Nonparametric Covariance Estimation for Longitudinal Data via Penalized Tensor Product Splines

Tayler A. Blake* Yoonkyung Lee†

June 1, 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

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. Estimation of population covariance matrices from samples

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

of multivariate data has been important for methods 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 establishing independence or conditional independence through graphical models, constructing discriminant functions as in linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) for the classification of Gaussian data, building confidence intervals for component means and contrasts, and constructing a low-dimensional representation of data via principal components analysis (PCA). One may note that the last two techniques require an estimate of the covariance matrix, and the first two require estimation of the inverse.

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.

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. See Pourahmadi [2011] for a survey of approaches to covariance estimation from the generalized linear modeling and regularization perspectives. 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 Wu and Pourahmadi [2003],Huang et al. [2007]. To our knowledge, however, no previous work has applied smoothing to 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. The Cholesky decomposition enables covariance estimation through the estimation of a 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, $\Omega = \Sigma^{-1}$, in terms of its Cholesky decomposition (see Pourahmadi [2007] 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 $Y = (Y_1, Y_2, \ldots, Y_m)^T$ be defined on a probability space with some probability measure $\mathcal{P}$ corresponding to the multivariate Normal distribution with mean 0 and covariance Σ, and let $Y_1, Y_2, \ldots, Y_m$ have associated 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

$$D = T\Sigma T^T \tag{3}$$

An attractive feature of this reparameterisation is that, regardless of the modelling approach, the estimated covariance matrix is guaranteed to be positive definite. The unconstrained regression coefficients $\{\phi_{jk}\}$ are referred to as the *generalized autoregressive parameters* (GARPs). The $\{\sigma_j^2\}$ 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

$$\Omega = \Sigma^{-1} = T^T D^{-1} T. \tag{4}$$

Rather than estimating a specific covariance matrix for data observed on a fixed, regular grid, we aim to estimate a smooth covariance function. This accomodates data which may consist of observations on multiple subjects measured at potentially unequally spaced and individual-specific times. In estimation of the means μ of p-vectors of i.i.d. variables, the Gaussian white noise model [9] is the appropriate infinite-dimensional model into which all objects of interest are embedded. In estimation of matrices, a natural analogue is the space B(l2,l2), which we write as B, of bounded linear operators from l2 to l2. These can be represented as matrices [cite *Regularized estimation of large covariance matrices by Bickel and Levina - section 4.*]

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. We assume that $G(s,t)$ satisfies some smoothness conditions, where smoothness is defined in terms of square integrability of certain derivatives. [TODO: clean up statement about smoothness of covariance function; integrability of covariance function of a stochastic process?] 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.

Extending 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. The procedures presented by Fan and Zhang [2000] and Huang et al. [2002] utilize varying coefficient models for modeling the mean of longitudinal data; parameterizing the covariance matrix according to 3 allows us to exploit these models in covariance estimation for such data as well. 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

$$y(t_j) = \sum_{k=1}^{j-1} \phi(t_j, t_k)\, y(t_k) + \sigma(t_j)\, \epsilon(t_j) \qquad j = 1, \ldots, m, \tag{5}$$

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

We represent the varying coefficient function and the innovation variances using tensor product smoothing splines and penalized tensor product B-splines alongside penalties to induce simplicity in ϕ and σ^2 to produce final covariance estimates exhibiting the desired null structure. For ease of exposition, we first focus our attention on the estimation of ϕ assume that $\sigma^2(t)$ is fixed and known; we will later propose an iterative procedure for simultaneous estimation of σ^2 and ϕ. Recasting the problem as the estimation of model 5 allows us access to the existing set of tools developed in the bivariate smoothing literature; our approach provides a flexible, comprehensive framework for covariance estimation.

3 Penalized Maximum Likelihood Estimation of ϕ

We employ maximum likelihood for the estimation of the varying coefficient function $\phi(t, s)$ and the innovation variance function $\sigma(t)$, though neither the derivation the form of model 1 nor model 5 via the Cholesky decomposition rely on any assumptions about the distribution of Y. Fixing σ_j^2, for a sample of N i.i.d. observations $Y_1, Y_2, \ldots, Y_N$ from a multivariate Gaussian distribution, the negative log-likelihood as a function of ϕ_{jk} corresponds to the usual error sums of squares and is proportional to

$$-2L(y_1, y_2, \ldots, y_N, \Phi) \propto \sum_{i=1}^{N} \sum_{j=2}^{m_i} \sigma(t_j)^{-2} \left(y_{ij} - \sum_{k=1}^{j-1} \phi(t_{ij}, t_{ik})\, y_{ik} \right)^2 \tag{6}$$

where

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

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, a condition that is potentially worsened when one or more subjects has at least one observation time that is unique from the set of observation times common across subjects. 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.

4 Representation of ϕ as a smooth function

To impose structure on the estimated varying coefficient function, we augment the negative log-likelihood **??** with penalty functional, which discourages the flexibility of the fitted function. We take the estimator of ϕ to minimize

$$-2L + \lambda J_\phi(\phi). \tag{7}$$

The first term in 7 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. 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 $s - t$ input axis, we reparameterize ϕ and express the coefficient function in terms of

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

Writing ϕ in terms of the rotation gives the reparameterized coefficient function

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

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

$$-2L + \lambda^* J_{\phi^*}(\phi^*). \tag{9}$$

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 Gu [2002], modeling

$$\phi^*(l, m) = \mu + \phi_l(l) + \phi_m(m) + \phi_{lm}(l, m). \tag{10}$$

As in Gu [2002], Craven and Wahba [1978],[more Wahba citations here], we consider functions ϕ^* belonging to a reproducing kernel Hilbert space (r.k.h.s.), $\mathcal{H}$. We equip each l and m with corresponding univariate Hilbert spaces, $\mathcal{H}_l$ and $\mathcal{H}_m$, choosing to let $\mathcal{H}_l$ correspond to the second-order Sobolev space $W_2(0,1)$ and $\mathcal{H}_m$ to the first-order Sobolev space $W_1(0,1)$, where

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

for $m = 1, 2$. Each space $\mathcal{H}_l$, $\mathcal{H}_m$ is endowed with inner product product

$$\left\langle f, g \right\rangle = \sum_{\nu=0}^{m-1} \left(\int_0^1 f^{(\nu)}(x)\, dx\right)\left(\int_0^1 g^{(\nu)}(x)\, dx\right) + \int_0^1 f^{(m)} g^{(m)} dx \tag{11}$$

The space of bivariate functions $\mathcal{H}$ can be constructed from the tensor product of the univariate function spaces for l and m:

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

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 Wu and Pourahmadi [2003], Huang et al. [2007] . 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 10 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 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, spanned by $\tau_1, \tau_2, \ldots \tau_M$, 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 $\mathcal{H}_1$:

$$\begin{aligned} J(\phi) &= ||P_1\phi^*||^2 \\ &= ||P_1\phi_l||^2 + ||P_1\phi_m||^2 + ||P_1\phi_{lm}||^2 \end{aligned} \tag{12}$$

The decomposition of $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ can be characterized by the decompositions of $\mathcal{H}_l$ and $\mathcal{H}_m$ induced by J:

$$\begin{aligned} \mathcal{H}_l &= \mathcal{H}_{l0} \oplus \mathcal{H}_{l1} \\ \mathcal{H}_m &= \mathcal{H}_{m0} \oplus \mathcal{H}_{m1} \end{aligned} \tag{13}$$

As $\lambda \to \infty$, the penalty term dominates the objective function in 9, forcing the minimizer to adopt the functional form of the $\mathcal{H}_0$. The parameterization in 10 allows us to easily construct penalties to so that for large values of λ, the fitted function will correspond to [cite the simple parametric and semiparametric models of Pourahmadi, Wu, etc as well as the null models proposed by others utilizing smoothing methods]. We consider specification of the penalty so that the null space excludes any functions ϕ^* which are non-constant in m, letting

$$\mathcal{H}_{m0} = \{1\} \tag{14}$$

Additionally, we let ϕ^* which are linear in lag l to incur zero penalty, letting

$$\mathcal{H}_{l0} = \{1\} \oplus \{k_1\}, \tag{15}$$

where $k_\nu = B_\nu/\nu!$ are scaled Bernoulli polynomials satisfying

$$\begin{aligned} B_0(x) &= 1, \\ \frac{d}{dx} B_j(x) &= j B_{j-1}(x). \end{aligned}$$

The penalized spaces $\mathcal{H}_{l1}$, $\mathcal{H}_{m1}$, defined as the subspaces orthogonal to $\mathcal{H}_{l0}$ and $\mathcal{H}_{m0}$ respectively, satisfy

$$\begin{aligned} \mathcal{H}_{l1} &= \{\phi_l : \int_0^1 \phi_l{}^{(\nu)}(l)\, dl = 0, \;\; \nu = 0, 1\} \\ \mathcal{H}_{m1} &= \{\phi_m : \int_0^1 \phi_m(m)\, dm = 0\} \end{aligned}$$

Using the properties of tensor product spaces, we may write $\mathcal{H}_0$ and $\mathcal{H}_1$ in terms of the elements defining the marginal subspaces:

$$\begin{aligned} \mathcal{H}_0 &= \{1\} \oplus \{k_1\} \\ \mathcal{H}_1 &= \mathcal{H}_{l1} \oplus \mathcal{H}_{m1} \oplus [\{k_1\} \otimes \mathcal{H}_{m1}] \oplus [\mathcal{H}_{l1} \otimes \mathcal{H}_{m1}.] \end{aligned}$$

Table 1: Tensor product space $\mathcal{H}$

	$\{1\}$	$\{k_1(l)\}$	$\{\mathcal{H}_{l1}\}$
$\{1\}$	$\{1\}$	$\{k_1(l)\}$	$\{\mathcal{H}_{l1}\}$
$\{\mathcal{H}_{m1}\}$	$\{\mathcal{H}_{m1}\}$	$\{\mathcal{H}_{m1}\} \otimes \{k_1(l)\}$	$\{\mathcal{H}_{m1}\} \otimes \{\mathcal{H}_{l1}\}$

The subspaces of $W_1[0,1] \otimes W_2[0,1]$ by the tensor product of the marginal subspaces of $\mathcal{H}_l$, $\mathcal{H}_m$.

Table 1 shows how the space of two-dimensional functions is constructed by taking tensor products of each of the subspaces which define the two univariate spaces, $\mathcal{H}_l$ and $\mathcal{H}_m$. One may show that the reproducing kernels for $\mathcal{H}_{l0}$ and $\mathcal{H}_{l1}$ are given by $R_l^0(l,l') = \sum_{\nu=0}^{1} k_\nu(l)\, k_\nu(l')$ and $R_l^1(l,l') = k_2(l)\, k_2(l') - k_4([l-l'])$, respectively, where $[z]$ denotes the integer part of $z \in \mathbb{R}$. The reproducing kernel for the full marginal space $\mathcal{H}_l$ is simply the sum of the reproducing kernels for each of the subspaces:

$$R_l(l,l') = \sum_{\nu=0}^{1} k_\nu(l)\, k_\nu(l') + k_2(l)\, k_2(l') - k_4([l-l']).$$

One can also show that the reproducing kernels for $\mathcal{H}_{m0}$ and $\mathcal{H}_{m1}$ are given by $R_m^0(m,m') = 1$ and $R_m^1(m,m') = k_1(m)\, k_1(m') + k_2(m)\, k_2(m') - k_4([m-m'])$, and similarly, the reproducing kernel R_m for the full marginal space $\mathcal{H}_m$ is taken to be the sum of the kernels for each subspace. The tensor product reproducing kernel for $W_1[0,1] \otimes W_2[0,1]$ is obtained by simply taking the product of each of the reproducing kernels for the univariate function spaces:

$$R((l,l'),(m,m'))$$

Table 2 shows the decomposition of the reproducing kernel for the space of bivariate functions into the product of the reproducing kernels for each of the univariate subspaces.

Table 2: Tensor product reproducing kernel $R((l,m),(l',m'))$

	$\{1\}$	$\{k_1(l)\}$	$\{\mathcal{H}_{l1}\}$
$\{1\}$	$\{1\}$	$k_1(l)\, k_1(l')$	R_1^l
$\{\mathcal{H}_{m1}\}$	$R_m^1(m,m')$	$R_m^1(m,m')\, k_1(l)\, k_1(l')$	$R_m^1(m,m')\, R_l^1(l,l')$

For $\mathcal{H}$ defined as above, our goal is to find $\phi^* \in \mathcal{H}$ that minimizes

$$-2L + \lambda||P_1\phi^*||^2 = \sum_{i=1}^{N}\sum_{j=2}^{m_i} \sigma(t_j)^{-2}\left(y_{ij} - \sum_{k=1}^{j-1}\phi^*(l_{ijk}, m_{ijk})\, y_{ik}\right)^2 + \lambda||P_1\phi^*||^2 \qquad (16)$$

where (l_{ijk}, m_{ijk}) are the result of transforming subject i's j and k^{th} observation times, t_{ij} and t_{ik}. Let B denote the $p \times M$ matrix with columns correponding to the M basis functions spanning $\mathcal{H}_0$ evaluated at the $p = \sum_i = 1^N \binom{n_i}{2}$ within-subject pairs of observed time points. In spirit of the result on the representation of a univariate $f\mathcal{H}$ given by citekimeldorf1971some, we can similarly show that the minimizer of 16 lies within a finite dimensional space despite the minimization being carried out over an infinite dimensional space, $\mathcal{H}$.

Theorem 4.1. *Let* $p = \sum_{i=1}^{N} \binom{n_i}{2}$ *be the total number of distinct within-subject pairs of design points, and index the transformed pairs* $(l,m)_i$, $i = 1, \ldots, p$. *Let* B *be the* $p \times 2$ *matrix with* $(i,j)^{th}$ *entry* $k_j\left((l,m)_i\right)$ *with rank* $r = 2$. *Then, the unique minimizer of the penalized likelihood* (**??**), $\phi^* \in \mathcal{H}$ *is of the form*

$$\phi^*(l,m) = d_0 + d_1 k_1(l) + \sum_{i=1}^{p} c_i R^1\left((l,m),(l,m)_i\right) \tag{17}$$

The proof is left to the appendix.

4.2 Demonstration of the smoothing spline fitting procedure: Examples

5 Optimal shrinkage of T: smoothness may not be enough

Estimating the varying coefficient function ϕ^* 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 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. The use of smoothing splines to estimate ϕ outlined in Section **??** yields smooth null models, but smoothness of the elements of the Cholesky factor alone may not lead to desirable structure in the inverse covariance matrix.

There has been a recent upsurge in both the theoretical analysis and study of the practical application of regularization procedures for large empirical covariance matrices, including Huang et al. [2006], Furrer and Bengtsson [2007], Fan et al. [2008], Ledoit and Wolf [2004]. These works examine different types of regularization imposed on different ; Furrer and Bengtsson [2007] propose stabilizing the sample covariance matrix by "tapering," or gradually shrinking the off-diagonal elements to zero. d'Aspremont et al. [2008] propose a sparse estimator by applying an L_1 penalty directly to the elements of the covariance matrix. Instead of regularizing the covariance matrix itself, others have opted to regularize its inverse; Wu and Pourahmadi [2003] use the Cholesky decomposition to band the inverse covariance matrix by setting certain diagonals of the Cholesky factor to zero. Huang et al. [2006] and Levina et al. [2008] use L_1 penalties to achieve parsimony of the entries of the Cholesky factor; sparsity of the Cholesky factor, however, does not necessarily imply sparsity in the inverse covariance matrix. Levina et al. [2008] propose banding the Cholesky factor using a nested Lasso penalty which yields sparse estimators of the precision matrix.

These approaches implicitly adopt different notions of sparsity. Like Huang et al. [2006] and Levina et al. [2008], we are interested in regularizing the inverse of the covariance matrix through the Cholesky factor (or rather, the function defining its elements). The elements of Ω can be

interpreted as the partial correlations between the pairs of observations, so it is reasonable to expect for y_i and y_j to be conditionally uncorrelated (or nearly so) for large $|i-j|$; this suggests that enforcing that the conditional dependence between observations decay to zero as l increases is a reasonable way to institute regularization. Pourahmadi [1999] was one of the first to hueristically argue that the GARPs, $\phi_{t,t-l}$ should be monotonically decreasing in absolute value as l increases. In effect, this is equivalent to proposing that the effect of y_{t-l} on y_t through model 5 should decrease as the time between the two measurements increases. They and several others propose regularized estimators of the inverse covariance by banding T: setting all elements of T beyond the K^{th} off-diagonal to zero, i.e. setting $\phi_{t,t-l}=0$ for $l>K$ for some choice of K. (See Wu and Pourahmadi [2003], Bickel and Levina [2008], and Huang et al. [2007].) In terms of model 1, this is equivalent to regressing y_t on only its K immediate predecessors and setting the regression coefficient for y_{t-l} to zero for $l>K$. We refer to this regularization as "banding the Cholesky factor," and it is particularly attractive because it has the effect of enforcing conditional independence between pairs of observations with measuring times that are more than K lags apart. To show this, we will establish an instrumental relationship between patterns of zeros in positive definite matrices and their Cholesky factors.

Proposition 5.1. *Let Ω denote a $m\times m$ positive definite matrix with elements ω_{ij} with modified Cholesky decomposition $T^TD^{-1}T$, where T is unit lower triangular. Let t_{ij} denote the ij^{th} element of T. For any column j and row $r(j)>j$, $\omega_{mj}=\cdots=\omega_{r(j),j}=0$ if and only if $t_{mj}=\cdots=t_{r(j),j}=0$.*

The proof is left to the appendix. Proposition 5.1 maintains that the modified Cholesky factor T with arbitrary column band lengths corresponds to inverse covariance matrix Ω with the same column band lengths, and hence the inverse covariance matrix is K-banded if and only if its Cholesky factor is K-banded. This notion is instrumental in justifying the following family of penalties:

$$J_{\mathcal{B}}=\sum_{l_i\in\mathcal{L}:l_i>l_0}|\mu^*+\phi_1^*(l_i)|^p \tag{18}$$

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

6 Banding the inverse covariance

$$J_{\mathcal{B}}=\int_{l_0}^{1}|\phi^*(l,m)|^p dl. \tag{19}$$

The form of these penalties is significantly different in nature from the penalty discussed in Section 2.1 and those typically encountered in the smoothing spline ANOVA setting. A function $\phi^*\in\mathcal{H}$ incurs penalty via 19 through its behaviour on only a subset of its domain, inducing an alternative class of function space decompositions than those induced by the traditional derivative-based penalties. Zero penalty is assigned to functions having bounded support:

$$\mathcal{H}_0(\mathcal{B}) = \{\phi^* : \phi^*(l,m) = 0 \text{ for all } l \geq l_0\}. \tag{20}$$

Decomposing $\mathcal{H}$ into $\mathcal{H}_0(\mathcal{B})$ and its orthogonal complement necessitates an alternative to the smoothing spline basis functions.

To further impose simplicity in the structure of the inverse covariance, we consider "banding" the functional components of these of these null models; specifically, if we consider any ϕ^* corresponding to a Toeplitz precision matrix, that is any ϕ^* of the form

$$\phi^*(l,m) = \mu^* + \phi_1^*(l) \tag{21}$$

we propose truncating the functional lag components: the overall mean and the main effect of l to zero for any $l > l_0$ for some truncation point $l_0 \in (0,1)$. We consider the class of penalty functions that can be written in terms of an L_p norm of the sum of the overall mean and the functional main effect of l. We follow in the work of citehuang2006covariance and consider penalties which may be written

Bickel and Levina [2008] discuss 2004; 2008a; Wu and Pourahmadi, 2009) and generally those based on the Cholesky decomposition of the covariance matrix or its inverse (Pourahmadi, 1999, 2000; Rothman et al. 2010) do impose an order among the components of Y and are not permutation-invariant.

where $\mathcal{L}$ denotes the observed values of l, so that any ϕ^* to which J_2 assigns zero penalty is one that inherits nonzero contribution from stationary functional components only for lags $l \leq l_0$. We focus our attention to two important members of this family of penalties: the L_2 penalty and the L_1 penalty, given by

$$\begin{aligned} J_{2,(2)} &= \sum_{l_i \in \mathcal{L}: l_i > l_0} \left(\mu^* + \phi_1^*(l_i)\right)^2 \quad \text{and} \\ J_{2,(1)} &= \sum_{l_i \in \mathcal{L}: l_i > l_0} \left|\mu^* + \phi_1^*(l_i)\right| \end{aligned} \tag{22}$$

respectively. These penalties will induce shrinkage in the autoregressive coefficient function ϕ^* (and hence in the overall inverse covariance function) as in ridge regression and LASSO, respectively. Considering both types of regularization introduced in this section and in the previous, any ϕ^* belonging to the set of models incurring zero penalty from both J_1 and J_2 may be written

$$\phi^*(l,m) = \begin{cases} d_0 + d_1 k_1(l), & l \leq l_0 \\ 0, & l > l_0 \end{cases}$$

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

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 **??**, they

7 P-splines

7.1 Truncated Power Basis

7.2 B-spline Basis

7.3 Difference penalties

8 Appendix

Proof of Theorem 4.1

Then we may verify that any $\phi^* \in \mathcal{H}$ can be written

$$\phi^*(l, m) = d_0 + d_1 k_1(l) + \sum_{i=1}^{n} c_i R_1((l, m), (l_i, m_i)) + \rho(l, m)$$

where $\rho \perp \mathcal{H}_0 = \{1\} \oplus \{k_1\},\ span\{R_1((l_i, m_i), \cdot)\}$. We do so by demonstrating that ρ does not improve the first term in (**??**) (the data fit functional) and only adds to the penalty term, $J(\phi^*)$. Consequently, if $\hat{\phi}^*$ is the minimizer of (**??**), then $\rho = 0$. Using the properties of reproducing kernels, we can rewrite ϕ^* as an inner product of itself with R:

$$
\begin{aligned}
\phi^{*}\left(l_j, m_j\right) &= \left\langle R\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), \phi^{*}(\cdot,\cdot)\right\rangle \\
&= \Big\langle R_0\left(\left(l_j, m_j\right), (\cdot,\cdot)\right) + R_1\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), d_0 + d_1 k_1(\cdot) \\
&\qquad\qquad + \sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right), (\cdot,\cdot)\right) + \rho\left((\cdot,\cdot)\right)\Big\rangle \\
&= \left\langle R_0\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), d_0 + d_1 k_1(\cdot)\right\rangle + \left\langle R_0\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), \sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right), (\cdot,\cdot)\right)\right\rangle \\
&\quad + \left\langle R_0\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), \rho\left((\cdot,\cdot)\right)\right\rangle + \left\langle R_1\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), d_0 + d_1 k_1(\cdot)\right\rangle \\
&\quad + \left\langle R_1\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), \sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right), (\cdot,\cdot)\right)\right\rangle + \left\langle R_1\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), \rho\left((\cdot,\cdot)\right)\right\rangle \\
&= \left\langle R_0\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), d_0 + d_1 k_1(\cdot)\right\rangle + \left\langle R_1\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), \sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right), (\cdot,\cdot)\right)\right\rangle \\
&\quad + \underbrace{\left\langle R_0\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), \rho(\cdot,\cdot)\right\rangle}_{0} + \underbrace{\left\langle R_1\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), \rho(\cdot,\cdot)\right\rangle}_{0} \\
&= d_0 + d_1 k_1(\cdot) + \sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right), \left(l_j, m_j\right)\right)
\end{aligned}
$$

Rewriting the data fit functional, we have that

$$
\begin{aligned}
&\sum_{i=1}^{N}\sum_{j=1}^{n_i} \sigma_{ij}^{-2}\left(y\left(t_{ij}\right) - \sum_{k=1}^{j-1} \phi^{*}\left(t_{ij}, t_{ik}\right) y\left(t_{ik}\right)\right)^2 \\
= &\sum_{i=1}^{N}\sum_{j=1}^{n_i} \sigma_{ij}^{-2}\left(y\left(t_{ij}\right) - \sum_{k=1}^{j-1} \left\langle R\left(\left(l_{jk}^{i}, m_{jk}^{i}\right), (\cdot,\cdot)\right), \phi^{*}(\cdot,\cdot)\right\rangle y\left(t_{ik}\right)\right)^2
\end{aligned}
$$

which is free of ρ. Consider the contribution of any nonzero ρ to $J\left(\phi^{*}\right)$:

$$
\begin{aligned}
J\left(\phi^{*}\right) &= ||P_1\phi^{*}||^2 \\
&= \left\langle \sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right), (\cdot,\cdot)\right) + \rho(\cdot,\cdot), \sum_{j=1}^{N_{\phi^*}} c_j R_1\left(\left(l_j, m_j\right), (\cdot,\cdot)\right) + \rho(\cdot,\cdot)\right\rangle \\
&= || \sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right), (\cdot,\cdot)\right) ||^2 + ||\rho||^2
\end{aligned}
$$

Thus, including ρ in ϕ^{*} only increases the penalty without improving (decreasing) the data fit

functional, so we indeed have that the minimizer of (**??**) has the form

$$\phi^* (l, m) = d_0 + d_1 k_1 (l) + \sum_{i=1}^{N_{\phi^*}} c_i R_1 \left((l, m), (l_i, m_i) \right) \tag{23}$$

Proof: Proposition 5.1

Proof: Using the expression

$$\sigma^{ij} = \sum_{k=i}^{p} d_{ii} t_{ki} t_{kj}$$

it follows immediately that $t_{pj} = \cdots = t_{r(j),j} = 0$ implies that $\sigma^{pj} = \cdots = \sigma^{r(j),j} = 0$.

From citewatkins2004fundamentals, we can show that we can sequentially derive the elements of T and D according to

$$d_{ii} = \sqrt{\sigma^{ii} - \sum_{k=1}^{i-1} t_{ki}^2}$$

$$t_{ij} = \frac{1}{d_{ii}} \left(\sigma^{ij} - \sum_{k=1}^{i-1} t_{ki} t_{kj} \right)$$

We proceed by induction. For the first row of T^T,

References

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

P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403, 1978.

A. d'Aspremont, O. Banerjee, and L. El Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66, 2008.

J. Fan and J.-T. Zhang. Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62 (2):303–322, 2000.

J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

R. Furrer and T. Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255, 2007.

C. Gu. Smoothing spline anova models, 2002.

J. Z. Huang, C. O. Wu, and L. Zhou. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89(1):111–128, 2002.

J. Z. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, pages 85–98, 2006.

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

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

E. Levina, A. Rothman, and J. Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, pages 245–263, 2008.

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

M. Pourahmadi. Cholesky decompositions and estimation of a covariance matrix: orthogonality of variance-correlation parameters. *Biometrika*, pages 1006–1013, 2007.

M. Pourahmadi. Covariance estimation: The glm and regularization perspectives. *Statistical Science*, pages 369–387, 2011.

W. B. Wu and M. Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844, 2003.