Nonparametric Covariance Estimation for Longitudinal Data via Penalized Tensor Product Splines

Tayler A. Blake[*]   Yoonkyung Lee[†]

January 9, 2018

Abstract

With high dimensional longitudinal and functional data becoming much more common, there is a strong need for methods of estimating large covariance matrices. Estimation is made difficult by the instability of sample covariance matrices in high dimensions and a positive-definite constraint we desire to impose on estimates. A Cholesky decomposition of the covariance matrix allows for parameter estimation via unconstrained optimization as well as a statistically meaningful interpretation of the parameter estimates. Regularization improves stability of covariance estimates in high dimensions, as well as in the case where functional data are sparse and individual curves are sampled at different and possibly unequally spaced time points. By viewing the entries of the covariance matrix as the evaluation of a continuous bivariate function at the pairs of observed time points, we treat covariance estimation as bivariate smoothing.

Within regularization framework, we propose novel covariance penalties which are designed to yield natural null models presented in the literature for stationarity or short-term dependence. These penalties are expressed in terms of variation in continuous time lag and its orthogonal complement. We present numerical results and data analysis to illustrate the utility of the proposed method.

keywords: non-parametric, covariance, longitudinal data, functional data, splines, reproducing kernel Hilbert space

1   Introduction

An estimate of the covariance matrix or its inverse is required for nearly all statistical procedures in classical multivariate data analysis, time series analysis, spatial statistics and, more recently, the growing field of statistical learning. Covariance estimates play a critical role in the

---

[*]The Ohio State University, 1958 Neil Avenue, Columbus, OH 43201
[†]The Ohio State University, 1958 Neil Avenue, Columbus, OH 43201

1

performance of techniques for clustering and classification such as linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), factor analysis, and principal components analysis (PCA), analysis of conditional independence through graphical models, classical multivariate regression, prediction, and Kriging. Covariance estimation with high dimensional data has has recently gained growing interest; it is generally recognized that there are two primary hurdles responsible for the difficulty in covariance estimation: the instability of sample covariance matrices in high dimensions and a positive-definite constraint we wish estimates to obey.

Prevalent technological advances in industry and many areas of science make high dimensional longitudinal and functional data a common occurrence, arising in numerous areas including medicine, public health, biology, and environmental science with specific applications including fMRI, spectroscopic imaging, gene microarrays among many others, presenting a need for effective covariance estimation in the challenging situation where parameter dimensionality p is possibly much larger than the number of observations, n.

We consider two types of potentially high dimensional data: the first is the case of functional data or times series data, where each observation corresponds to a curve sampled densely at a fine grid of time points; in this case, it is typical that the number of time points is larger than the number of observations. The second is the case of sparse longitudinal data where measurement times may be almost unique yet sparsely distributed within the observed time range for each individual in the study. In this case, the nature of the high dimensionality may not be a consequence of having more measurements per subject than the number of subjects themselves, but rather because when pooled across subjects, the total number of unique observed time points is greater than the number of individuals. Several approaches have been taken in effort to overcome the issue of high dimensionality in covariance estimation. Regularization improves stability of covariance estimates in high dimensions, particularly in the case where the parameter dimensionality p is much larger than the number of observations n. Regularization of the covariance matrix and its Cholesky decomposition has been explored extensively through various approaches including banding, tapering, kernel smoothing, penalized likelihood, and penalized regression; see citetpourahmadi2011covariance for a comprehensive overview.

To overcome the hurdle of enforcing covariance estimates to be positive definite, several have considered modeling various matrix decompositions including variance-correlation decomposition, spectral decomposition, and Cholesky decomposition. The Cholesky decomposition has received particular attention, as it which allows for a statistically meaningful interpretation as well as an unconstrained parameterization of elements of the covariance matrix. This parameterization allows for estimation to be accomplished as simply as in least squares regression. If we assume that the data follow an autoregressive process with (possibly) heteroskedastic errors, then the two matrices comprising the Cholesky decomposition, the Cholesky factor (which diagonalizes the covariance matrix) and diagonal matrix itself, hold the autoregressive coefficients and the error variances, respectively.

In longitudinal studies, the measurement schedule could consist of targeted time points or could consist of completely arbitrary (random) time points. If either the measurement schedule

has targeted time points which are not necessarily equally spaced or if there is missing data, then we have what is considered incomplete and unbalanced data. If the measurement schedule has arbitrary or almost unique time points for every individual so that at a given time point there could be very few or even only a single measurement, we must consider how to handle what we consider as sparse longitudinal data. We view the response as a stochastic process with corresponding continuous covariance function and the generalized autoregressive parameters as the evaluation of a continuous bivariate function at the pairs of observed time points rather than specifying a finite set of observations to be multivariate normal and estimating the covariance matrix. This is advantageous because it is unlikely that we are only interested in the covariance between pairs of observed design points, so it is reasonable to approach covariance estimation in a way that allows us to obtain an estimate of the covariance between two measurements at any pair of time points within the time interval of interest.

Through the Cholesky decomposition, we formulate covariance estimation as a penalized regression problem and propose novel covariance penalties designed to yield natural null models presented in the literature. By transforming the axes of the design points, we express these penalties in terms of two directions: the lag component and the additive component and characterize the solution coefficient function in terms of a functional ANOVA decomposition. Some have sidestepped the issue of high dimensionality by prescribing simple parametric models for the elements of the Cholesky decomposition. **?**, Pourahmadi [1999], and **?** have elicited stationary parametric models for the generalized autoregressive coefficients, letting the GARPs depend only on the distance between two time points. To induce the structural simplicity of such stationary models with the flexibility of a nonparametric approach, we penalize all functional components but that corresponding to the lag component so that the set of null models is comprised of stationary models. Huang et al. [2007] follow the hueristic argument presented in Pourahmadi [1999] that the generalized autoregressive parameters are monotone decreasing in as lag increases and set off-diagonal elements of either the covariance matrix or the Cholesky factor corresponding to large lags to zero. Rather than shrinking element of the Cholesky factor to zero after particular value of l, we choose to softly enforce monotonicity in l by penalizing order restriction as in the work of Tibshirani et al. [2011].

The rest of the paper is organized as follows: Section 2 summarizes the general penalized estimation approach and introduces the transformed design coordinates and penalties for stationarity and non-monotonicity. Section 3 presents a detailed discussion of optimization and computational issues. Section 4 presents a simulation study and a real example to examine the performance of our methods as well as others. Section 5 concludes with discussion and future work.

2 Covariance estimation: a review

Parametric models are frequently used for modeling covariance structure in the longitudinal data setting. Simple models which depend on a small number of parameters are commonly found in the literature such as those corresponding to compound symmetry and autoregressive models of order

k, where k is small. However, model misspecification can lead to considerably biased estimates. Alternately, several have proposed applying nonparametric methods directly to elements of the sample covariance matrix or a function of the sample covariance matrix. Diggle and Verbyla (1998) introduced a nonparametric estimator obtained by kernel smoothing the sample variogram and squared residuals. Yao, Mueller, and Wang applied a local linear smoother to the sample covariance matrix in the direction of the diagonal and a local quadratic smoother in the direction orthogonal to the diagonal to account for the presence of additional variation due to measurement error. [REVIEW 2009 WU AND POURAHMADI METHOD: banding the sample covariance matrix. Under the assumption of short range dependency, they show that their estimator converges to the true covariance matrix for a broad class of nonlinear processes.] The estimates yielded by these approaches, however, are not guaranteed to be positive definite.

To satisfy the positive-definiteness constraint, methods have been developed and applied to certain reparameterizations of the covariance structure. Chiu, Leonard, and Tsui modeled the matrix logarithm of the covariance matrix. Early nonparametric work using the spectral decomposition of the covariance matrix included that of Rice and Silverman (1991) which discussed smoothing and smoothing parameter choice for eigenfunction estimation for regularly-spaced data. Staniswalis and Lee (1998) extended kernel-based smoothing of eigenfunctions to functional data observed on irregular grids. However, when the data are sparse in the sense that there are few repeated within-subject measurements and measurement times are quite different from subject-to-subject, approximation of the functional principal component scores defined by the Karhunen-Loeve expansion of the stochastic process by usual integration is unsatisfactory and requires numerical quadrature. Many have explored regression-based approaches using the Spectral decomposition, framing principal components analysis as a least-squares optimization problem. Among many others, Zou, Hastie and Tibshirani (2006) imposed penalties on regression coefficients to induce sparse loadings. [REVIEW THE METHODS OF HUANG, KAUFMAN, YAO HERE]

We adopt the approach based on the Cholesky decomposition. The modified Cholesky decomposition (MCD) has received much attention in the covariance estimation literature, as it ensures positive-definite covariance estimates, and, unlike the spectral decomposition whose parameters follow an orthogonality constraint, the Cholesky decomposition are unconstrained and have an attractive statistical interpretation as particular regression coefficients and variances. The covariance matrix Σ of a zero-mean random vector $Y = (y_1, \ldots, y_m)'$ has the following unique modified Cholesky decomposition (Newton, 1988)

$$T\Sigma T' = D,$$

where T is a lower triangular matrix with 1?s as its diagonal entries and $D = \text{diag}\left(\sigma_1^2, \ldots, \sigma_m^2\right)$ is a diagonal matrix. An attractive feature of this decomposition is that unlike the entries of Σ, the subdiagonal entries of T and the log of the diagonal elements of D, $\log\left(\sigma_m^2\right)$, $t = 1, \ldots, m$, are not constrained. This permits one to impose structures on the unconstrained parameters without worrying about the resulting estimator not satisfying the positive-definiteness constraint. Denote estimators of T and D in 1 by $\hat{T}$ and $\hat{D}$, which may be obtained by fitting linear models or some other structural models; then an estimator of Σ given by $\Sigma = \hat{T}^{-T}\hat{D}\hat{T}^{-T}$ is guaranteed to be

positive-definite. From this perspective, covariance modeling can be considered an extension of generalized linear models **?**. Factoring Σ as in 1 provides a link function $g(\Sigma) = (T, \log(D))$ where $\log(D) = \text{diag}(\log(\sigma_1^2), \ldots, \log(\sigma_m^2))$. Parametric, nonparametric, or Bayesian models may then be applied to the unconstrained entries of T and $\log(D)$. Whereas other decompositions are permutation-invariant, the interpretation of the regression model induced by the MCD assumes a natural (time) ordering among the variables in Y.

Several others have considered this approach to covariance estimation; Kaufman et al. [2008] assume a stationary process, restricting covariance estimates to a specific class of functions. They as well as Huang, Liu, and Liu Huang et al. [2007] follow the hueristic argument presented by Pourahmadi [1999] that $\phi_{t,t-l}$ is monotone decreasing in l and set off-diagonal elements of either the covariance matrix or the Cholesky factor corresponding to large lags to zero. As in Huang et al. [2007], Kaufman et al. [2008], and Yao et al. [2005], we treat covariance estimation as a function estimation problem where the covariance matrix is viewed as the evaluation of a smooth function at particular design points.
including Bickel and Levina [2008] and Huang et al. [2006] have proposed nonparametric estimators of a specific covariance matrix (or its inverse) rather than the parameters of a covariance function.

Yao et al. [2005] do not utilize the Cholesky parameterization, and their estimates are not guaranteed to be positive definite. We combine the advantages of bivariate smoothing as in Yao et al. [2005] with the added utility of the Cholesky parameterization in Huang et al. [2007]; in doing so, we present a flexible and coherent approach to covariance estimation, while simultaneously we ensuring positive definiteness of estimates.Rather than shrinking element of the Cholesky factor to zero after a particular value of l, we choose to softly enforce monotonicity in l by using a hinge penalty as in the work of Tibshirani et al. [2011].

3   The Cholesky Decomposition and the MLE for Σ

Let $Y = (y_1, y_2, \ldots, y_m)'$ denote a mean zero random vector with variance-covariance matrix Σ, which we can think of as the time-ordered measurements on one subject in a longitudinal study. To present a comprehensive overview our estimation procedure, we begin with the representation of the covariance matrix, Σ, in terms of its Cholesky decomposition. Decomposing Σ in such a way allows for both an unconstrained parameterization and statistically meaningful interpretation of covariance parameters. For any positive definite matrix Σ, there exists a unique lower triangular matrix T with diagonal entries equal to 1 which diagonalizes Σ:

$$T\Sigma T^T = D \tag{1}$$

The convenient statistical interpretation of the parameters of the covariance matrix then comes if we consider, for $t = 2, \ldots, m$, regressing y_t on its predecessors $y_1, \ldots, y_{t-1}$, letting

$$y_i = \sum_{j=1}^{i-1} \phi_{ij} y_j + \sigma_i \epsilon_i, \tag{2}$$

5

where var $(\epsilon_i) = \sigma_i^2$. If we take the i-j^{th} element T to be $-\phi_{ij}$ for $j < i$, and take the ith diagonal entry of D to be var $(\epsilon_i) = \sigma_i^2$, a vectorized expression for Model 2 is given by

$$\epsilon = TY. \tag{3}$$

and taking covariances on both sides of (??), we see that T and D satisfy 1. Immediately, we have that $\Sigma^{-1} = T'D^{-1}T$. The regression coefficients $\{\phi_{ij}\}$ are referred to as the *generalized autoregressive parameters* (GARPs), and the $\{\sigma_{ij}\}$ are referred to as the *innovation variances* (IVs.) Assuming that Y follows a multivariate normal distribution, the loglikelihood function $\ell(Y, \Sigma)$ satisfies

$$-2\ell(Y, \Sigma) = \log|\Sigma| + Y'\Sigma Y \tag{4}$$

From 1, we have that

$$|\Sigma| = |D| = \prod_{i=1}^{m} \sigma_i^2$$

and

$$\Sigma^{-1} = T'D^{-1}T.$$

Thus, 4 can be written in terms of the prediction errors and their variances of the non-redundant entries of (T, D):

$$-2\ell(Y, \Sigma) = \log|D| + Y'T'D^{-1}TY$$
$$= \sum_{i=1}^{m} \log\sigma_i^2 + \sum_{i=1}^{m} \frac{\epsilon_i^2}{\sigma_i^2}, \tag{5}$$

where $\epsilon_1 = y_1$ and $\epsilon_i = y_i - \sum_{j=1}^{i-1}\phi_{ij}y_j$. Maximum likelihood estimation or any of its penalized variants may then be employed to obtain estimates of T and D.

Unlike many of those before who have used the Cholesky decomposition as a means of modeling Σ, we allow observed time points to be individual-specific and not necessarily regularly spaced. Let $Y_1, \ldots, Y_N$ denote a random sample of mean zero vectors of longitudinal measurements taken on N subjects having common covariance structure Σ. We allow subject i to have observation vector $y_i = (y_{i1}, \ldots, y_{i,m_i})'$ with corresponding vector of observation times $(t_{i1}, \ldots, t_{i,m_i})'$. Accommodating the subject-specific sample sizes and measurement times requires merely adding a subscript, and Model 2 becomes

$$y_{ij} = \sum_{k=1}^{j-1} \phi_{ijk}y_{ik} + \sigma_{ij}\epsilon_{ij}, \tag{6}$$

where ϕ_{ijk} is the autoregressive coefficient corresponding to the pair of measurements observed at time t_{ij} and t_{ik}. A vectorized representation of Model 6 can be obtained as before by adding the necessary parameters to T and D.

6

Removing the restriction that subjects having common covariance structure also share common, equally-spaced observation times encourages a functional representation of the parameters defining the Cholesky decomposition of Σ. Rather than a vector of longitudinal data points, we view the random vectors Y_i and $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{i,m-i})'$ as discrete renditions of the stochastic processes: $Y(t)$ and $\epsilon(t)$. We assume $Y(t)$ has covariance function $G(t, s)$ and that $\epsilon(t)$ follows a zero mean Gaussian white noise process with unit variance. Under mild assumptions regarding the behaviour of Y, then $G(t, s)$ satisfies some smoothness conditions, where smoothness is defined in terms of square integrability of certain derivatives. We view the entries of Σ as values of G evaluated at the distinct pairs of within-subject observed time points. If we consider the Cholesky

decomposition of Σ within such functional context, it is natural to extent the same notion to the elements of T and D. We view the GARPs $\{\phi_{ijk}\}$ and innovation variances as the evaluation of the smooth functions $\phi(t, s)$ and $\sigma^2(t)$ at observed time points, where

$$\phi_{ijk} = \phi(t_{ij}, t_{ik}), \quad t_{ik} < t_{ij}$$

and

$$\sigma_{ij}^2 = \sigma^2(t_{ij}).$$

With this definition of ϕ and σ^2, Model 6 becomes

$$y(t_{ij}) = \sum_{k=1}^{j-1} \phi(t_{ij}, t_{ik}) y(t_{ik}) + \sigma(t_{ij}) \epsilon(t_i) \quad i = 1, \ldots, N, \tag{7}$$

Our formulation transforms the task of estimating a covariance matrix is to estimating the function $\phi(t, s)$ (and $\sigma(t)$) using bivariate (univariate) smoothing. Like other nonparametric situations, we make no assumption about the functional form of ϕ other than that ϕ is smooth, with smoothness defined in terms of square integrability of certain derivatives. Our proposed method for covariance estimation defines a flexible, general framework which makes all of the existing techniques for penalized regression accessible for the seemingly far different task of estimating a covariance matrix.

To add to the interpretability of our proposed models, we propose a rotation of the (t, s) axes:

$$l = t - s$$
$$m = \frac{t + s}{2}$$

The l axis, the "lag" between time points t and s is of special interest within the context of covariance modeling. The class of isotropic covariance functions M is isotropic are defined as those functions $G(t, s) = G(||t - s||)$ that depend on time points t and s only through the Euclidean distance $||t - s||$ between them. Isotropic covariance functions obey some notion of parsimony,

so it is convenient to apply regularized estimation to ϕ in terms of the transformed points (l, m). Along with this class of functions, others have proposed parsimonious models for the elements of the Cholesky decomposition in terms of the difference between Re-expressing ϕ in terms of these new arguments, our goal is to estimate

$$\phi^*(l, m) = \phi^*\left(s - t, \frac{1}{2}(s + t)\right) = \phi(s, t) \tag{8}$$

Rather than imposing structure on the unconstrained values of $\phi(s, t)$, we instead consider a rotation of the axes of the input space and re-express ϕ in terms of the transformed coordinates $\begin{bmatrix} l \\ m \end{bmatrix} = \begin{bmatrix} s - t \\ \frac{(s+t)}{2} \end{bmatrix}$

$$\begin{aligned} \phi(s, t) &= \phi^*\left(s - t, \frac{s + t}{2}\right) \\ &= \phi^*(l, m) \end{aligned}$$

While our framework allows for estimation of the autoregressive coefficient function and the innovation variance function via any nonparametric regression setup, we focus on two primary approaches for representing ϕ and σ. First, we assume that ϕ belongs to a reproducing kernel Hilbert space, $\mathcal{H}$ and employ the smoothing spline methods of Kimeldorf and Wahba (see Kimeldorf and Wahba [1971] and Wahba [1990] for comprehensive presentation.) To enhance the statistical interpretability of model parameters, we decompose ϕ into functional components similar to the notion of the main effect and the interaction terms in classical analysis of variance. We adopt the smoothing spline analogue of the classical ANOVA model proposed by Gu (see Gu [2013]), and estimation is achieved through similar computational strategies.

References

Peter J Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227, 2008.

Chong Gu. *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media, 2013.

Jianhua Z Huang, Naiping Liu, Mohsen Pourahmadi, and Linxu Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, pages 85–98, 2006.

Jianhua Z Huang, Linxu Liu, and Naiping Liu. Estimation of large covariance matrices of longitudinal data with basis function approximations. *Journal of Computational and Graphical Statistics*, 16(1):189–209, 2007.

Cari G Kaufman, Mark J Schervish, and Douglas W Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103 (484):1545–1555, 2008.

George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

Mohsen Pourahmadi. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690, 1999.

Ryan J Tibshirani, Holger Hoefling, and Robert Tibshirani. Nearly-isotonic regression. *Technometrics*, 53(1):54–61, 2011.

Grace Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.