Nonparametric Covariance Estimation for Longitudinal Data via Penalized Tensor Product Splines

Tayler A. Blake[*]   Yoonkyung Lee[†]

May 21, 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

1

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. 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. [add citation.] There is an extensive 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 address 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. We model the generalized autoregressive parameters as the evaluation of a continuous bivariate function at the pairs of observed time points. This allows us to accommodate problematic sampling as well as estimate

the covariance between two measurements at any pair of time points within the time interval of interest, and not just at the observed pairs of time points. 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). Decomposing the precision matrix 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 unit lower triangular matrix T with diagonal entries equal to 1 which diagonalizes Σ:

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

Let $Y = (y_1, y_2, \ldots, y_m)^T$ denote a random vector of measurements taken at times $t_1 < t_2 < \cdots < t_m$, with

$$E[Y] = 0.$$

The entries of T and D are easily interpretable if we 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 $y_1 = e_1$. The regression coefficients $\{\phi_{jk}\}$ are referred to as the *generalized autoregressive parameters* (or GARPs), and the $\{\sigma_j\}$ are referred to as the *innovation variances* (or IVs.) 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 be the unit lower triangular matrix with elements $T_{jk}, j, k = 1, \ldots, m$, where

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

Then model 1 as follows:

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

Taking covariances on both sides of (2), we have

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

Immediately, we can express the precision matrix (and thus its inverse) entirely in terms of the regression parameters defining Model **??**:

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

To accommodate unequally spaced and subject-specific observation times, we consider Y and e as vectorized the stochastic processes: $Y(t)$ and $e(t)$. We assume $Y(t)$ has 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, and similarly, we treat the elements of the precision matrix Ω as the values of some smooth function, $\omega(s, t)$ evaluated at observed time points. A natural extention of the same perspective to the elements of the Cholesky decomposition T and D leads us to the varying coefficient (VC) models first introduced by Hastie and Tibshirani, which assume that regressors and response variables vary according to an *indexing variable*. These models are particularly attractive because they permit 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

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

for $t_1 < t_2 < \cdots < t_m$. Several approaches to estimate the coefficient function $\phi(\cdot, \cdot)$ have been proposed. citewu2003nonparametric, for example, used locally weighted polynomials to smooth down the sub-diagonals of T. citehuang2007estimation smoothed the sub-diagonals of T using univariate smoothing splines. Within our formulation, the task of estimating a covariance matrix is equivalent to estimating the function $\phi(s, t)$ using bivariate smoothing. For ease of exposition, we assume that $\sigma^2(t)$ is fixed and known. 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 and let ϕ belong to a reproducing kernel Hilbert space, $\mathcal{H}$.

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. The autoregressive coefficients are often referred to in the literature as the *generalized autoregressive parameters*, or *GARPs*, and the error variances are often called the *innovation variances*, or *IVs*.

3 Enforcing Smoothness of ϕ^*

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¿ Previous work employs a number of approaches including banding, tapering, kernel smoothing, penalized likelihood, and penalized regression; see citetpourahmadi2011covariance for a comprehensive overview.

3.1 Outline Smoothing Spline Approach

3.2 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

$$
\begin{aligned}
J(\phi^*) &= \int_{l_0}^1 |\mu^* + \phi_1^*(l)|^p \, dl \\
&= \int_0^1 |\mu^* + \phi_1^*(l)|^p I(l > l_0) \, dl
\end{aligned}
\tag{5}
$$

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 kth 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 5, they

4   P-splines

4.1   Truncated Power Basis

4.2   B-spline Basis

4.3   Difference penalties