Statistica Sinica **20** (2010), 469-479

NONPARAMETRIC COVARIANCE MODEL

Jianxin Yin[1], Zhi Geng[1], Runze Li[2] and Hansheng Wang[1]

[1]*Peking University and* [2]*Pennsylvania State University*

Abstract: There has been considerable attention paid to estimation of conditional variance functions in the literature. We propose a nonparametric model for the conditional covariance matrix. A kernel estimator is developed, its asymptotic bias and variance are derived, and its asymptotic normality is established. A data example is used to illustrate the proposed procedure.

Key words and phrases: Conditional variance, heteroscedasticitym kernel regression, nonparametric covariance model, volatility.

1. Introduction

In addition to estimation of means, estimation of variance functions or covariance matrices is a common statistical problem that has application in many different fields. They include, but are not limited to, graphical modeling (Edwards (2000), and Drton and Perlman (2004)), longitudinal data analysis (Diggle and Verbyla (1998), and Smith and Kohn (2002)), risk management (Ledoit and Wolf (2004)), and machine learning (Bilmes (2000)). Covariance matrix estimation has received a lot of attention in the recent literature (Pourahmadi (1999, 2000), Huang, Liu, Pourahmadi and Liu (2006), Meinshausen and Buhlmann (2006), Bickel and Levina (2008), Levina, Rothman and Zhu (2007), and Lam and Fan (2007)). Estimation of covariance functions, correlation functions, and multivariate spectra is also an active research area. Nonparametric and semiparametric estimation of covariance and correlation functions for longitudinal data have been carefully studied (Wu and Pourahmadi (2003), Yao, Müller and Wang (2005a,b), and Fan, Huang and Li (2007)). Dai and Guo (2004) and Rosen and Stoffer (2007) advocated the use of Cholesky decomposition for spectral analysis of multivariate time series. Recently, Li, Wang, Hong, Turner, Lupton and Carroll (2007) proposed a nonparametric kernel estimator for the correlation function of a random field.

Nonparametric regression models have been used in various areas. A number of estimation procedures for nonparametric mean regression have been extensively studied. To our best knowledge, however, there are few references available for nonparametric models for a conditional covariance matrix, although there are some references on nonparametric conditional variance functions; see Ruppert, Wand, Holst and Hossjer (1997), Fan and Yao (1998) and references therein. This paper intends to fill this gap by developing a nonparametric model for a conditional covariance matrix. The proposed model can be regarded as a natural extension of existing nonparametric models for conditional variance. Specifically, for two random vectors X and U, we model the conditional covariance of X given U as $\mathrm{Cov}\,(X|U) = \Sigma(U)$, whose every component is assumed to be an unknown but smooth function of U. We propose a kernel estimator for $\Sigma(U)$ estimation. The sampling properties of the proposed estimator are systematically studied: we derive the asymptotic bias and variance of the kernel estimator, and establish its asymptotic normality. Our findings indicate that, without knowing the true regression function, we may estimate the conditional covariance matrix asymptotically as well as if we knew the true regression function in advance. We further investigate its asymptotic behavior under Stein loss and quadratic loss (Muirhead (1982)). For a practical implementation, a local likelihood-based cross validation method is developed for an automatic selection of the bandwidth.

The rest of the article is organized as follows. Section 2 introduces the model, notation, and the estimator. The asymptotic properties of the estimator are studied in this section. Numerical studies based on the Boston Housing data set are presented in Section 3. The article concludes with a short discussion in Section 4. All technical details, and a simulation study, are left to a separate Technical Appendix that can be found at `http://www.stat.sinica.edu.tw/statistica`.

2. Nonparametric Covariance Model

2.1. The model and estimator

Let $X = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ be a p-dimensional random vector and $U = (U_1, \ldots, U_q)^\top \in \mathbb{R}^q$ be the associated *index* random vector. A nonparametric covariance model assumes that, conditional on $U = u$, X follows a distribution with mean $m(u) = \{m_1(u), \ldots, m_p(u)\}^\top$ and covariance $\Sigma(u) = \{\sigma_{j_1 j_2}(u)\}$. Furthermore, we assume that both $m(u)$ and $\Sigma(u)$ are unknown but smooth functions of u. In this paper, we consider $q = 1$ only. Extension to multivariate U is straightforward in theory, but it is less useful in practice due to the curse of dimensionality.

To estimate $\Sigma(u)$ consistently, we develop a Nadaraya-Watson (NW) kernel estimator for both $m(\cdot)$ and $\Sigma(\cdot)$. Let $K(u)$ be a symmetric kernel density function and $K_h(u) = h^{-1}K(u/h)$ be the scaled kernel function with a bandwidth $h > 0$. To motivate our estimate, we temporarily assume that given U, X follows a normal distribution, although the conditional normality assumption is unnecessary. Suppose (X_i, U_i) with $X_i = (X_{i1}, \ldots, X_{ip})^\top$, $i = 1, \ldots, n$, is a random sample from the population (X, U). The kernel method is to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\left[\left\{X_i - m(u)\right\}^\top \Sigma^{-1}(u)\left\{X_i - m(u)\right\} - \log\left(|\Sigma^{-1}(u)|\right)\right]K_h(U_i - u), \quad (2.1)$$

where u is an arbitrary point in the support of U.

By minimizing the objective function (2.1), the resulting NW kernel estimators are:

$$\hat{m}(u) = \left\{\sum_{i=1}^{n} K_h(U_i - u)X_i\right\}\left\{\sum_{i=1}^{n} K_h(U_i - u)\right\}^{-1}, \quad (2.2)$$

$$\hat{\Sigma}(u) = \left[\sum_{i=1}^{n} K_h(U_i - u)\left\{X_i - \hat{m}(U_i)\right\}\left\{X_i - \hat{m}(U_i)\right\}^\top\right]\left\{\sum_{i=1}^{n} K_h(U_i - u)\right\}^{-1}. \quad (2.3)$$

Since the kernel estimate for the regression function has been well studied, we study only the asymptotic properties of the proposed conditional covariance estimator in the next section. In (2.2) and (2.3) the same bandwidth h is used. This is unnecessary. Note that one may calculate the estimator $\hat{m}(\cdot)$ and $\hat{\Sigma}(\cdot)$ separately. This allows us to easily use different bandwidths for different components of the conditional mean and also the conditional covariance, in order to be adaptive to different smoothnesses. For example, we may use different bandwidths h_{1j} for estimation of the j-th component of the regression function, i.e.,

$$\hat{m}_j(u) = \left\{\sum_{i=1}^{n} K_{h_{1j}}(U_i - u)X_{ij}\right\}\left\{\sum_{i=1}^{n} K_{h_{1j}}(U_i - u)\right\}^{-1}, \quad (2.4)$$

for $j = 1, \ldots, p$. One may further employ the local linear regression to estimate the mean function. Local linear estimators have several nice properties, such as high statistical efficiency in an asymptotic minimax sense, design adaptivity, and automatic correction of boundary effects (Fan (1993), and Cheng, Fan and Marron (1997)). Since the main goal of this paper is to estimate the conditional covariance matrix, our attention will focus on the NW kernel estimate. The results in the next section are still valid for local linear regression under certain conditions.

Remark 1. As to the estimation of the conditional covariance matrix, one may use different bandwidths for different elements of $\Sigma(u)$. However, the resulting estimate with different bandwidths cannot be guaranteed to be positive definite (Li et al. (2007)). In practice, positive definiteness is a desirable property and we suggest using the same bandwidth for all elements.

Remark 2. As an alternative, one may extend the local linear regression to estimate the conditional covariance matrix. To develop a sensible local linear estimator for the conditional covariance matrix without destroying positive-definiteness seems not to be very straightforward. For this reason, we advocate the use of the kernel estimate in (2.3).

2.2. Theoretical properties

Since the kernel regression $\hat{m}_j(\cdot)$ has been extensively studied, we study only the asymptotic properties of the conditional covariance estimator in (2.3). Without loss of generality, it is assumed that the bandwidth h_{1j}'s in (2.4) are the same and equal the bandwidth h for $\hat{\Sigma}(u)$ in our theoretic development. The following regularity conditions are imposed, they are not the weakest possible but they facilitate the proofs.

Regularity conditions

(C1) (The density of the *index* variable) U has a compact support and a probability density $f(U)$, bounded away from 0 and with two continuous derivatives.

(C2) (The moment requirement) For any $1 \le j_1,\ j_2 \le p$, there exists a constant $\delta \in [0, 1)$ such that $\sup_u E\{|X_{j_1}(u)X_{j_2}(u)|\}^{2+\delta} < \infty$.

(C3) (Smoothness of the conditional mean) The conditional mean $m_j(\cdot)$ has two continuous derivatives.

(C4) (Smoothness of the conditional variance) $E\{X_{j_1}^{k_1} X_{j_2}^{k_2} X_{j_3}^{k_3} X_{j_4}^{k_4} | U = u\}$ has two continuous derivatives in u for $k_1, k_2, k_3, k_4 \in \{0, 1\}$, where $j_1,\ j_2,\ j_3$, and j_4 are not necessarily different.

(C5) (The bandwidth) $h \to 0$ and $nh^5 \to c > 0$ for some $c > 0$.

(C6) (The kernel function) $K(u)$ is a bounded probability density function symmetric about 0. For the δ in (C2), $\int K^{2+\delta}(v)v^j dv < \infty$ for $j = 0, 1, 2$. For two arbitrary indices u_1 and u_2, $|K(u_1) - K(u_2)| \le K_c|u_1 - u_2|$ for some $K_c > 0$.

Conditions (C1) and (C2) are rather standard technical assumptions (Li (2006)). Conditions (C3) and (C4) are necessary smoothness constraints (Fan (1993), and Yao and Tong (1998)). By (C5) we know that the optimal convergence rate is $n^{-1/5}$. Condition (C6) is a standard requirement for the kernel function (Yao and Tong (1996)), which is trivially satisfied by the Gaussian and Epanechnikov kernels.

Let $\hat{\Sigma}(u) = \{\hat{\sigma}_{j_1 j_2}(u)\}$ with $1 \le j_1, j_2 \le p$. For an arbitrary function $g(u)$, $\dot{g}(u)$ and $\ddot{g}(u)$ denote its first and second order derivatives, respectively. Let $\nu_0 = \int_{-\infty}^{\infty} K^2(u)du$ and $\mu_2 = \int_{-\infty}^{\infty} u^2 K(u)du$. Then the asymptotic behavior of $\hat{\sigma}_{j_1 j_2}(u)$ is characterized by the following theorem.

Theorem 1. *Under* (C1)–(C6),

$$\sqrt{nh}\Big\{\hat{\sigma}_{j_1j_2}(u)-\sigma_{j_1j_2}(u)-\theta_n\Big\}\to N\Big(0,f^{-1}(u)\nu_0\omega_{j_1j_2}(u)\Big)$$

in distribution, where

$$\theta_n=\frac{h^2\mu_2}{2}\Big\{\ddot{\sigma}_{j_1j_2}(u)+2\dot{\sigma}_{j_1j_2}(u)\frac{\dot{f}(u)}{f(u)}\Big\}, \tag{2.5}$$

$f(u)$ *is the probability density function of* U *evaluated at* $U=u$, $\varepsilon_{j_1j_2}(i)=\{X_{ij_1}-m_{j_1}(U_i)\}\{X_{ij_2}-m_{j_2}(U_i)\}-\sigma_{j_1j_2}(U_i)$, *and* $\omega_{j_1j_2}(U_i)\triangleq \mathrm{Var}\,(\varepsilon_{j_1j_2}(i)\big|U_i)$.

The proof can be found in Appendix B of the separate Technical Appendix. By Theorem 1, we know that the variance of $\hat{\sigma}_{j_1j_2}(u)$ is of the order $(nh)^{-1}$ and its bias is of the order h^2. Thus, the optimal nonparametric convergence rate of $n^{-2/5}$ can be achieved by setting $h\propto n^{-1/5}$. The asymptotic bias and variance in Theorem 1 are the same as those for the one replacing $\hat{m}(\cdot)$ with the true regression function $m(\cdot)$ in $\hat{\Sigma}(u)$. In other words, Theorem 1 demonstrates that without knowing the regression function, we may estimate the conditional covariance matrix asymptotically as well as if we knew the true regression function in advance. From our theoretic derivation, if all bandwidths $h_{1j}=O(n^{-1/5})$ are used, the results in Theorem 1 are still valid. This implies that no undersmooth of $\hat{m}$ is required in the estimation of the conditional covariance matrix.

It is certainly of interest to evaluate the global convergence of $\hat{\Sigma}(u)$. Here we consider two widely used loss function

$$\Delta_1(u)=E\{tr\{\Sigma^{-1}(u)\hat{\Sigma}(u)\}-\log|\Sigma^{-1}(u)\hat{\Sigma}(u)|\}-p, \tag{2.6}$$

$$\Delta_2(u)=E\Big[tr\{(\hat{\Sigma}(u)\Sigma^{-1}(u)-I)^2\}\Big], \tag{2.7}$$

where $tr(A)$ denotes the trace of an arbitrary matrix A. The losses Δ_1 and Δ_2 are known as the Stein loss and the quadratic loss, respectively. See, for example, Muirhead (1982).

Theorem 2. *Under* (C1)–(C6),

$$\Delta_1(u)=0.5\Delta_2(u)\{1+o(1)\}, \tag{2.8}$$

$$\Delta_2(u)=\frac{h^4}{4}\{\mu_2\}^2C_1(u)+\frac{1}{nhf(u)}\nu_0\Big[\Phi(u,u)-p\Big]+o\Big(h^4+(nh)^{-1}\Big), \tag{2.9}$$

where $C_1(u)=tr\{[\ddot{\Sigma}(u)\Sigma^{-1}(u)+2\dot{f}(u)/f(u)\dot{\Sigma}(u)\Sigma^{-1}(u)]^2\}$ *and* $\Phi(U_i,u)=E_{X|U}\{[(X_i-m(U_i))^{\top}\Sigma^{-1}(u)(X_i-m(U_i))]^2\}$.

The proof of Theorem 2 is given in Appendix C of the Technical Appendix. By (2.8), we know that using the Stein loss or the quadratic loss is not crucial because they are asymptotically equivalent to each other by just a constant and some negligible term. By (2.9), we know that the quadratic loss defined above has two different components. One is due to the variance having the order of $(nh)^{-1}$ and the other one is due to the bias having the order of h^4. Once again the optimal nonparametric convergence rate of $n^{-2/5}$ can be achieved by setting $h \propto n^{-1/5}$.

2.3. The bandwidths selection

Bandwidth selection for the kernel regression estimator and the local linear estimator have been well studied. One may directly use the existing results to select a bandwidth h_{1j} for $\hat{m}_j(\cdot)$ in (2.4). As to selection of bandwidth for $\hat{\Sigma}(u)$, one may develop a plug-in bandwidth based on the quadratic loss or the Stein loss as the plug-in bandwidth was for the kernel estimator regression function; see Chapter 4 of Fan and Gijbels (1996). For simplicity, we consider the log likelihood type leave-out-one criterion

$$\mathrm{CV}_\Sigma(h) = \frac{1}{n}\sum_{i=1}^{n}\left[\Big\{X_i - \hat{m}(U_i)\Big\}^\top \hat{\Sigma}^{-1}_{(-i)}(U_i)\Big\{X_i - \hat{m}(U_i)\Big\} + \log\Big(|\hat{\Sigma}_{(-i)}(U_i)|\Big)\right],$$

where $\hat{\Sigma}_{(-i)}$ is the estimate computed according to (2.3) but without the ith observation. Hereafter, the optimal bandwidth for the conditional covariance estimation can be determined by minimizing $\mathrm{CV}_\Sigma(h)$.

3. Numerical Example

Monte Carlo simulation was conducted to assess the finite sample performance of the proposed estimate. Our simulation results are summarized in Appendix D of the Technical Appendix and are consistent with the theoretical results of last section. In this section, we focus on the Boston Housing dataset that contains a total of 506 observations. The Gaussian kernel function is used for both our Monte Carlo simulation study and the Boston Housing data analysis.

For illustration purposes, we consider four social economics variables: CRIM (crime rate by town), TAX (full-value property-tax rate), PTRATIO (pupil-teacher ratio by town), and MEDV (median value of owner-occupied homes). Fan and Huang (2005) used $\sqrt{\mathrm{LSTAT}}$ as the *index* variable U, where LSTAT denotes the percentage of lower status of the population. Thus, we are able to study how the correlation structure of those X-variables changes as the percentage of lower status varies. Nevertheless, if we use the same U-variable, the second term of (2.5) suggests added bias in the local estimation. To reduce possible bias, we

Table 3.1. Sample Correlation Coefficients.

	CRIM	TAX	PTRATIO	MEDV
CRIM	1	0.5828	0.2899	-0.3883
TAX		1	0.4609	-0.4685
PTRATIO			1	-0.5078
MEDV				1

transform U so that it follows a uniform distribution over [0,1]. To this end, let U be the rank of LSTAT divided by the total sample size. Then U is uniformly distributed over [0,1]. Thus, the second term in (2.5), $2\dot{\sigma}_{j_1j_2}(u)\dot{f}(u)/f(u)$, is 0. Lastly, all X-variables are standardized to have zero mean and unit variance.

To confirm that the covariance matrix $\Sigma(u)$ is indeed varying according to u, we randomly split the dataset into two parts with equal sample sizes. Those two subsamples serve, respectively, as the training data and the testing data. Based on the training data, both $m(u)$ and $\Sigma(u)$ can be estimated. Denote these estimates by $\hat{m}(u)$ and $\hat{\Sigma}(u)$. We evaluate forecasting error by the out-of-sample loss measure

$$\Delta_{\text{out}} = \frac{1}{n^*}\sum_{i=1}^{n^*}\left[\Big\{X_i^* - \hat{m}(U_i^*)\Big\}^\top \hat{\Sigma}^{-1}(U_i^*)\Big\{X_i^* - \hat{m}(U_i^*)\Big\} + \log\Big(|\hat{\Sigma}(U_i^*)|\Big)\right],$$

where (X_i^*, U_i^*), $1 \le i \le n^*$, stands for the testing data. Intuitively, if both $m(u)$ and $\Sigma(u)$ are estimated accurately, we then should expect a reasonably good out-of-sample fit by treating those estimates as if they were the parameters, which in turn should produce reasonably small values for the negative log-likelihood loss. For a reliable evaluation, we replicated such an experiment a total of 200 times, which gave a median $\Delta_{\text{out}} = 0.5557$. To examine whether the conditional covariance $\Sigma(u)$ truly varies according to u, we replicated the same experiment but with $\hat{\Sigma}(U)$ in (2.3) replaced by $\tilde{\Sigma} = n^{-1}\sum_{i=1}^{n}\{X_i - \hat{m}(U_i)\}\{X_i - \hat{m}(U_i)\}^\top$ (i.e., a non-varying constant covariance). Then the resulting median Δ_{out} was 0.8575, much greater than 0.5557 of the nonparametric covariance model. We conclude that the covariance structure of those four social economics variables varies as the proportion of the lower status population (i.e., U) changes.

Next, we obtain the covariance matrix $\hat{\Sigma}(u)$ based on the whole dataset from which the pairwise nonparametric correlation coefficients, denoted by $\hat{\rho}_{j_1j_2}(u)$, were computed. They are reported in Figure 3.1. Figure 3.1 also gives the pointwise 90% confident intervals $\hat{\rho}_{j_1j_2}(u) \pm 1.64\widehat{\text{SE}}\{\hat{\rho}_{j_1j_2}(u)\}$, where the standard error estimate $\widehat{\text{SE}}\{\hat{\rho}_{j_1j_2}(u)\}$ was obtained based on 200 bootstrap experiments. For comparison purposes, the sample correlation coefficients are also presented in Table 1. Comparing the results in Figure 3.1 and Table 3.1, a number of findings can be obtained. For example, from Table 3.1, we have that sample correlation

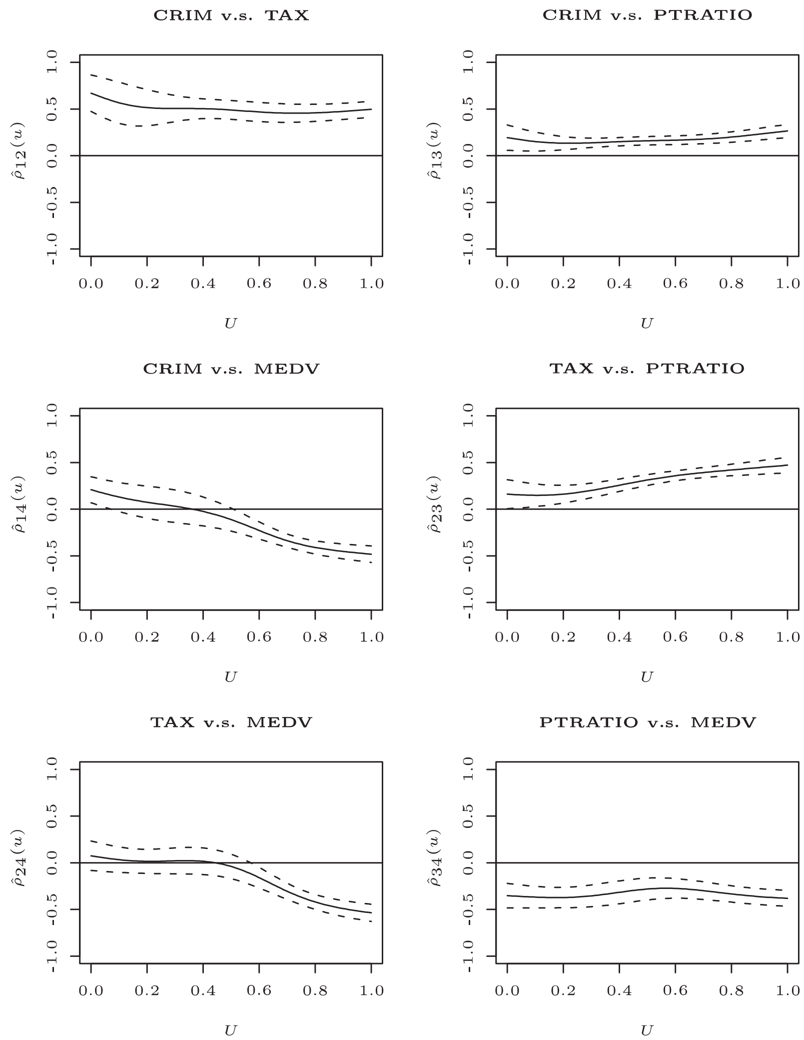

Figure 3.1. The Estimated Correlation Coefficients for the Boston Housing Data.

coefficient between the crime rate (CRIM) and the housing value (MEDV) is negative. On the other hand, by Figure 3.1, we further see the correlation coefficient has a decreasing trend as lower status increases, see the middle left panel. Furthermore, from the bottom-left panel, we find that the correlation between tax rate (TAX) and housing value (MEDV) has a decreasing trend as the lower status increases. We remark that these findings are not available from the simple sample correlation coefficient matrix as presented in Table 3.1.

Remark 3. One referee questions how the estimation bias might affect our conclusions. First, by Theorems 1 and 2, we know that as long as the bandwidth is selected appropriately, the magnitude of bias should be comparable to that of the standard error $\widehat{\text{SE}}\{\hat{\rho}_{24}(u)\}$, whose average is about 0.0707. Thus, we expect that an overall decreasing trend, as exhibited by $\hat{\rho}_{24}(0) - \hat{\rho}_{24}(1) = 0.6111$ for example, should dominate potential estimation bias. Thus, we are reasonably confident that an overall decreasing trend holds for $\hat{\rho}_{24}(u)$, even after eliminating the potential estimation bias. Similar comments apply to $\hat{\rho}_{14}$, the estimated correlation coefficient between CRIM and MEDV.

4. Concluding Remarks

It is of interest to estimate covariance and correlation functions for functional data and time series, see Yao, Müller and Wang (2005a,b), Fan, Huang and Li (2007), and Li et al. (2007). Estimating nonparametric covariance and correlation functions for functional data, longitudinal data, and multivariate time series is an interesting topic for future research.

Cholesky decomposition has been used in estimation of covariance matrices and covariance functions (Dai and Guo (2004), Huang et al. (2006), and Rosen and Stoffer (2007)). Estimation procedures based on Cholesky decomposition are particularly useful when there is a natural order among the variables of interest. It would be interesting to extend Cholesky decomposition techniques to estimation of the nonparamatric conditional covariance matrix. It is worth pointing out that different X-variable order might lead to a different Cholesky decomposition based estimate with a finite sample size. Thus, it is important to resolve issues related to bandwidth flexibility, positive definiteness, and permutation invariance (Rothman, Bickel, Levina and Zhu (2007)) under a unified framework. Further studies along this line are needed.

Appendix

The technical proofs of Theorems 1, 2, and a detailed simulation study are given in the on-line supplement material (i.e., the Technical Appendix) available at `http://www.stat.sinica.edu/statistica`. Also see Yin, Geng, Li and Wang (2008).

Acknowledgement

The authors are very grateful to the Editor, an Associate Editor, and two referees for helpful comments and constructive advice, which lead to a substantially improved manuscript. Jianxin Yin and Zhi Geng's researches are supported in part by NSFC (10771007) and NSFC (10721403). Jianxin Yin and Hansheng Wang's researches are supported in part by NSFC (10771006). Runze Li's research was supported mainly by National Institute on Drug Abuse, NIH, 1 R21 DA024260, and partially by National Science Foundation grant DMS-0348869.

References

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

Bilmes, J. A. (2000). Factorized sparse inverse covariance matrices. *IEEE International Conference on Acoustics, Speech, and Signal Processing.*

Cheng, M. Y., Fan, J. and Marron, J. S. (1997). On automatic boundary corrections. *Ann. Statist.* **25**, 1691-1708.

Dai, M. and Guo, W. (2004). Multivariate spectral analysis using Cholesky decomposition. *Biometrika* **91**, 629-643.

Diggle, P. and Verbyla, A. (1998). Nonparametric estimation of covariance structure in longitudinal data. *Biometrics* **54**, 401-415.

Drton, M. and Perlman, M. (2004). Model selection for gaussian concentration graphs. *Biometrika* **91**, 591-602.

Edwards, D. M. (2000). *Introduction to Graphical Modeling.* Springer, New York.

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.

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**, 645-660.

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

Lam, C. and Fan, J. (2007). Sparsistency and rates of covariance in large covariance matrices estimation. Manuscript.

Ledoit, O. and Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. *J. Portfolio Management* **4**, 110-119.

Levia, E., Rothman, A. J. and Zhu, J. (2007). Sparse estimation of large covariance matrices via a nested lasso penalty. *Ann. Appl. Stat.* **2**, 245-263.

Li, Q. (2006). *Nonparametric Econometrics: Theory and Practice.* Princeton University Press, Princeton, New Jersey.

Li, Y., Wang, N., Hong, M., Turner, N. D., Lupton, J. R. and Carroll, R. J. (2007). Nonparametric estimation of correlation functions in longitudinal and spatial data, with application to colon carcinogenesis experiments. *Ann. Statist.* **35**, 1608-1642.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34**, 1436-1462.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theorey.* Wiley, New York.

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. (2007). Sparse permutation invariant covariance estimation. Manuscript, Department of Statistics, University of Michigan.

Ruppert, D., Wand, M., Holst, U., and Hössjer, O. (1997). Local polynomial variance function estimation. *Technometrics* **39**, 262-273.

Smith, M. and Kohn, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *J. Amer. Statist. Assoc.* **97**, 1141-1153.

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

Yao, F., Müller, H. G. and Wang, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100**, 57-590.

Yao, F., Müller, H. G. and Wang, J.-L. (2005b). Functional regression analysis for longitudinal data. *Ann. Statist.* **33**, 2873-2903.

Yao, Q. and Tong, H. (1996). Asymmetric least squares regression estimation: a nonparametric approach. *J. Nonparametr. Stat.* **6**, 273-292.

Yao, Q. and Tong, H. (1998). Cross-validatory bandwidth selections for regression estimation based on dependent data. *J. Statist. Plann. Inference* **68**, 387-415.

Yin, J., Geng, Z., Li, R. and Wang, H. (2008). Nonparametric covariance model. Technical Report 08-90. The Methodology Center, The Pennsylvania State University, University Park, Pennsylvania.

School of Mathematical Sciences, Peking University, Beijing, P. R. China, 100871.

E-mail: jianxinyin@gmail.com

School of Mathematical Sciences, Peking University, Beijing, P. R. China, 100871.

E-mail: zgeng@math.pku.edu.cn

Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, Pennsylvania, 16802 USA.

E-mail: rli@stat.psu.edu

Guanghua School of Management, Peking University, Beijing, P. R. China, 100871.

E-mail: hansheng@gsm.pku.edu.cn

(Received January 2008; accepted June 2008)