Nonparametric Covariance Estimation for Longitudinal Data via Penalized Tensor Product Splines

Tayler A. Blake* Yoonkyung Lee†

May 24, 2017

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 performance of techniques for clustering and classification such as linear discriminant analysis (LDA), quadratic

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

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 is possibly much larger than the number of observations. Additional difficulty due to constraints required to yield positive definite estimates make covariance estimation a potentially complex optimization problem. Further, most existing approaches to covariance estimation require data to be sampled at regular grid (time) points, with subjects sharing a set of common observation points. However, in many practical situations, data are irregularly sampled, and subjects may share few common observation times, and methods are needed to accommodate for data collected in this way.

To address the challenge of enforcing positive definiteness, 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.

[ADD citation for the first to introduce using the modified cholesky decomposition]

It is well known that the sample covariance matrix is unstable in high dimensions, and there is an extensive existing body of work addressing the issue of high dimensionality in the context of covariance estimation. [cite the pourahmadi survey paper here.] However, much of this work addresses high dimensionality arising from functional or times series data sampled on a dense, regular grid. With such data, it is typical that the number of time points is larger than the number of observations. Few have addressed the challenges posed by 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, 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. Incomplete and unbalanced data arise when measurement schedules with targeted time points which are not necessarily equally spaced or if there is missing data. Sparse longitudinal data arise when the measurement schedule has arbitrary or almost unique time points for every individual. A given time point may have very few individuals with corresponding measurements.

We sidestep both issues of high dimensionality and irregularly sampled data by viewing the response as a stochastic process having continuous covariance function. Recent work outlines the use of function estimation for smoothing elements of the covariance matrix, including [cite all the smoothing papers here.] To our knowledge, however, no previous work has applied bivari-

ate smoothing both dimensions of the Cholesky factor. We model the generalized autoregressive parameters using tensor product splines. Viewing covariance modeling as bivariate function estimation both accommodates irregularly sampled curved and permits interpolation and extrapolation of the covariance function between two measurements at any pair of time points within the time interval of interest rather than at observed pairs of time points only. Through the Cholesky decomposition, we carry out estimation by the estimation of varying coefficient model. A transformation of the design point axes allows for an ANOVA-like decomposition of the coefficient function into two components, corresponding to the lag between time points and an additive component. Through this general framework, we can easily impose penalties on fitted functions to yield natural null models presented in the literature.

2 Cholesky Decomposition of Σ

To present a comprehensive overview our estimation procedure, we begin with the representation of the inverse covariance matrix, Σ^{-1}, in terms of its Cholesky decomposition (see citetpourahmadi2007cholesky for a detailed discussion.) In the section to follow, we will demonstrate that this parameterization of the precision matrix is particularly attractive due to both the computational advantages as well as the convenient modeling interpretation it permits. For any positive definite matrix Σ, there exists a unique unit lower triangular matrix T with diagonal entries equal to 1 which diagonalizes Σ:

$$T\Sigma T^T = D$$

If we assume that the data having covariance matrix Σ follow an autoregressive model, then the entries of the Cholesky factor T and D enjoy a useful interpretation. Let $Y = (y_1, y_2, \ldots, y_m)^T$ denote a mean-zero random vector of observations with corresponding measurement times

$$t_1 < t_2 < \cdots < t_m.$$

Consider regressing y_j on its predecessors:

$$y_j = \sum_{k=1}^{j-1} \phi_{jk} y_k + \sigma_j e_j, \qquad j = 2, \ldots, m, \tag{1}$$

where we define $y_1 = e_1$. Standard regression theory gives us that if $\{\phi_{jk}\}$ are the coefficients of the linear least squares predictor of y_j based on its predecessors, then the residuals $e = (e_1, e_2, \ldots, e_m)^T$ have diagonal covariance. Let T denote the $m \times m$ matrix with elements

$$T_{jk} = \begin{cases} -\phi_{jk} & j > k \\ 1 & j = k \\ 0 & otherwise, \end{cases}$$

for $j, k = 1, \ldots, m$. Then in matrix notation, model 1 may then be written

$$e = TY, \tag{2}$$

Taking covariances on both sides of 2, we have

$$D = T\Sigma T^T$$

The regression coefficients $\{\phi_{jk}\}$, which are unconstrained, are referred to as the *generalized autoregressive parameters* (GARPs). The $\{\sigma_j\}$ are called the *innovation variances* (IVs.) Unconstrained estimation of the $\{\sigma_k^2\}$ is achieved by log transformation; we leave these details for section 2.) Expressing the precision matrix in terms of the GARPs and IVs, we have

$$\Omega = \Sigma^{-1} = T^T D^{-1} T. \tag{3}$$

To extend this estimation framework to accommodate observations on multiple subjects which may be taken at unequally spaced and individual-specific observation times. Rather than m-dimensional vectors, consider Y and e as the values of the stochastic processes $Y(t)$ and $e(t)$ at the set of observation times. We assume that $Y(t)$ is equipped with covariance function $G(s,t)$, and

$$e(s) \sim \mathcal{WN}(0,1)$$

is a zero mean Gaussian white noise process with unit variance. For a well-behaved process Y, we may assume that $G(s,t)$ satisfies some smoothness conditions, where smoothness is defined in terms of square integrability of certain derivatives. The entries of Σ, then, correspond to G evaluated at the distinct pairs of observed time points. Similarly, we treat the elements of the precision matrix Ω as the values of some smooth function, $\omega(s,t)$ evaluated at observed pairs of time points. Extention of this perspective to the elements of D and the elements of the Cholesky factor T leads us to the varying coefficient (VC) models first introduced by Hastie and Tibshirani. A generalization of traditional linear regression models, varying coefficient models offer more flexibility than their static analogues by allowing the effect of covariates to change smoothly with the value other variables. Both regressors and response variables are assumed to vary according to an *indexing variable*, which is particularly attractive because this permits interpolation of regressors and response variables at values of this indexing variable where there is either missing data of only a single observation and slope estimation is not feasible. Replacing $\{\phi_{jk}\}$ and $\{\sigma_j\}$ with smooth functions, we model

$$y(t_j) = \sum_{k=1}^{j-1} \phi(t_j, t_k)\, y(t_k) + \sigma(t_j)\, \epsilon(t_j) \qquad j = 1, \ldots, m, \tag{4}$$

for $t_1 < t_2 < \cdots < t_m$. Within our proposed framework, the task of estimating a covariance matrix is equivalent to estimating the function $\phi(s,t)$. We explore using both smoothing splines and B-spline expansions for function approximation; to induce smoothness and parsimony of estimated Cholesky For ease of exposition, we first assume that $\sigma^2(t)$ is fixed and known; we will later relax this assumption and discuss the estimation of σ^2 and ϕ simultaneously. We assume nothing about the functional form of ϕ other than that ϕ is smooth, with smoothness, which is defined in terms of

families of penalties which we will discuss in detail in sections to follow. Our approach presents a flexible, comprehensive framework for covariance estimation in a wide variety of contexts in which the data may be generated according to a broad class of dependency structures.

3 Penalized Maximum Likelihood Estimation of ϕ

Along with citethuang2006covariance, citetlevina2008sparse, and citetpourahmadi2000maximum we employ the Gaussian log-likelihood as a goodness of fit measure for the varying autoregressive coefficient function, $\phi(t,s)$ and the innovation variance function $\sigma(t)$, though neither the derivation of model 1 nor model 4 rely on any assumptions about the distribution of e. Fixing σ_j^2, the negative loglikelihood as a function of ϕ_{jk} corresponds to the usual error sums of squares. Under the gaussian assumption, for fixed $\left\{\sigma_j^2\right\}$, the negative log-likelihood of N i.i.d. vectors of observations $y_1, y_2, \ldots, y_N$ is proportional to

$$-2L\left(y_1, y_2, \ldots, y_N, \Phi\right) \propto \sum_{i=1}^{N} \sum_{j=2}^{m_i} \sigma\left(t_j\right)^{-2} \left(y_{ij} - \sum_{k=1}^{j-1} \phi\left(t_{ij}, t_{ik}\right) y_{ik} \right)^2 \tag{5}$$

where

$$y_i = \left(y_{i1}, y_{i2}, \ldots, y_{i,m_i}\right), \quad i = 1, \ldots, N$$

denotes the vector of observations for subject i with corresponding measurement times

$$t_{i1} < t_{i2} < \cdots < t_{i,m_i}.$$

The form of the likelihood of $y_1, \ldots, y_N$ indicates that we allow both the number of measurements as well as the observation times to varying across subjects. The $\{t_{ij}\}$ need not be evenly-spaced within or across individuals. In the case that subjects share a common set of observation times $t_1 < \cdots < t_m$, it is well known that the MLE for Σ, $S = \sum_{i=1}^{N} y_i y_i^T$ is highly unstable in high dimensions. [cite those who have proposed mitigating this using penalized maximum likelihood for the distinct elements of T.] This condition is potentially worsened when one or more subjects has at least one unique observation time. To mitigate instability due to high dimensionality and simultaneously permit the estimation of $\phi(\cdot,\cdot)$ as a smooth bivariate function, we obtain a covariance estimator by applying bivariate smoothing of the elements of the Cholesky factor. We impose two families of penalties on fitted functions, leading to two distinct parameterizations of the smoothed function [blah blah something about flexibility and penalty being problem-specific.]

4 Representation of ϕ as a smooth function

To impose structure on the estimated autoregressive function, we append a penalty functional to the negative log-likelihood **??**. that discourages the flexibility of the fitted function. We take the estimator of ϕ to minimize

$$-2L + \lambda J_\phi\left(\phi\right). \tag{6}$$

The first term in 6 discourages the lack of fit of ϕ to the data, and λ is a smoothing parameter which controls the tradeoff between the lack of fit and amount of regularization imposed on the fitted function through the penalty, J_ϕ.The task of estimating of $\phi(t,s)$ is an inherently different problem than the estimation of an arbitrary bivariate function. Since ϕ explicitly defines an inverse covariance function, imposing specific types of structure on ϕ is of particular interest. Covariance models for longitudinal or time series data are commonly defined in terms of lag, or in the continuous case, the difference between two measurement times. By transforming the original $t-s$ axis according to

$$\begin{aligned} l &= s - t, \text{ and} \\ m &= \frac{1}{2}(s+t), \end{aligned}$$

we may parameterize the coefficient function in terms of components defined by functions of l and m. Writing phi in terms of the rotation gives the reparameterized coefficient function

$$\phi^*(l,m) = \phi^*\left(s-t, \frac{1}{2}(s+t)\right) = \phi(s,t). \tag{7}$$

We define our estimator $\hat{\phi^*}$ as the minimizer of

$$-2L + \lambda^* J_{\phi^*}(\phi^*). \tag{8}$$

4.1 Smoothing spline ANOVA models

We consider models that capture the marginal effects of l and m, as well as interaction between the two directions. We first consider the smoothing spline ANOVA decomposition of citetgu2002smoothing, modeling

$$\phi^*(l,m) = \mu^* + \phi_1^*(l) + \phi_2^*(m) + \phi_{12}^*(l,m). \tag{9}$$

As in citegu2002smoothing, citecraven1978smoothing,[more Wahba citations here], we consider functions ϕ^* belonging to a reproducing kernel Hilbert space (r.k.h.s.), $\mathcal{H}$. We equip l and m each with r.k.h.s., $\mathcal{H}_l$ and $\mathcal{H}_m$; the space of bivariate functions $\mathcal{H}$ can be constructed from the tensor product of the univariate function spaces for l and m:

$$\mathcal{H} = \mathcal{H}_l \otimes \mathcal{H}_m.$$

One choice for the marginal spaces is the second-order Sobolev space, letting $\mathcal{H}_l = \mathcal{H}_m = W_2(0,1)$, where

$$W_2(0,1) = \{f: \quad f, f' \text{absolutely continuous}, \int_0^1 \left(f^{(2)}\right)^2 dt < \infty\}.$$

4.2 Decomposition of $\mathcal{H}$ via J_{ϕ^*}

Several have proposed methods for applying regularization of Cholesky decomposition including banding, tapering, kernel smoothing, penalized likelihood, and penalized regression. See [ ] Within the function estimation paradigm, a number of approaches to estimate the coefficient function $\phi(\cdot,\cdot)$ have been proposed including See citewu2003nonparametric, citehuang2007estimation . Common techniques for inducing structure to produce simple and stable covariance estimates include shrinking estimated functions or the elements of the covariance matrix itself so that the resulting dependency structure corresponds to parsimonious covariance models frequently adopted in the time series and longitudinal data literature. [CITE PAPERS PROPOSING PARSIMONIOUS MODELS FOR phi ij] The ANOVA model in 9 allows us to easily specify penalties J that encourage estimates to adhere to the structure of these models. [cite some general time series/longitudinal sources] When ϕ^* corresponds to the simple models of the form (**??**), the bivariate function may be written in terms of only its first argument. . .

The penalty functional penalty functional J induces a decomposition of $\mathcal{H}$ as a direct sum of two subspaces:

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1,$$

where $\mathcal{H}_0$ denotes the null space of J, which is composed of functions such that $J(\phi^*) = 0$, and $\mathcal{H}_1$ is the subspace orthogonal to $\mathcal{H}_0$. Let $P_1\phi^*$ denote the projection of ϕ^* onto the penalized space $\mathcal{H}_1$. We can express J in terms of the projection of $\phi^* \in \mathcal{H}$ onto $\mathcal{H}_1$:

$$\begin{aligned} J(\phi) &= ||P_1\phi^*||^2 \\ &= ||P_1\phi_1^*||^2 + ||P_1\phi_2^*||^2 + ||P_1\phi_{12}^*||^2 \end{aligned} \tag{10}$$

The penalty functional J induces a decomposition of $\mathcal{H}$ as follows: $\mathcal{H}_l = \mathcal{H}_l^0 \oplus \mathcal{H}_l^1$ and $\mathcal{H}_m = \mathcal{H}_m^0 \oplus \mathcal{H}_m^1$ where let $\mathcal{H}_l^0 = \{1\} \oplus \{k_1\}$, $\mathcal{H}_m^0 = \{1\}$, and where $\{k_r\}$ denotes the subspace spanned by k_r. $\mathcal{H}_l^1$ and $\mathcal{H}_m^1$ are the subspaces orthogonal to $\mathcal{H}_l^0$ and $\mathcal{H}_m^0$, respectively:

$$\mathcal{H}_l^1 = \{\phi_1^* : \int_0^1 \phi_1^{*(\nu)}(l)\, dl = 0, \;\; \nu = 0, 1\}$$

$$\mathcal{H}_m^1 = \{\phi_2^* : \int_0^1 \phi_2^*(m)\, dm = 0\}$$

Using the properties of tensor product spaces, we may decompose $\mathcal{H} = \mathcal{H}^0 \oplus \mathcal{H}^1$ where

$$\begin{aligned} \mathcal{H}^0 &= \{1\} \oplus \{k_1\} \\ \mathcal{H}^1 &= \mathcal{H}_l^1 \oplus \mathcal{H}_m^1 \oplus \left[\{k_1\} \otimes \mathcal{H}_m^1\right] \oplus \left[\mathcal{H}_l^1 \otimes \mathcal{H}_m^1\right] \end{aligned}$$

To find the solution $\hat{\phi}^*$ which is the stage-wise minimizer of (**??**): we first set $\lambda_2 = 0$ and find $\tilde{\phi}^*$ which minimizes (**??**):

$$-2L+\lambda_1 J_1(\phi^*) = \sum_{i=1}^{N}\sum_{j=2}^{p_i}\sigma(t_j)^{-2}\left(y(t_{ij}) - \sum_{k=1}^{j-1}\phi(t_{ij},t_{ik})\,y\right)^2 + \lambda_1\left(||P_1\phi_1^*||^2 + ||P_1\phi_2^*||^2 + ||P_1\phi_{12}^*||^2\right) \tag{11}$$

4.3 Outline Smoothing Spline Approach

4.4 The truncated power basis and an alternative decomposition of $\mathcal{H}$

The estimation of $\phi^*(l,m)$ is quite different from the usual problem of estimating an arbitrary bivariate function via smoothing. In the case of the latter, we most typically treat both arguments equally in terms of regularization, but in the case of covariance estimation and the generalized coefficient function equal treatment of l and m in terms of penalization perhaps is not the most appropriate approach. The lag component, l, has particularly significant meaning in terms of the covariance function and thus also in terms of ϕ^* and is of considerable more interest than the orthogonal component, m. As discussed in Section 2, we can define an entire class of stationary functional autoregressive models using only the l direction, and additionally, as discussed in Section 3, there is a natural expectation about the functional form of the autoregressive coefficient function (and hence covariance) as a function of l, making imposing that conditional dependence between observation decay as l and the time between observations increase a reasonable way to institute regularization.

This latter notion is instrumental in justifying the family of penalties

$$J_{2,(p)} = \sum_{l_i\in\mathcal{L}:l_i>l_0} |\mu^* + \phi_1^*(l_i)|^p$$

which we may view as a design-driven way of implementing the regularization which may be imposed by the penalty functionals taking the form

$$\begin{aligned} J(\phi^*) &= \int_{l_0}^{1} |\mu^* + \phi_1^*(l)|^p \, dl \\ &= \int_{0}^{1} |\mu^* + \phi_1^*(l)|^p I(l > l_0) \, dl \end{aligned} \tag{12}$$

The penalty functionals given by (**??**) motivate a different decomposition of $\mathcal{H}$ than the derivative-based penalty. The form of (**??**) is significantly different in nature from the penalty discussed in Section 2.1 and those typically encountered in the setting smoothing spline ANOVA models, particularly because (**??**) effects only a subset of the domain for l. Therefore, an appropriate decomposition of the function space into the null space of J and the penalized space should perhaps be formulated in terms of basis functions for the lag component, l with domains which do not include the entire unit interval.

The truncated power basis, as in their use in defining polynomial regression splines, enjoy a particular ease of interpretation, as the coefficient β_{i+k} may be identified as the size of the jump at x_i in the k^{th} derivative of f. This fact is especially useful when tracking change points or, in general, any abrupt changes in the regression curve. If we reflect these basis functions about each of their corresponding knot points and denote these reflections $\{T_{ik}^{-}\}$, then expressing the regularization corresponding to the penalty functionals (**??**) becomes quite natural in terms of the reflected basis functions $(\cdot - l_1)_{-}^{k}, \ldots, (\cdot - l_n)_{-}^{k}$, where $(\alpha)_{-} = \max(-\alpha, 0)$. While the truncated power basis initially appears very attractive for representing functions in terms of the decomposition induced by penalties of the same form as that in Equation 12, they

5 P-splines

5.1 Truncated Power Basis

5.2 B-spline Basis

5.3 Difference penalties