opmad | Operating profit margin after depreciation |
| gpm | Gross profit margin |
| ptm | Pretax profit margin |
| cf | Cash flow margin |
| roa | Return on assets, net profit/total assets |
| gprof | Gross profit/total assets |
| cash-lt | Cash balance/total liabilities |
| debt-ebitda | Total debt/earnings before interest |
| cash-debt | Cash flow/total debt |
| it-ppent | Total liabilities/total tangible assets |
| at-turn | Asset turnover |
| rect-turn | Receivables turnover |
| pay-turn | Payable turnover |
| ptb | Price/book value |

Table 9. The monthly mean, standard deviation (SD) and Sharpe ratio (SR) of the portfolio returns.

| Description | Mean | SD | SR |
|-------------|------|----|----|
| Linear covariance model (QMLE) | 0.014 | 0.049 | 0.270 |
| Linear covariance model (OLS) | 0.011 | 0.040 | 0.209 |
| Exponential covariance model (QMLE) | 0.012 | 0.030 | 0.366 |
| Exponential covariance submodel (QMLE) | 0.012 | 0.024 | 0.484 |

In this example, we collect the monthly stock returns of 400 stocks in the U.S. stock market from January 2016 to July 2018, where the data were downloaded from Wharton Research Data Services. The response variable consists of $p = 400$ stocks’ monthly returns, and there are 31 months in total. Table 8 presents 22 covariates as auxiliary information for constructing the weight matrices, and they are commonly used in the literature (see, e.g., Chan, Karceski, and Lakonishok 1998; Hou, Xue, and Zhang 2020). Denote the vector of stock returns and the auxiliary covariates at month $i$ by $Y_i \in \mathbb{R}^p$ and $X_{k(i)}^i \in \mathbb{R}^p$, respectively, for $i = 1, \ldots, 31$ and $k = 1, \ldots, 22$, where the $X_{k(i)}^i$s are evaluated at month $i - 1$. Hence, the auxiliary covariates are all observed one month prior to $Y_i$, and it is reasonable to treat the auxiliary covariates as fixed. We next construct the weight matrices $W_k^{i(i)}$. Let the off-diagonal elements of $W_k^{i(i)}$ at month $i$ be $\exp(-10(X_{k(j_1)}^i - X_{k(j_2)}^i)^2)$ if $|X_{k(j_1)}^i - X_{k(j_2)}^i| < \tau(k)$ for some threshold value $\tau(k)$ and 0 otherwise, where $1 \leq j_1 \neq j_2 \leq p$ and $\tau(k)$ is selected such that the density of $W_k^{i(i)}$ (the ratio of nonzero elements in $W_k^{i(i)}$) is 10%. In addition, set the diagonal elements to be zeros. We then fit each of the first 30 months data ($i = 1, \ldots, 30$) by the generalized covariance model (2.1), respectively, with the identity link function and the exponential link function. As a result, we obtain QMLE and the OLS estimates for the linear covariance model and QMLE for the exponential covariance model. For the sake of simplicity, their resulting covariance estimates are denoted by $\hat{\Sigma}^{i(i)}$.

Based on the estimated covariance matrices $\hat{\Sigma}^{i(i)} (i = 1, \ldots, 30)$, we follow the DeMiguel et al.’s (2009) approach and obtain the minimum variance portfolio weights by minimizing the variance of the portfolio, $\{w^{i(i)}_1 \Sigma^{i(i)} w^{i(i)}_1\}$, such that $\{w^{i(i)}_1 \Sigma^{i(i)} w^{i(i)}_1\}^T 1 = 1$, where the portfolio weight $w^{i(i)}_1 \in \mathbb{R}^p$ and $1 = (1, \ldots, 1)^T \in \mathbb{R}^p$. After algebraic simplification, the weight of the minimum variance portfolio is $\hat{\gamma}^{i(i)} = \hat{\Sigma}^{i(i)} - 1/1^T \hat{\Sigma}^{i(i)} - 1$. Subsequently, we compute the out-of-sample portfolio return at month $i + 1$ based on the optimal portfolio weight $\hat{\gamma}^{i(i)}$ that is $r_i = \{\hat{\gamma}^{i(i)}\}^T Y_{i+1}$ for $i = 1, \ldots, 30$. To assess the out-of-sample performance of the portfolio return, we consider three performance measures, namely, the sample mean (Mean) of $r_i$, the sample standard deviation (SD) of $r_i$ and the Sharpe ratio (SR) of $r_i$. Note that the Sharpe ratio is the excess return of the investment portfolio over the risk-free rate by adjusting SD, where the risk-free rate is proxied by the return of 1-month treasury rate.
regression model (Izenman 1975), and bilinear regression models (von Rosen 2018). The second is conducting covariance matrix analysis for multivariate discrete responses (Liang, Zeger, and Qaqish 1992). The third is extending our model to accommodate random weight matrices. The last is combining our model with the factor model (Bai and Ng 2002; Cai, Han, and Pan 2020) to study high dimensional covariance matrix with divergent eigenvalues. We believe these efforts would further increase the usefulness of generalized covariance models.

Appendix

This Appendix includes three components (Appendices A–C) to show Theorems 1, 2, and 6, respectively. To save space, 12 lemmas used for proving the theorems are relegated to the supplementary materials. In addition, the proofs of Theorems 3–5 are also presented in the supplementary materials.

Appendix A: Proof of Theorem 1

To prove this theorem, we consider the following two steps. The first step demonstrates that \( \hat{\beta}_Q \) is \( \sqrt{p}/K \)-consistent, while the second step shows the asymptotic normality of \( \hat{\beta}_Q \).

Step I. To complete this step, it suffices to follow the approach of Fan and Peng (2004) to show that, for any given \( \varepsilon > 0 \), there is a large constant \( C \) such that

\[
\Pr \left\{ \sup_{|u| = C} \epsilon_Q(\beta_0^0 + \alpha_p u) < \epsilon_Q(\beta_0) \right\} \geq 1 - \varepsilon
\]

where \( p \) is sufficiently large, where \( \alpha_p = \sqrt{K/p} \). This implies that, with probability tending to 1, there is a localizer \( \hat{\beta}_Q \) in the ball \( \| \beta_0 + \alpha_p u \| \leq C \) such that \( \| \hat{\beta}_Q - \beta_0 \| < C \alpha_p \). By the Taylor series expansion, we have

\[
\epsilon_Q(\beta_0^0 + \alpha_p u) - \epsilon_Q(\beta_0) = \alpha_p^T \frac{\partial^2 \epsilon_Q(\beta_0)}{\partial \beta} u + \frac{1}{2} \alpha_p^T \left\{ \frac{\partial^2 \epsilon_Q(\beta_0)}{\partial \beta \partial \beta} \right\} u + R_p(u)
\]

where

\[
R_p(u) = \frac{1}{6} \alpha_p^2 \sum_{j=0}^{K} \sum_{k=0}^{K} \sum_{l=0}^{K} \sum_{m=0}^{K} \frac{3}{2} \epsilon_Q(\beta^*) - \frac{\partial^2 \epsilon_Q(\beta^*)}{\partial \beta_j \partial \beta_k} \beta_j \beta_k
\]

and \( \beta^* \) lies between \( \beta_0 \) and \( \beta_0^0 + \alpha_p u \).

Applying similar techniques to those used in the proof of Lemma 4 of the supplementary materials, we have

\[
M_1 = (pK)^{1/2} \alpha_p C \left\{ (pK)^{-1/2} \frac{\epsilon_Q(\beta_0^0)}{\epsilon_Q(\beta_0)} \right\} = \frac{1}{2} K CO_p(1)
\]

In addition, it can be shown that

\[
M_2 = -\frac{1}{4} \sqrt{p} \alpha_p^2 u^T \left( \frac{1}{4} \epsilon_Q^2 + \frac{1}{2} \frac{\partial^2 \epsilon_Q(\beta_0)}{\partial \beta \partial \beta} \right) u \leq -\frac{1}{4} \sqrt{p} \alpha_p^2 \epsilon_Q^2 + \epsilon_Q(\beta_0) \leq -\frac{1}{4} K \epsilon_Q^2 + \epsilon_Q(\beta_0)
\]

where \( \varphi_0 > 0 \). Moreover, we have

\[
R_p(u) \leq \frac{1}{6} \sqrt{p} \alpha_p^2 \| u \|^3 \left\{ \sum_{j=0}^{K} \sum_{l=0}^{K} \sum_{m=0}^{K} \frac{3}{2} \epsilon_Q(\beta^*) - \frac{\partial^2 \epsilon_Q(\beta^*)}{\partial \beta_j \partial \beta_k} \beta_j \beta_k \right\}.
\]

By Conditions 3–4, it can be shown that \( \frac{1}{6} \frac{\partial^2 \epsilon_Q(\beta^*)}{\partial \beta \partial \beta} \) is bounded uniformly. Accordingly, by Condition 1, \( R_p(u) \leq O_p(p^{-1/2} K^2 C^3) = o_p(K) \).

Combining the above results, we have \( K^{-1/2} \{ M_1 + M_2 + R_p(u) \} \leq \frac{1}{4} C O_p(1) - \frac{1}{4} C \varphi_0 + o_p(K) \), which is a quadratic function of \( C \). Hence, as long as \( C \) is sufficiently large, we have

\[
\sup_{|u| = C} \epsilon_Q(\beta_0^0 + \alpha_p u) - \epsilon_Q(\beta_0) \leq \sup_{|u| = C} K \left\{ \frac{1}{2} C O_p(1) - \frac{1}{4} C \varphi_0 + o_p(1) \right\} < 0,
\]

with probability tending to 1, which completes the proof of Step I.

Step II. Using the result of Step I and applying the Taylor series expansion, we have

\[
\frac{\partial \epsilon_Q(\hat{\beta}_Q)}{\partial \beta} = \frac{\partial \epsilon_Q(\beta_0)}{\partial \beta} + \frac{\partial^2 \epsilon_Q(\beta_0)}{\partial \beta \partial \beta} \beta_0 \beta_0 + R_p(\beta_0)
\]

where

\[
R_p = \frac{1}{2} \left\{ I_{K+1} \otimes (\hat{\beta}_Q - \beta_0)^T \right\} \frac{1}{\beta_0} \text{vec} \left( \frac{\partial^2 \epsilon_Q(\beta_0)}{\partial \beta \partial \beta} \right) \beta_0 - \beta_0
\]

and \( \beta^* \) lies between \( \beta_0 \) and \( \hat{\beta}_Q \). After algebraic simplification, we obtain

\[
\| \hat{\beta}_Q - \beta_0 \| \leq \sqrt{\frac{K^2 \epsilon_Q^2}{p \alpha_p^2} + \frac{1}{2} \frac{\partial^2 \epsilon_Q(\beta_0)}{\partial \beta \partial \beta} \| \beta_0 \| + R_p(\beta_0)}
\]

\[
= \sqrt{\frac{K^2 \epsilon_Q^2}{p \alpha_p^2} + \frac{1}{2} \frac{\partial^2 \epsilon_Q(\beta_0)}{\partial \beta \partial \beta} \| \beta_0 \| + R_p(\beta_0)}
\]

\[
= \frac{1}{\sqrt{p}} \epsilon_Q(\beta_0) + o_p(1)
\]

and

\[
\| \hat{\beta}_Q - \beta_0 \| \leq \sqrt{\frac{K^2 \epsilon_Q^2}{p \alpha_p^2} + \frac{1}{2} \frac{\partial^2 \epsilon_Q(\beta_0)}{\partial \beta \partial \beta} \| \beta_0 \| + R_p(\beta_0)}
\]

\[
= \frac{1}{\sqrt{p}} \epsilon_Q(\beta_0) + o_p(1)
\]

Accordingly, we have

\[
-(p/K)^{-1/2} \frac{1}{p} \frac{\partial^2 \epsilon_Q(\beta_0)}{\partial \beta \partial \beta} \beta_0 \beta_0 \leq \frac{1}{2} \frac{\partial^2 \epsilon_Q(\beta_0)}{\partial \beta \partial \beta} \| \beta_0 \| + o_p(1)
\]

\[
= \frac{1}{2} \frac{\partial^2 \epsilon_Q(\beta_0)}{\partial \beta \partial \beta} \| \beta_0 \| + o_p(1)
\]

By (A.1), we further have

\[
-(p/K)^{-1/2} \frac{1}{p} \frac{\partial^2 \epsilon_Q(\beta_0)}{\partial \beta \partial \beta} \beta_0 \beta_0 \leq \frac{1}{\sqrt{p}} \frac{\partial \epsilon_Q(\beta_0)}{\partial \beta} - \hat{R}_p \sqrt{\frac{p}{K}}
\]

Note that, by Condition 1, \( \hat{R}_p(\sqrt{p}) = o_p(K^2/p^{1/2}) = o_p(1) \). This, together with the above results and (A.2), implies that

\[
\frac{1}{2} \frac{\partial^2 \epsilon_Q(\beta_0)}{\partial \beta \partial \beta} \| \beta_0 \| + o_p(1)
\]

Finally, by Lemma 4 again, we complete the proof of Step II.
Appendix B: Proof of Theorem 2

By the Taylor series expansion, we have that \( \Sigma(\hat{\beta}_Q) - \Sigma_0 = \sum_{k=0}^{K} \frac{1}{2} \Sigma(\hat{\beta}_Q - \beta_0)^T \Sigma(\hat{\beta}_Q) \hat{\beta}_Q - \beta_0 \), where \( \hat{\beta}_Q \) lies between \( \beta_0 \) and \( \hat{\beta}_Q \). In addition, employing Condition 3, we obtain \( \sup_{k, \beta \in U_3} \frac{1}{2} \Sigma(\hat{\beta}_Q)^T \Sigma(\hat{\beta}_Q) \leq \sup_{k, \beta \in U_3} \frac{1}{2} \Sigma(\hat{\beta}_Q)^T \Sigma(\hat{\beta}_Q) \leq \infty \). The above results, together with Theorem 1, imply that \( \| \Sigma(\hat{\beta}_Q) - \Sigma_0 \|_2 \leq \sup_{k, \beta \in U_3} \frac{1}{2} \Sigma(\hat{\beta}_Q)^T \Sigma(\hat{\beta}_Q) \| \hat{\beta}_Q - \beta_0 \|_1 \leq \tau_0 \sqrt{K+1} \| \hat{\beta}_Q - \beta_0 \|_2 = O_p(K^{-1/2}). \)

We next study the asymptotic property of \( \Sigma^{-1}(\hat{\beta}_Q) \). Applying the Taylor series expansion, we obtain that
\[
\Sigma^{-1}(\hat{\beta}_Q) = \Sigma_0^{-1} - \sum_{k=0}^{K} \hat{\beta}_Q - \beta_0 \Sigma^{-1}(\hat{\beta}_Q) \Sigma^{-1}(\hat{\beta}_Q),
\]
where \( \hat{\beta} \) lies between \( \hat{\beta}_Q \) and \( \beta \). By Conditions 3 and 4, there exists a constant \( \epsilon_{\max} > 0 \) such that, for all \( k = 0, \ldots, K, \| \Sigma^{-1}(\hat{\beta}_Q)^{K} \beta \Sigma^{-1}(\hat{\beta}_Q) \|_2 \leq \epsilon_{\max} \). The above results, in conjunction with Theorem 1, lead to
\[
\| \Sigma^{-1}(\hat{\beta}_Q) - \Sigma_0^{-1} \|_2 \leq \epsilon_{\max} \| \hat{\beta}_Q - \beta_0 \|_1 \leq \epsilon_{\max} \sqrt{K+1} \| \hat{\beta}_Q - \beta_0 \|_2 = O_p(K^{-1/2}),
\]
which completes the entire proof.

Appendix C: Proof of Theorem 6

Employing the Taylor series expansion, we have that, for any \( G \in \{G_1, G_2\} \),
\[
\ell_{QG}(\hat{\beta}_G) = \ell_{QG}(\beta_G) + (\hat{\beta}_G - \beta_G)^T \frac{\partial \ell_{QG}(\beta_G^*)}{\partial \beta_G} + \frac{1}{2} (\hat{\beta}_G - \beta_G)^T \frac{\partial^2 \ell_{QG}(\beta_G^*)}{\partial \beta_G^T \partial \beta_G} (\hat{\beta}_G - \beta_G) + O_G,
\]
where \( O_G \) satisfies
\[
O_G = \frac{1}{2} \sum_{j=0}^{K} \sum_{k=0}^{K} (\hat{\beta}_G - \beta_G)^T (\beta_G - \beta_G) \frac{\partial^2 \ell_{QG}(\hat{\beta}_G)}{\partial \beta_G^T \partial \beta_G} (\hat{\beta}_G - \beta_G).
\]
and \( \hat{\beta}_G \) lies between \( \beta_G^* \) and \( \hat{\beta}_G \). Applying similar techniques to those used in the proof of Theorem 1, we can verify that \( O_G = o_p(K) \). In addition, by Lemma 9, \( \| \hat{\beta}_G - \beta_G^* \|_2 = O_p(K^{1/2}) \). The above results, in conjunction with Lemmas 8 and 9, lead to
\[
(np)^{-1/2} \left( \ell_{QG}(\hat{\beta}_G) - \ell_{QG}(\beta_G^*) \right) \leq K^{1/2} \| \hat{\beta}_G - \beta_G^* \|_2 (np)^{-1/2} \frac{\partial \ell_{QG}(\beta_G^*)}{\partial \beta_G} \| + (np)^{1/2} \| \hat{\beta}_G - \beta_G^* \|_2 (np)^{-1/2} \left( \ell_{QG}(\hat{\beta}_G) - \ell_{QG}(\beta_G^*) \right) = O_p \left( K^2 (np)^{-1/2} \right) \to 0 \text{ by Condition 1.}
\]
Based on the above results, we obtain that
\[
(np)^{-1/2} \left( \ell_{QG}(\beta_G) - \ell_{QG}(\hat{\beta}_G) \right) = (np)^{-1/2} \left( \ell_{QG}(\beta_G^*) - \ell_{QG}(\hat{\beta}_G) \right) + o_p(1).
\]
By the definition of quasi-loglikelihood function, we then have
\[
(np)^{-1/2} \left( \ell_{QG}(\beta_G) - \ell_{QG}(\hat{\beta}_G) \right) - (np)^{-1/2} E \left( \ell_{QG}(\beta_G) - \ell_{QG}(\hat{\beta}_G) \right)
\]
which completes the proof.

Supplementary Materials

The supplementary material consists of nine sections. Section S.1 presents a detailed illustration of G and its related derivatives. Section S.2 introduces an example that satisfies Conditions 2–5. Section S.3 presents twelve useful lemmas; Sections S.4–S.6 demonstrate Theorems 3–5, respectively; Section S.7 extends our model to accommodate \( n \geq 1 \); Section S.8 introduces the proofs of Theorems 7 and 8 that are proposed in Section S.7; Section S.9 presents the simulation results for the mixture normal and the standardized exponential distributions.

Acknowledgments

The authors are grateful to the editor, associate editor, and anonymous referees for their insightful comments and constructive suggestions.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

Xinyan Fan’s research was supported by the National Natural Science Foundation of China (NSFC, 12201626), and the Public Computing Cloud, Renmin University of China. Wei Lan’s research was supported by the National Natural Science Foundation of China (NSFC, 71991472, 12171395, 11931014), the Joint Lab of Data Science and Business Intelligence at Southwest University of Finance and Economics, and the Fundamental Research Funds for the Central Universities (JBK1800602). Tao Zou’s research was supported by ANU College of Business and Economics Early Career Researcher Grant, the RSFAS Cross Disciplinary Grant.

References

Adamczak, R. (2015), “A Note on the Hanson-Wright Inequality for Random Vectors with Dependencies,” Electronic Communications in Probability, 20, 1–13. [38]
Anderson, T. W. (1973), “Asymptotically Efficient Estimation of Covariance Matrices with Linear Structure,” The Annals of Statistics, 1, 135–141. [36]
Bai, J., and Ng. S. (2002), “Determining the Number of Factors in Approximate Factor Models,” Econometrica, 70, 191–221. [46]
Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015), Hierarchical Modeling and Analysis of Spatial Data (2nd ed.), Boca Raton, FL: Chapman and Hall/CRC. [36]
