Statistica Sinica **25** (2015), 1249-1263
doi:http://dx.doi.org/10.5705/ss.2013.129

LOCAL LINEAR ESTIMATION OF COVARIANCE MATRICES VIA CHOLESKY DECOMPOSITION

Ziqi Chen and Chenlei Leng

Central South University and University of Warwick

Abstract: An important problem in multivariate statistics is the estimation of covariance matrices. We consider a class of nonparametric covariance models in which the entries in the covariance matrix depend on covariates. Previously, the locally constant approach was used for estimating this matrix due to its simplicity. However, to ensure the positive definiteness of the resulting estimator, a single bandwidth parameter was used for estimating all the elements in this matrix. We propose to use the locally linear method, a technique known to outperform local constant estimation, for estimating the elements after the modified Cholesky decomposition. The proposed estimator is guaranteed to be positive definite, allows different degrees of smoothing for different elements, possesses good theoretical properties, and performs well in numerical studies. An application to the Boston housing data is provided to illustrate the finite-sample performance of the proposed method.

Key words and phrases: Covariance matrix, local constant estimator, local linear estimator, modified Cholesky decomposition.

1. Introduction

Estimation of covariance matrices is a problem of fundamental importance in multivariate statistics. A great deal of research has been done in developing models and approaches for estimation when the structure is complicated (Pourahmadi (1999); Fan, Huang, and Li (2007)) and dimensionality is large (Bickel and Levina (2008); Levina, Rothman, and Zhu (2008)). A major difficulty in modelling covariances, as opposed to the mean, is that the resulting estimates must be positive definite. This requirement puts severe restriction on the feasibility of developing flexible models and any associated estimation method.

A usual assumption is that covariances between the variables are constant. This is seldom true. Consider, for example, a genetic network that is used for studying genes interactions. It is conceivable that correlations between these genes depend on such biomedical factors as blood pressure, hormone level and other important biomedical factors, possibly as functions of time (Kolar et al. (2010)). In finance, covariances between different assets are constantly changing in response to policies and markets, or simply time (Engle (2002)). In this

paper, we consider a flexible nonparametric covariance model that allows entries in the covariance matrix to depend on covariates in a data-adaptive fashion. Fan, Huang, and Li (2007) studied a model that only allows marginal variances to be covariate-dependent. Yin et al. (2010) proposed the Nadaraya-Watson kernel estimator for this model.

In particular, Yin et al. (2010) employed a locally constant argument in forming an estimator. To ensure that the local constant estimator is positive definite, a single bandwidth is used for estimating all the components. In practice, different components of the nonparametric covariance matrix may have different degrees of smoothness. Moreover, locally linear estimators are superior compared to the locally constant ones in terms of smaller biases, higher statistical efficiency in an asymptotic minimax sense, and better boundary properties (Fan and Gijbels (1996)). Though it is much more difficult to develop local linear estimators for covariance matrices. As pointed out by Yin et al. (2010), it is a major challenge to develop local linear estimators for these matrices that are positive definite.

We apply local linear kernel regression to each component after the modified Cholesky decomposition of the nonparametric covariance matrix (Pourahmadi (1999); Leng, Zhang, and Pan (2010)) and obtain the so-called local linear estimator of such a matrix. We show that the proposed estimator is guaranteed to be positive definite, allows different degrees of smoothness for different components, possesses good theoretical properties and performs well in numerical studies. In the longitudinal data context where variables are naturally ordered, Wu and Pourahmadi (2003) applied locally constant estimation to smooth rows or columns in this decomposition. For the spectral estimator of a multivariate stationary time series, Dai and Guo (2004) proposed to smooth the Cholesky decomposition of an initial estimate of the multivariate spectrum, and Rosen and Stoffer (2007) proposed a Bayesian approach to estimate the components of modified Cholesky decomposition of the inverse of the spectral matrix. We organize our paper as follows. In Section 2, based on the modified Cholesky decomposition, we propose a local linear estimator of the conditional covariance matrix. The asymptotic properties of the estimators are given in Section 3. In Section 4, we report on simulation studies conducted to evaluate the performances of the proposed method. The proposed approach is further illustrated with a dataset. A brief discussion is presented in Section 5. Two lemmas and the detailed proofs of Theorems 1, 2, and 3 are in the Supplementary Material.

2. Local Linear Estimation

Let $Y=(y_1,\ldots,y_p)^T$ be a p-dimensional random vector and $U=(u_1,\ldots,u_q)^T$ be the associated index random vector. We model the conditional mean and the conditional covariance of Y given U as $m(U)=(m_1(U),\ldots,m_p(U))^T$ and $\Sigma(U)$,

respectively. Suppose (Y_i, U_i) with $Y_i = (y_{i1}, \ldots, y_{ip})^T$ is a random sample from the population (Y, U) with $Y|U \sim N(m(U), \Sigma(U))$, for $i = 1, \ldots, n$. We are interested in the estimation of the conditional covariance matrix $\Sigma(u)$. In this paper, we only consider $q = 1$.

We assume that the mean function $m(u)$ is estimated by $\hat{m}(u)$ and discuss how to estimate it at the end of this section. Given $U = u$, Yin et al. (2010) proposed to estimate $\Sigma(u)$ by minimizing

$$\sum_{i=1}^{n} K_h(U_i - u)\Big[\{Y_i - m(u)\}^T \Sigma^{-1}(u)\{Y_i - m(u)\} + \log(|\Sigma(u)|)\Big], \tag{2.1}$$

where $K_h(u) = h^{-1}K(u/h)$ with $K(\cdot)$ being a kernel function and h any appropriate bandwidth. The resulting estimator, known as the Nadaraya-Watson kernel estimator or the local constant estimator, of the conditional covariance matrix is

$$\hat{\Sigma}_{LC}(u) = \Big[\sum_{i=1}^{n} K_h(U_i - u)\{Y_i - m(U_i)\}\{Y_i - m(U_i)\}^T\Big]\Big\{\sum_{i=1}^{n} K_h(U_i - u)\Big\}^{-1}.$$

To ensure the positive definiteness of the local constant estimator, the same bandwidth h is used for smoothing all the entries (Claeskens and Aerts (2000)). However, different components of $\Sigma(u)$ may have different degrees of smoothness. We seek locally linear estimators as more desirable than the local constant estimators in terms of smaller biases, higher statistical efficiency in an asymptotical minimax sense and better boundary properties (Fan (1993); Fan and Gijbels (1996)). A natural question arises on how to develop such estimators for covariance matrices that are positive definite.

To guarantee the positive definiteness of the estimated conditional covariance matrix, we make use of the modified Cholesky decomposition by decomposing $\Sigma(u)$ as $P(u)\Sigma(u)P(u)^T = D(u)$, where $P(u)$ is a lower unitriangular matrix with the (j, l)-th below diagonal entry being $-\phi_{jl}(u)$ and $D(u) = \mathrm{diag}(\sigma_1^2(u), \ldots, \sigma_p^2(u))$. This decomposition has a clear statistical interpretation (Pourahmadi (1999)). The below diagonal entries of $P(U_i)$ are the negatives of the autoregressive coefficients $\phi_{jl}(U_i)$ in the autoregressive models

$$\hat{y}_{ij} = m_j(U_i) + \sum_{l=1}^{j-1} \phi_{jl}(U_i)\{y_{il} - m_l(U_i)\}.$$

Thus the ordinary regression coefficients of the linear regression of y_{ij} on its predecessors $y_{i(j-1)}$, …, y_{i1} are the conditional (given U_i) autoregressive coefficients $\phi_{jl}(U_i)$. The diagonal entries $\sigma_j^2(U_i)$ of $D(U_i)$ are the conditional innovation variances $\sigma_j^2(U_i) = \mathrm{var}(\epsilon_{ij}|U_i)$ with $\epsilon_{ij} = y_{ij} - \hat{y}_{ij}$. This decomposition is attractive

in that $\phi_{jl}(u)$ and $\log\{\sigma_j^2(u)\}$ are unconstrained. Let $\nu_j(u) := \log\{\sigma_j^2(u)\}$. If we have estimators $\hat{\phi}_{jl}(u)$ and $\hat{\nu}_j(u)$ of $\phi_{jl}(u)$ and $\nu_j(u)$, we immediately obtain estimators $\hat{P}(u)$ and $\hat{D}(u)$ of $P(u)$ and $D(u)$, and an estimator of $\Sigma(u)$ can be obtained as $\hat{\Sigma}(u) = \hat{P}^{-1}(u)\hat{D}(u)\hat{P}^{-1}(u)^T$, which is positive definite. The modified Cholesky decomposition was first studied in longitudinal data when the order of the multivariate variables was known. For our problem, it is used as an intermediate step in estimating $\Sigma(u)$ when variables need not be ordered. We illustrate later that the ordering of the variables has little effect on the performance of the estimator.

By the modified Cholesky decomposition of $\Sigma(U_i)$, $\Sigma^{-1}(U_i) = P(U_i)^T D^{-1}(U_i) P(U_i)$, the objective function (2.1) becomes

$$\sum_{i=1}^{n} K_h(U_i - u)\Big[\{Y_i - m(U_i)\}^T P(U_i)^T D^{-1}(U_i) P(U_i)\{Y_i - m(U_i)\} + \log(|D(U_i)|)\Big], \tag{2.2}$$

which is equal to

$$\sum_{j=1}^{p}\sum_{i=1}^{n} K_h(U_i - u)\left\{\frac{[y_{ij} - m_j(U_i) - \sum_{l=1}^{j-1}\phi_{jl}(U_i)\{y_{il} - m_l(U_i)\}]^2}{\sigma_j^2(U_i)} + \log\sigma_j^2(U_i)\right\}. \tag{2.3}$$

Here the notation $\sum_{l=1}^{0}$ means zero throughout this paper.

Because of the orthogonality between the mean and the covariance matrix in normal regression (Claeskens and Aerts (2000); Ye and Pan (2006)), we can replace $m(u)$ in (2.2) or (2.3) by some consistent estimate $\hat{m}(u)$ for estimating the autoregressive coefficient functions $\phi_{jl}(u)$. If $D(u)$ is taken in (2.2) or (2.3) to be locally constant, then the innovation variance functions do not affect the estimation of the autoregressive coefficient functions, leading to the usual local likelihood

$$\sum_{i=1}^{n} K_h(U_i - u)\Big[y_{ij} - \hat{m}_j(U_i) - \sum_{l=1}^{j-1}\{\phi_{jl}(u) + \phi_{jl}(u)'(U_i - u)\}\{y_{il} - \hat{m}_l(U_i)\}\Big]^2. \tag{2.4}$$

If $\hat{r}_i := Y_i - \hat{m}(U_i)$, then $\hat{r}_{ij} = y_{ij} - \hat{m}_j(U_i)$. If $X_i^{(j)} := (\hat{r}_{i1}, \ldots, \hat{r}_{i(j-1)}, (U_i - u)\hat{r}_{i1}, \ldots, (U_i - u)\hat{r}_{i(j-1)})$, $\Phi_j(u) := (\phi_{j1}(u), \ldots, \phi_{j(j-1)}(u))^T$, then minimization of (2.4) leads to the locally linear estimator of $\Phi_j(u)$,

$$\hat{\Phi}_j(u) = (I_{(j-1)}, \mathbf{0}_{(j-1)})\Big\{\sum_{i=1}^{n} K_h(U_i - u) X_i^{(j)T} X_i^{(j)}\Big\}^{-1} \sum_{i=1}^{n} K_h(U_i - u) X_i^{(j)T} \hat{r}_{ij}, \tag{2.5}$$

where $I_{(j-1)}$ is a $(j-1)\times(j-1)$ identity matrix and $\mathbf{0}_{(j-1)}$ is a $(j-1)\times(j-1)$ zero matrix. With $\hat{\Phi}_j(u)$, for $j = 2, \ldots, p$, we obtain the locally linear estimator $\hat{P}(u)$ of $P(u)$.

From (2.3), we use the following to estimate the innovation variance $\sigma_j^2(u)$:

$$\sum_{i=1}^{n} K_h(U_i - u)\Big\{\frac{\Big[y_{ij} - m_j(U_i) - \sum_{l=1}^{j-1}\phi_{jl}(U_i)\{y_{il} - m_l(U_i)\}\Big]^2}{\sigma_j^2(U_i)} + \log\sigma_j^2(U_i)\Big\}. \tag{2.6}$$

We fix $m_j(u) \equiv \hat{m}_j(u)$ and set $\phi_{jl}(u) \equiv \hat{\phi}_{jl}(u)$. Define $\hat{\epsilon}_{ij} := y_{ij} - \hat{m}_j(U_i) - \sum_{l=1}^{j-1}\hat{\phi}_{jl}(U_i)\{y_{il} - \hat{m}_l(U_i)\}$. Then (2.6) becomes

$$\sum_{i=1}^{n} K_h(U_i - u)\Big\{\frac{\hat{\epsilon}_{ij}^2}{\sigma_j^2(U_i)} + \log\sigma_j^2(U_i)\Big\}. \tag{2.7}$$

Due to the orthogonality among the mean, the autoregressive coefficient and the innovation variance (Pourahmadi (2000)), it is reasonable to use (2.7) to estimate the innovation variances. Unfortunately, the explicit introduction of local linearity into the estimation of $\sigma_j^2(u)$ via (2.7) does not guarantee positivity. However, $\log\sigma_j^2(u)$ is unconstrained. We follow Yu and Jones (2004) to overcome this difficulty by modeling $\log\sigma_j^2(u)$ as locally linear, and then minimizing

$$\sum_{i=1}^{n} K_h(U_i - u)\Big\{\hat{\epsilon}_{ij}^2\exp\{-\nu(u) - \nu'(u)(U_i - u)\} + \nu(u) + \nu'(u)(U_i - u)\Big\} \tag{2.8}$$

to obtain the local linear estimator of the innovation variance $\hat{\sigma}_j^2(u) = e^{\hat{\nu}(u)}$. Then we have the local linear estimator of $D(u)$, $\hat{D}(u)$. By the modified Cholesky decomposition, a locally linear estimator of $\Sigma(u)$ is basically $\hat{\Sigma}(u) = \hat{P}^{-1}(u)\hat{D}(u)\hat{P}^{-1}(u)^T$, which is positive definite.

In (2.5) and (2.8), we use the same bandwidth h but this is unnecessary. To adapt to different smoothness, we can use different bandwidths for different components of the conditional covariance matrix. In particular, we can introduce different bandwidths for all $\hat{\Phi}_j(u)$ (for $j = 2, \ldots, p$) as in (2.5), and all $\hat{\sigma}_j^2(u)$ (for $j = 1, \ldots, p$) as in (2.8). To choose the right amount of smoothness, we use leave-one-out cross validation for estimating the autoregressive coefficient $\Phi_j(u)$ (for $j = 2, \ldots, p$),

$$\hat{h}_j^{AR} = \arg\min_h \sum_{i=1}^{n}\Big[y_{ij} - \hat{m}_j(U_i) - \sum_{l=1}^{j-1}\hat{\phi}_{jl}(U_i; h)^{(-i)}\{y_{il} - \hat{m}_l(U_i)\}\Big]^2,$$

where $\hat{\phi}_{jl}(u; h)^{(-i)}$ is the estimate of $\phi_{jl}(u)$ obtained by leaving out the ith observation according to (2.5) with the bandwidth h. Similarly, the leave-one-out cross-validation bandwidth for estimating the innovation variance $\sigma_j^2(u)$ (for $j = 1, \ldots, p$) is given by

$$\hat{h}_j^{IV} = \arg\min_h \sum_{i=1}^{n} \Big\{ \frac{\hat{\epsilon}_{ij}^2}{\hat{\sigma}_j^2(U_i;h)^{(-i)}} + \log\{\hat{\sigma}_j^2(U_i;h)^{(-i)}\} \Big\},$$

where $\hat{\sigma}_j^2(u;h)^{(-i)}$ is the estimate of $\sigma_j^2(u)$ derived without the ith observation according to (2.8) with the bandwidth h.

We now discuss the estimation of $m(u)$. We use a local linear estimator for the conditional mean $m(u)$. If we take $\Sigma(u)$ in (2.1) to be the identity matrix I_p, the mean function $m(\cdot)$ is estimated by by minimizing

$$\sum_{i=1}^{n} K_h(U_i - u)\Big[\{Y_i - m_0 - m_1(U_i - u)\}^T\{Y_i - m_0 - m_1(U_i - u)\}\Big],$$

which yields the local linear estimator of $m(u)$,

$$\hat{m}(u) = (I_p, \mathbf{0}_p)\Big\{ \sum_{i=1}^{n} K_h(U_i - u) \begin{pmatrix} 1 & (U_i - u) \\ (U_i - u) & (U_i - u)^2 \end{pmatrix} \otimes I_p \Big\}^{-1} \times \sum_{i=1}^{n} K_h(U_i - u)\{(1 \quad (U_i - u))^T \otimes I_p\} Y_i.$$

3. Asymptotic Results

We study the asymptotic properties of the proposed estimators in Theorem 1 and Theorem 2 in this section. We investigate the global convergence of the proposed conditional covariance matrix estimator under the Kullback-Leibler loss (Yuan and Lin (2007); Levina, Rothman, and Zhu (2008)) and the Frobenius loss (Bickel and Levina (2008)) in Theorem 3. Some technical conditions are imposed; they may not be the weakest possible conditions, but are imposed to facilitate the proofs.

Regularity conditions.

(a) $U_1, \ldots, U_n$ are independently and identically sampled from a density $f(\cdot)$ with compact support Ω; f is twice continuously differentiable, and is bounded away from 0 on its support.

(b) The kernel function $K(\cdot)$ is a symmetric density function about 0. There exists some $s > 0$ such that $\int K(u)^{2+s}u^j du < \infty$, for $j = 0, 1, 2$. Moreover, $\sup_u K(u) < K_1 < \infty$ and $\sup_u |K'(u)| < K_2 < \infty$.

(c) The bandwidth satisfies $h \to 0$ and $nh^5 \to c > 0$ for some $c > 0$, as $n \to \infty$.

(d) The mean function $m(u)$, $\Phi_j(u)$ $(j = 2, \ldots, p)$, and the innovation variance function $\sigma_j^2(u)$ $(j = 1, \ldots, p)$ have continuous second order derivatives.

Let $\Sigma(u)_{(j-1,j-1)}$ denote the $(j-1)$-th main submatrix of $\Sigma(u)$ for $j = 2, \ldots p$. Let $\mu_2 := \int u^2 K(u)du$, $\gamma_0 := \int K^2(u)du$, and $c_n := h^2 + \{\log(1/h)/(nh)\}^{1/2}$. Define $r_i := Y_i - m(U_i)$, so $r_{ij} = y_{ij} - m_j(U_i)$. Define $\epsilon_{ij} := r_{ij} - \sum_{l=1}^{j-1} \phi_{jl}(U_i) r_{il}$.

Theorem 1. *If* (a)−(d) *hold, we have*

$$\hat{\Phi}_j(u) - \Phi_j(u) = f^{-1}(u)\{\Sigma(u)_{(j-1,j-1)}\}^{-1}\frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u)\begin{pmatrix} r_{i1} \\ \cdot \\ \cdot \\ \cdot \\ r_{i(j-1)} \end{pmatrix}\epsilon_{ij}$$

$$+\frac{1}{2}\mu_2\Phi_j^{''}(u)h^2 + o_p(c_n), \tag{3.1}$$

which holds uniformly in u, for $j = 2, \ldots, p$.

Theorem 2. *If* (a)−(d) *hold, we have*

$$\hat{\sigma}_j^2(u) - \sigma_j^2(u) = f^{-1}(u)\frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u)\Big\{\epsilon_{ij}^2 - \sigma_j^2(U_i)\Big\}$$

$$+\frac{1}{2}\mu_2[\log\{\sigma_j^2(u)\}]^{''}\sigma_j^2(u)h^2 + o_p(c_n)$$

which holds uniformly in u, for $j = 1, \ldots, p$.

Remark 1. The first term of (3.1) dominates the asymptotic variance of $\hat{\Phi}_j(u)$ while the second term dominates the asymptotic bias. The asymptotic variance and the asymptotic bias of $\hat{\Phi}_j(u)$ are both independent of $\hat{m}(u)$. We have that the asymptotic result of $\hat{\Phi}_j(u)$ does not change when we replace $\hat{m}(u)$ by the Nadaraya-Watson kernel estimator $\hat{m}_{LC}(u)$ in (2.5) from the proof of Theorem 1. Further, the asymptotic variance and the asymptotic bias of $\hat{\Phi}_j(u)$ using $\hat{m}(u)$ or $\hat{m}_{LC}(u)$ are equal to the asymptotic variance and the asymptotic bias of the estimator of $\Phi_j(u)$ based on the true conditional mean $m(u)$. Similarly, the asymptotic variance and the asymptotic bias of $\hat{\sigma}_j^2(u)$ are independent of $\hat{m}(u)$.

We now investigate the accuracy in estimating the covariance matrix $\Sigma(u)$. We define the Kullback-Leibler loss and the Frobenius loss to evaluate it:

$$L(\Sigma(u), \hat{\Sigma}(u))_{KL} = E\Big[\text{trace}\{\Sigma(u)\hat{\Sigma}^{-1}(u)\} - \log|\Sigma(u)\hat{\Sigma}^{-1}(u)|\Big] - p,$$

$$L(\Sigma(u), \hat{\Sigma}(u))_F = E\Big[\text{trace}[\{\hat{\Sigma}^{-1}(u) - \Sigma^{-1}(u)\}^2]\Big].$$

Theorem 3. *Under* (a)−(d), *we have*

$$L(\Sigma(u), \hat{\Sigma}(u))_{KL} = \frac{1}{4}\mu_2^2 h^4 trace\Big\{P^{''}(u)^T P^{''}(u)\Sigma(u)D^{-1}(u) + \frac{1}{2}D_{B1}^2(u)\Big\}$$

$$+\frac{\gamma_0}{nhf(u)}trace\Big\{P^*(u)\Sigma(u)D^{-1}(u) + I\Big\} + o(h^4 + \frac{1}{nh}),$$

where $P^*(u)$ *is a diagonal matrix with first diagonal entry* 0 *and* j*-th diagonal entry*

$\sigma_j^2(u) trace[\{\Sigma(u)_{(j-1,j-1)}\}^{-1}]$ *for* $j = 2, \ldots, p$, *while* $D_{B1}(u)$ *is a diagonal matrix with* j*-th diagonal entry* $[\log\{\sigma_j^2(u)\}]''$, *and*

$$
\begin{aligned}
&L(\Sigma(u), \hat{\Sigma}(u))_F \\
&= \frac{1}{4}\mu_2^2 h^4 trace\Big\{2P''(u)P''(u)P(u)^T P(u)^T D^{-2}(u) + 2P''(u)^T P''(u)\Sigma^{-1}(u)D^{-1}(u) \\
&\quad + \Sigma^{-2}(u)D_{B1}^2(u) + 2P''(u)\Sigma^{-1}(u)P(u)D^{-1}(u)D_{B1}(u)\Big\} \\
&\quad + \frac{\gamma_0}{nhf(u)} trace\Big\{P^*(u)\Sigma^{-1}(u)D^{-1}(u) + 2\Sigma^{-2}(u)\Big\} + o(h^4 + \frac{1}{nh}).
\end{aligned}
$$

Remark 2. The accuracy of $\hat{\Sigma}(u)$ does not depend on $\hat{m}(u)$. Furthermore, the accuracy of the estimator for the conditional covariance matrix based on $\hat{m}(u)$ or $\hat{m}_{LC}(u)$ is the same as the one based on the true conditional mean $m(u)$.

Remark 3. Our proposed estimator of the nonparametric covariance matrix has convergence rate $h^4 + 1/(nh)$ under the Kullback-Leibler loss and the Frobenius loss, in which $1/(nh)$ comes from the asymptotic variance term and h^4 comes from the asymptotic bias term. This is the familiar bias-variance trade-off and the optimal convergence rate is achieved when $h \propto n^{-1/5}$.

Our proposed estimator is permutation invariant in a loose sense. First, the rate of convergence in Theorem 3 does not depend on the ordering of the variables, although different orderings give different scaling constants in the asymptotic bias and variance term. Second, our approach is based on the local linear estimating procedure, and allows one to apply different degrees of smoothness to different components of the decomposition, making our approach more adaptive than the local constant approach. Third, our numerical results suggest that the accuracy of our proposed estimators changes little when variables are permuted; see Study 1-4 in the next section. We conclude that, although our proposed method uses the modified Cholesky decomposition which depends on the ordering of the variables as an intermediate in estimating $\Sigma(u)$, the ordering has little effect on performance. Rothman et al. (2008) proposed a different approach for estimating permutation-invariant concentration matrices.

4. Numeric Studies

In this section, we report on several simulation studies to evaluate the finite sample performances of our proposed estimators in Sections 2, and illustrate the proposed approach on a data set. For brevity, we refer to our approach as the local linear estimator (LL) and that of Yin et al. (2010) as the local constant estimator (LC).

For each simulation study, we generated 200 datasets, each consisting of $n = 300$ observations. We sampled $U_i,\ i = 1, \ldots, n$, independently from the truncated normal distribution with density function $f_0(u) = \exp(-u^2/2)/\int_{-1}^{1} \exp(-t^2/2)dt$, for $-1 \le u \le 1$ and zero elsewhere. The response variable was generated according to $Y_i \sim N(m(U_i), \Sigma(U_i))$. Since our emphasis is on estimating the conditional covariance matrix $\Sigma(u)$, we used the local linear estimate $\hat{m}(u)$ for estimating the mean function in both methods. For all simulations, we set $p = 5$. We used the Kullback-Leibler and Frobenius losses as the criteria to compare the two estimators. Specifically, for each dataset, we calculated the median of n Kullback-Leibler losses and Frobenius losses for each method, defined as

$$\begin{aligned} \text{Median Kullback-Leibler Loss} &= \text{median}\{\bigtriangledown_{KL}(U_i),\ i = 1, \ldots, n\}, \\ \text{Median Frobenius Loss} &= \text{median}\{\bigtriangledown_{F}(U_i),\ i = 1, \ldots, n\}, \end{aligned}$$

where $\bigtriangledown_{KL}(u)$ and $\bigtriangledown_{F}(u)$ are the Kullback-Leibler and Frobenius losses for an estimator $\hat{\Sigma}(u)$. For brevity, "Median Kullback-Leibler Loss" and "Median Frobenius Loss" are referred to as MKLL and MFL, respectively. In order to overcome the impact of the covariance matrix estimators on boundary points, we summarize the simulation results using the sample median instead of the sample mean to be consistent with Yin et al. (2010). For all the studies, we look at the original order $Y_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4}, y_{i5})^T$ and a random permutation of Y_i for each generated dataset; this is useful for investigating the sensitivity of the proposed method to permutation of the variables.

Study 1. We considered a nonparametric covariance model by setting the mean function as $m(u) = \mathbf{0} = (0, 0, 0, 0, 0)^T$ and $\Sigma(u) = P^{-1}(u)D(u)P^{-1}(u)^T$ where $D(u)$ is a diagonal matrix with diagonal entries $(\exp(u/2), \cos(\pi u) + 1.1, \exp(u/2), \cos(\pi u) + 1.1, \exp(-u/2))$ and

$$P(u) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ -\phi(u) & 1 & 0 & 0 & 0 \\ -\phi(u)/2 & -\phi(u)/2 & 1 & 0 & 0 \\ -\phi(u)/4 & -\phi(u)/4 & -\phi(u)/4 & 1 & 0 \\ -\phi(u))/8 & -\phi(u)/8 & -\phi(u)/8 & -\phi(u)/8 & 1 \end{pmatrix}.$$

Here $\phi(u)$ is the density of the standard normal distribution. For this example, we assumed that $m(u)$ was known.

Study 2. The model used in the this study was identical to the model in Study 1, except that the mean function $m(u)$ was set to $(\cos(u), \sin(u), \cos(u), \sin(u), \cos(u))^T$. This study was to investigate how the estimation of the mean function $m(u)$ affected the estimation of the covariance function $\Sigma(u)$.

Study 3. The mean function was set as $m(u) = \mathbf{0} = (0, 0, 0, 0, 0)^T$ and the conditional covariance matrix had an AR(1) structure, $\Sigma(u) = D^{1/2}(u)R(u)D^{1/2}(u)$, where $D(u)$ is defined in Study 1 and

$$R(u) = \begin{pmatrix} 1 & \phi(u) & \phi^2(u) & \phi^3(u) & \phi^4(u) \\ \phi(u) & 1 & \phi(u) & \phi^2(u) & \phi^3(u) \\ \phi^2(u) & \phi(u) & 1 & \phi(u) & \phi^2(u) \\ \phi^3(u) & \phi^2(u) & \phi(u) & 1 & \phi(u) \\ \phi^4(u) & \phi^3(u) & \phi^2(u) & \phi(u) & 1 \end{pmatrix}.$$

This study together with Study 1 and Study 2 investigated the performance of the proposed method when the variables had a natural ordering. The other purpose of this study was to investigate how the proposed method performed when the true covariance matrix did not admit an explicit modified Cholesky decomposition.

Study 4. The mean function was set as $m(u) = \mathbf{0} = (0, 0, 0, 0, 0)^T$ and the conditional covariance matrix had an exchangeable structure, $\Sigma(u) = D^{1/2}(u)R(u)$ $D^{1/2}(u)$, where $D(u)$ is defined in Study 1 and

$$R(u) = \begin{pmatrix} 1 & \phi(u) & \phi(u) & \phi(u) & \phi(u) \\ \phi(u) & 1 & \phi(u) & \phi(u) & \phi(u) \\ \phi(u) & \phi(u) & 1 & \phi(u) & \phi(u) \\ \phi(u) & \phi(u) & \phi(u) & 1 & \phi(u) \\ \phi(u) & \phi(u) & \phi(u) & \phi(u) & 1 \end{pmatrix}.$$

This study was designed to investigate the performance of the proposed method when the variables had no natural ordering. This study also investigated how the proposed method performed when the true covariance matrix did not have an explicit modified Cholesky decomposition.

Study 5. The mean function was set as $m(u) = \mathbf{0} = (0, 0, 0, 0, 0)^T$ and $\Sigma(u) = P^{-1}(u)D(u)P^{-1}(u)^T$ where $D(u)$ is identical to that of Study 1 and

$$P(u) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ -\cos(u) & 1 & 0 & 0 & 0 \\ -\cos(u) & -\sin(u) & 1 & 0 & 0 \\ -\cos(u) & -\sin(u) & -\cos(u) & 1 & 0 \\ -\cos(u) & -\sin(u) & -\cos(u) & -\sin(u) & 1 \end{pmatrix}.$$

Here the correlations between y_4 and y_5 were larger than 0.958 when $0.5 < U \leq 1$. This study was used to show that high correlations among components of Y affect the performance of the proposed estimator.

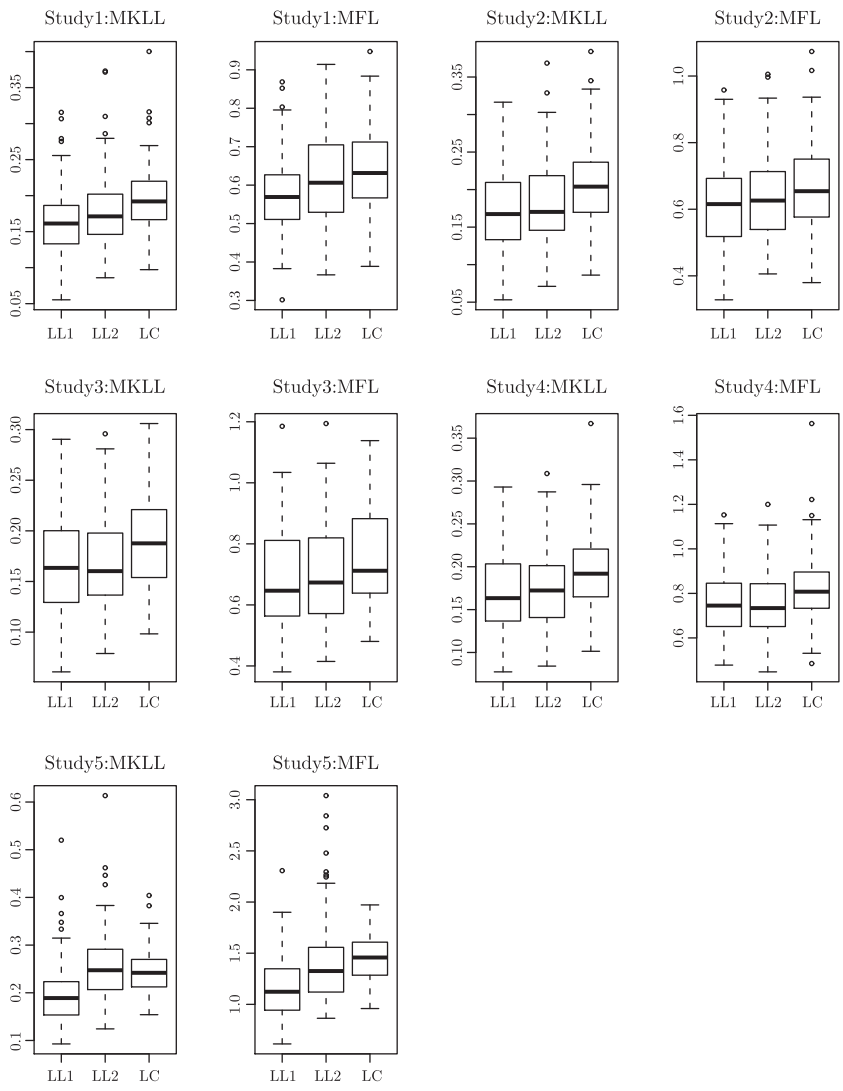

Figure 1. Box-plots of 200 median Kullback-Leibler losses (MKLLs) and 200 median Frobenius losses (MFLs) over 200 datasets for the local linear approach (LL) and the local constant approach (LC) for the five simulation studies.

The results for the above five simulation studies are summarized in Figure 1, where "LL1" is the LL method using the original order, and "LL2" is the LL method after a random permutation of the variables. First, studies 1–4 show that the proposed local linear method outperforms the local constant estimator in terms of the Kullback-Leibler and Frobenius losses, sometimes by a large margin. Second, the estimation of the mean function does not change this pattern from

the results of Study 2. Third, the results of Study 1–4 indicate that the superior performance of the local linear estimator continues with permuted variables, and that the Kullback-Leibler and Frobenius losses of the proposed estimators were not seriously affected by the permutation. Fourth, even if the generating covariance matrix did not admit a clear modified Cholesky decomposition, the local linear method continued to outperform in studies 3 and 4. Fifth, even though the variables have no natural ordering, the local linear estimating method based on the modified Cholesky decomposition works very well in Study 4. Finally, when we permuted the variables randomly, Study 5 has the local linear method performing worse than the local constant approach. This is caused by the high correlation between y_4 and y_5 in estimating the conditional autoregressive coefficients. It was previously observed that this could yield biased and inefficient estimators (Kumar (1975)).

Boston Housing Data

We illustrate the proposed local linear method by an application to the Boston housing dataset that contains a total of 506 observations (Fan and Huang (2005); Yin et al. (2010)). We considered five social economic variables: crime rate (y_1), full-value property-tax rate (y_2), pupil-teacher ratio (y_3), median value of owner-occupied homes (y_4) and average number of rooms per dwelling (y_5). Let $Y = (y_1, y_2, y_3, y_4, y_5)^T$ and take the index random variable as the square root of the percentage of lower status (U). The y-variables were standardized before analysis. The conditional covariance model is denoted as $\mathrm{Var}(Y|U) = \Sigma(U)$, and we investigated the change of the correlation structure of Y in response to a change in the percentage of lower status.

A total of 450 observations were randomly chosen as training data and the remaining 56 observations were used as testing data. Using the training data, we obtained the local linear and the local constant estimators of the conditional covariance matrix. For a fair comparison, we used the local linear estimator $\hat{m}(\cdot)$ for estimating the mean in both methods, based on the training data. The prediction performances were measured by the log-likelihood-like loss measure

$$\Delta = \frac{1}{n^*}\sum_{i=1}^{n^*}\Big[\{Y_i^* - \hat{m}(U_i^*)\}^T\hat{\Sigma}^{-1}(U_i^*)\{Y_i^* - \hat{m}(U_i^*)\} + \log(|\hat{\Sigma}(U_i^*)|)\Big],$$

where $\hat{\Sigma}(\cdot)$ is the covariance matrix estimator based on the training data and (Y_i^*, U_i^*) $(i = 1, \ldots, n^*)$ is the testing data. This procedure was replicated 100 times. The median of the 100 Δs was 0.300 when the local constant method was used to estimate $\Sigma(u)$, and the median was 0.024 when the local linear method is used. We also note that the median of the 100 Δs was 1.560 if the covariance estimator

was assumed constant and was estimated via $n^{-1}\sum_{i=1}^{n}\{Y_i-\hat{m}(U_i)\}\{Y_i-\hat{m}(U_i)\}^T$ where $(Y_i,\ U_i)$ $(i = 1,\ldots,n)$ were from the training data. This discrepancy suggests that it is preferable to take a nonparametric modelling approach to estimate the covariance matrix of the five social economic variables.

Since the variables in $Y = (y_1, y_2, y_3, y_4, y_5)^T$ have no natural ordering, we investigated the impact of permutation and found that the proposed local linear method still outperformed the local constant method. For each data in the above 100 runs, we randomly permuted the order of the variables and calculated the Δ quantity as before. The median of the 100 Δs was 0.052 for the local linear covariance estimator, smaller than the 0.300 for the local constant estimator and 1.560 for the global constant covariance estimator.

5. Discussion

The Cholesky decomposition of covariance is considered appropriate when variables have a natural ordering (Pourahmadi (1999, 2000); Levina, Rothman, and Zhu (2008)). We show, however, that serving as an intermediate step, the decomposition continues to be useful for estimating covariances, even when variables have no natural ordering. It is of interest to extend the current work to estimate multiple covariance matrices (Guo et al. (2011)), to deal with structured matrices (Levina, Rothman, and Zhu (2008)), and to investigate alternative decompositions for this task (Rothman, Levina, and Zhu (2010); Zhang and Leng (2012)).

In the paper, we only consider the estimation of a conditional covariance matrix when $q = 1$ and, in principle, the technique can be generalized to $q > 1$. However, multivariate kernel smoothing may suffer from the curse of dimensionality, and is less useful (Fan and Gijbels (1996)). To overcome the dimensionality problem, one can try a semiparametric model for each entry of the conditional covariance matrix, which combines the flexibility of nonparametric regression and parsimony of linear regression. Further studies along this line are needed.

Acknowledgement

We thank the joint Editor, an associate Editor and two reviewers for their constructive comments. Chen's research is supported in part by National Nature Science Foundation of China(no. 11401593), Specialized Research Fund for the Doctoral Program of Higher Education of China (no. 20130162120086), China Postdoctoral Science Foundation (no. 2013M531796), and China Postdoctoral Science Foundation (no. 2014T70778).

References

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199-227.

Claeskens, G. and Aerts, M. (2000). On local estimating equations in additive multiparameter models. *Statist. Probab. Lett.***49**, 139-148.

Dai, M. and Guo, W. (2004). Multivariate spectral analysis using Cholesky decomposition. *Biometrika* **1**, 629-643.

Engle, R. (2002). Dynamic conditional correlation: a simple class of multivariate GARCH models. *J. Busi. Econ. Statist.* **20**, 339-350.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196-216.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and its Applications.* Chapman and Hall, London.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11**, 1031-1057.

Fan, J., Huang, T. and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *J. Amer. Statist. Assoc.* **102**, 632-641.

Guo, J., Levina, L., Michailidis, G. and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98**, 1-15.

Kolar, M., Song, L., Ahmed, A. and Xing, E. (2010). Estimating time-varying networks. *Ann. Appl. Statist.* **4**, 94-123.

Kumar, T. K. (1975). Multicollinearity in regression analysis. *Rev. Econom. Statist.* **57**, 365-366.

Leng, C., Zhang, W. and Pan, J. (2010). Semiparametric mean-covariance regression analysis for longitudinal data. *J. Amer. Statist. Assoc.* **105**, 181-193.

Levina, E., Rothman, A. J. and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *Ann Appl. Statist.* **2**, 245-263.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterization. *Biometrika* **86**, 677-690.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix. *Biometrika* **87**, 425-435.

Rosen, O. and Stoffer, D. S. (2007). Automatic estimation of multivariate spectra via smoothing splines. *Biometrika* **94**, 335-345.

Rothman, A., Bickel, P., Levina, L. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic J. Statist.* **2**, 494-515.

Rothman, A., Levina, L. and Zhu, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika* **97**, 539-550.

Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90**, 831-844.

Ye, H. and Pan, J. (2006). Modelling of covariance structures in generalised estimating equations for longitudinal data. *Biometrika* **93**, 927-941.

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

Yu, K. and Jones, M. C. (2004). Likelihood-based local linear estimation of the conditional variance function. *J. Amer. Statist. Assoc.* **99**, 139-144.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**, 19-35.

Zhang, W. and Leng, C. (2012). A moving average Cholesky factor model in covariance modeling for longitudinal data. *Biometrika* **99**, 141-150.

School of Mathematics and Statistics, Central South University, Changsha, Hunan 410083, China.

E-mail: chenzq453@gmail.com

Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K.

E-mail: C.Leng@warwick.ac.uk

(Received May 2013; accepted August 2014)