Nonparametric Covariance Estimation with Shrinkage toward Stationary Models

Tayler A. Blake, *The Ohio State University*
Yoonkyung Lee, *The Ohio State University*

Technical Report No. 894

August, 2019

Department of Statistics
The Ohio State University
1958 Neil Avenue
Columbus, OH 43210-1247

Nonparametric Covariance Estimation with Shrinkage toward Stationary Models

Tayler A. Blake* Yoonkyung Lee*

Abstract

Estimation of an unstructured covariance matrix is difficult because of the challenges posed by parameter space dimensionality and the positive-definiteness constraint that estimates should satisfy. We consider a general nonparametric covariance estimation framework for longitudinal data using the Cholesky decomposition of a positive-definite matrix. The covariance matrix of time-ordered measurements is diagonalized by a lower triangular matrix with unconstrained entries that are statistically interpretable as parameters for a varying coefficient autoregressive model. Using this dual interpretation of the Cholesky decomposition and allowing for irregular sampling time points, we treat covariance estimation as bivariate smoothing and cast it in a regularization framework for desired forms of simplicity in covariance models. Viewing stationarity as a form of simplicity or parsimony in covariance, we model the varying coefficient function with components depending on time lag and its orthogonal direction separately and penalize the components that capture the nonstationarity in the fitted function. We demonstrate construction of a covariance estimator using the smoothing spline framework. Simulation studies establish the advantage of our approach over alternative estimators proposed in the longitudinal data setting. We analyze a longitudinal dataset to illustrate application of the methodology and compare our estimates to those resulting from alternative models.

Keywords: autoregression; Cholesky decomposition; covariance function; functional ANOVA model; stationary model

*Department of Statistics, The Ohio State University, Columbus, OH, USA

INTRODUCTION

Estimation of a covariance matrix is fundamental to the analysis of multivariate data for mean inference, discrimination, and dimension reduction. The two primary challenges in fulfilling this prerequisite are due to the total number of parameters to be estimated in relation to the data dimension, and a structural constraint for covariance. As compared to mean estimation, the number of parameters grows quadratically in the dimension, and these parameters must satisfy the positive-definiteness constraint. It is well known that the widely used sample covariance matrix, though positive-definite and unbiased for the population covariance matrix, is unstable in high dimensions (Lin, 1985; Johnstone, 2001). In the applied literature, it is common practice to specify a parametric model for the covariance structure by incorporating primary factors for variation in the data or those elements suggested by a study design. These models are typically parsimonious and require modest computational effort for estimation. However, specifying the appropriate covariance model is challenging even for the experts, and model misspecification can lead to considerably biased estimates.

On the other hand, several regularized estimators of the sample covariance have been proposed to balance the two extremes. There are several elementwise regularization methods for estimating a covariance matrix; see, for example, Bickel and Levina (2008); Rothman, Levina, and Zhu (2009); Cai, Zhang, and Zhou (2010). Methods for covariance estimation leveraging elementwise shrinkage are attractive, in part, because they typically present very low computational burden, but such estimators are not guaranteed to be positive-definite with finite sample sizes. A direct local polynomial smoothing of the sample covariance matrix proposed by Yao, Müller, and Wang (2005) does not ensure the positive-definiteness of the estimator either.

There has been a recent shift in covariance estimation toward regression-based approaches to eliminate the positive-definite constraint from estimation procedures altogether. Similar to this idea is the approach of modeling various matrix decompositions directly rather than the covariance matrix itself, including the spectral decomposition, the variance-correlation decomposition, and the Cholesky decomposition. The Cholesky decomposition in particular has recently received much attention because of its qualities that make it particularly at-

tractive for its use in covariance estimation for data with naturally ordered measurements such as time series or longitudinal data. The entries of the lower triangular matrix and the diagonal matrix of the modified Cholesky decomposition have statistical interpretations as autoregressive coefficients, or the *generalized autoregressive parameters* and prediction variances, or *innovation variances* when regressing a measurement on its predecessors. The unconstrained reparameterization and its statistical interpretability makes it easy to cast covariance modeling into the generalized linear model framework while guaranteeing that the resulting estimates are positive-definite. See Pourahmadi (2011) for a general overview of modeling the Cholesky decomposition.

In this paper, we extend the regression model associated with the Cholesky decomposition of a covariance matrix to a functional varying coefficient model. Treating covariance estimation as bivariate smoothing, our framework naturally accommodates unbalanced longitudinal data and employs regularization as in the usual function estimation setting. The outline of the article is as follows. In *the Cholesky Decomposition* section, we review the role of the modified Cholesky decomposition in the unconstrained reparameterization of a covariance matrix. In the next section, we present a functional varying coefficient model for the elements of the reparameterized covariance matrix and propose a reproducing kernel Hilbert space framework for estimation of the varying coefficient function. We then demonstrate estimation of the innovation variances via smoothing splines. Section *Simulation Studies* presents numerical studies comparing the performance of our estimator to other covariance estimators proposed in the literature. We apply our method to a dataset collected from a longitudinal study of cattle weights in *Data Analysis* section.

THE CHOLESKY DECOMPOSITION

For a positive-definite covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ for p variables, there exist a lower triangular matrix $T \in \mathbb{R}^{p\times p}$ with unit diagonal entries and a diagonal matrix $D \in \mathbb{R}^{p\times p}$ with positive entries such that

$$D = T\Sigma T'. \tag{1}$$

This representation (1) is commonly referred to as the modified Cholesky decomposition of Σ.

The lower triangular entries of T are unconstrained and can be interpreted as the coefficients of a particular regression model for ordered variables, and the diagonal of D can be interpreted as the prediction error variances associated with the same model. Let $Y = (y_1, \ldots, y_p)'$ denote a mean zero random vector with positive-definite covariance matrix Σ, and consider regressing y_t on its predecessors $y_1, \ldots, y_{t-1}$. Let $\hat{y}_t$ be the linear least-squares predictor of y_t based on previous measurements $y_{t-1}, \ldots, y_1$, and let $\mathrm{Var}(\epsilon_t) = \sigma_t^2$ denote the variance of the corresponding prediction error, where $\epsilon_t = y_t - \hat{y}_t$. Regression theory gives us that there exist unique scalars ϕ_{tj} so that

$$y_t = \begin{cases} \epsilon_t, & t = 1 \\ \sum_{j=1}^{t-1} \phi_{tj} y_j + \epsilon_t, & t = 2, \ldots, p, \end{cases} \tag{2}$$

and the prediction errors ϵ_t are mean zero and independently distributed. If we negate the regression coefficients ϕ_{tj} and place them in the lower triangle of T so that the (t, j) entry of T is $-\phi_{tj}$, and let $D = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_p)'$, then the sequence of regression models in (2) can be written in matrix form as

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

Taking covariances on both sides of (3) gives the modified Cholesky decomposition (1). Thus, modeling a covariance matrix is equivalent to fitting a sequence of $p - 1$ varying-coefficient and varying-order regression models. Since the ϕ_{tj} are regression coefficients, these and the $\log \sigma_t^2$, are unconstrained. The regression coefficients of the model in (2) and the prediction error variances are referred to as the *generalized autoregressive parameters* and *innovation variances* (Pourahmadi, 1999, 2000). The powerful implication of the regression framework of decomposition (1) is the accessibility of the entire portfolio of regression methods for the task of modeling covariance matrices. Moreover, the estimator $\hat{\Sigma}^{-1} = \hat{T}'\hat{D}^{-1}\hat{T}$ constructed from the unconstrained parameters, ϕ_{tj} and σ_t^2, is guaranteed to be positive-definite.

However, it is unclear how to apply model (2) to irregular or incomplete data without prior imputation. In most longitudinal studies, the functional trajectories of the involved smooth random processes are not directly observable, and often, the observed data are sparse

and irregularly spaced measurements of these trajectories. In the case that there is no fixed number of measurements and set of associated observation times for each subject, it is unclear how to define the discrete lag as in the usual formulation of autoregressive models. This makes treatment of individual subdiagonals of the Cholesky factor or the covariance matrix itself infeasible. To handle data collected in such a manner requires methods which are formulated in terms of continuous measurements. We address this concern by extending the framework supported by the unconstrained parameterization in (1) to naturally accommodate unbalanced longitudinal data. In the following section, we present a functional varying coefficient model for the elements of the Cholesky decomposition and propose regularization using a reproducing kernel Hilbert space framework.

A FUNCTIONAL VARYING-COEFFICIENT MODEL FOR THE MODIFIED CHOLESKY DECOMPOSITION

Given a sample of repeated measurements on N independent subjects, we model the observed data collected on an individual as a realization of a continuous-time stochastic process $Y(t)$ at discrete "time" points. In general, t doesn't need to be time, but for the ease of exposition, assume that measurements are indexed by time. Let $Y_i = (y(t_{i1}), \ldots, y(t_{i,p_i}))'$ denote measurements taken on the i^{th} subject at observation times $\mathcal{T}_i = \{t_{i1} < \cdots < t_{i,p_i}\}$, $i = 1, \ldots, N$. We assume that measurement times are drawn from $\mathcal{T} = [0, 1]$ without loss of generality.

We extend the linear model corresponding to the Cholesky decomposition (2) with the following functional varying-coefficient model:

$$y(t_{ij}) = \sum_{k<j} \phi(t_{ij}, t_{ik})\, y(t_{ik}) + \epsilon(t_{ij}), \qquad \begin{matrix} i = 1, \ldots, N \\ j = 2, \ldots, p_i, \end{matrix} \tag{4}$$

where the prediction errors $\epsilon(t)$ follow a zero-mean Gaussian process with variance function $\sigma^2(t)$. In the setting where sampling points are subject-specific and varying in length, the covariance function of the underlying process $Y(t)$, $\mathrm{Cov}(Y(t), Y(s))$ becomes the natural target of interest.

As parsimonious parametric models, Pourahmadi (2000) and Pan and Mackenzie (2003)

considered low-order polynomials of the lag between observed time points for the generalized autoregressive coefficient function ϕ and polynomials of time for log innovation variances in the analysis of longitudinal data. Further, Wu and Pourahmadi (2003) proposed local polynomial smoothers to individually estimate the sub-diagonals of T for modeling ϕ, imposing smoothness along the direction of lag. Short-term dependence could be another form of parsimony for covariance models, and can be realized by truncating the varying coefficient at certain time lag, which leads to a banded matrix (Huang, Liu, Pourahmadi, & Liu, 2006; Levina, Rothman, & Zhu, 2008).

The time lag or the sub-diagonal direction of T plays a prominent role in those parsimonious models for expressing the dependence structure. Rather than modelling the varying coefficient function ϕ directly, we reparameterize it explicitly in terms of lag and its orthogonal direction so that the fitted function can easily be used for suggesting parsimonious or structured models for the covariance function. Specifically, we take stationarity as a form of parsimony in covariance models, including those models parameterizing the elements of T as a function of the lag between observations in the literature (Pourahmadi, 1999; Pourahmadi & Daniels, 2002; Pan & Mackenzie, 2003; Leng, Zhang, & Pan, 2010). To facilitate such model specification, we transform inputs from a pair of time points (t, s) for $t > s$ to the lag, $l = t - s \in [0, 1]$, and average, $m = \frac{t+s}{2} \in [0, 1]$, in the additive direction as illustrated in Figure 1, and model ϕ in terms of the new arguments l and m:

$$\tilde{\phi}(l, m) \equiv \tilde{\phi}\left(t - s, \frac{1}{2}(t + s)\right) = \phi(t, s). \tag{5}$$

In other words, the composition of $\tilde{\phi}$ and the coordinate transformation yields ϕ. For brevity of notation, we will use ϕ to refer to the generalized autoregressive coefficient function taken as a function of either (t, s) or (l, m). Its arguments, unless specified explicitly, are to be interpreted depending on the context.

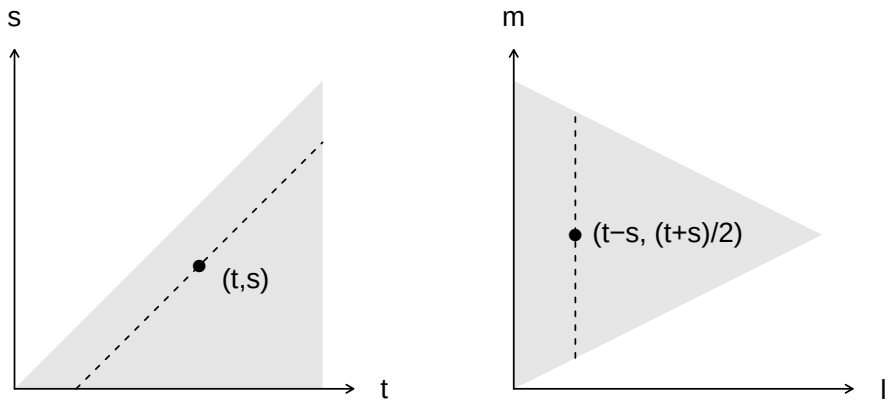


Figure 1: *Coordinate transformation of a pair of time points* (t, s) *for* $t > s$ *in the left panel to the lag and average in the additive direction,* (l, m)*, in the right panel.*

Model (4) corresponds to a stationary process when ϕ can be written as a function of lag l only and the innovation variances are constant in time t. For simplicity in the covariance model, we choose to regularize the autoregressive varying coefficient and the innovation variance function so that heavy penalization to both ϕ and σ^2 results in models which are close to stationary covariance matrices. To estimate $\phi(t, s)$ and $\sigma^2(t)$, we employ the smoothing spline framework (Wahba, 1990).

In particular, we model ϕ in a structured function space that allows decomposition of ϕ into functional components of lag l and additive direction m, and using the components, we specify penalties that naturally yield the aforementioned models in the literature as null models. For such a structural representation of ϕ, we adopt the smoothing spline ANOVA models in Gu (2013) taken as a functional analogue of the classical analysis of variance (ANOVA) model. They exhibit the same interpretability as their classical counterparts, allowing multivariate functions to be decomposed into components similar in spirit to the main effects and interaction terms associated with the ANOVA model. This property makes them especially useful for verifying or eliciting parametric models (Liu & Wang, 2004).

Two-Way Functional ANOVA Models

To model the varying coefficient function ϕ on $[0,1]^2$ using the smoothing spline ANOVA model framework, we first consider a univariate function space for lag l and additive direction m separately and take their tensor product. For example, the second-order Sobolev space $W_2[0,1] = \{f : [0,1] \to \mathbb{R} \mid f,\ f' \text{ absolutely continuous},\ \int_0^1 (f''(x))^2 dx < \infty\}$ can be taken as a model space for smooth univariate functions. When the curvature of f, $J(f) = \int_0^1 (f''(x))^2 dx$ is used as a roughness penalty functional for estimation of an unknown function from the space with data, the solution to the penalized least squares problem is known as a cubic smoothing spline. The function space $\mathcal{H} := W_2[0,1]$ can be equipped with inner product such that $\mathcal{H}$ as a Hilbert space is a direct sum of two orthogonal subspaces $\mathcal{H}_0$ and $\mathcal{H}_1$, the null space $\mathcal{H}_0$ consists of constant or linear functions taken as null models, and the penalty functional $J(f)$ corresponds to the squared norm of the projection of f onto $\mathcal{H}_1$ denoted by $\|P_1 f\|^2$. Further, with an appropriate averaging operator (e.g. $A(f) = \int_0^1 f(x)dx$) and a basis $k_1(\cdot)$ for linear functions in $\mathcal{H}_0$ satisfying $A(k_1) = 0$ (e.g. $k_1(x) = x - 1/2$), the null space $\mathcal{H}_0$ can be decomposed as a direct sum of $\{1\}$ and $\{k_1(\cdot)\}$. Thus, $\mathcal{H} = \{1\} \oplus \{k_1(\cdot)\} \oplus \mathcal{H}_1$ and each function $f(x)$ in $\mathcal{H}$ admits a unique representation of $c_0 + c_1 k_1(x) + f_1(x)$ with $c_0, c_1 \in \mathbb{R}$ and $f_1 \in \mathcal{H}_1$. This functional decomposition is akin to the one-way ANOVA model. In the representation, $c_1 k_1(x) + f_1(x)$ is treated as a functional main effect of x, and $c_1 k_1(x)$ and $f_1(x)$ are called parametric and nonparametric main effects, respectively.

Taking two structured function spaces for l and m, $\mathcal{H}^{[l]} = \{1\} \oplus \{k_1(l)\} \oplus \mathcal{H}_1^{[l]}$ and $\mathcal{H}^{[m]} = \{1\} \oplus \{k_1(m)\} \oplus \mathcal{H}_1^{[m]}$ as building blocks, we can define the tensor product space $\mathcal{H}^{[l]} \otimes \mathcal{H}^{[m]}$ and use it as a model space for bivariate ϕ. Analogous to the two-way ANOVA model, the subspaces of $\mathcal{H}^{[l]} \otimes \mathcal{H}^{[m]}$ define a unique decomposition of ϕ into the overall mean, main effects of l and m, and interaction of l and m: $\phi(l,m) = \mu + \phi_1(l) + \phi_2(m) + \phi_{12}(l,m)$.

In addition, we can specify the null space as the subspace with desired simple models (e.g. low-order polynomials of lag only), and use the functional norm associated with each subspace to define a general "roughness" penalty functional $J(\phi)$ for bivariate smoothing, which results in two-way smoothing spline ANOVA models. This penalized function esti-

mation framework is very flexible in the choice of a null space $\mathcal{H}_0$ and a penalty functional $J(\phi)$, allowing the user to adapt these choices to the context of data analysis and modeling.

Mathematically, smoothing spline ANOVA models are rooted in the theory of reproducing kernel Hilbert spaces (Aronszajn, 1950; Wahba, 1990; Berlinet & Thomas-Agnan, 2011). Reproducing kernels are essential to the characterization of function spaces, their subspaces, and related geometric notion of norms and projections. For clear exposition of the model fitting procedure, we briefly review some basic properties of reproducing kernel Hilbert spaces.

Reproducing Kernel Hilbert Spaces

A Hilbert space $\mathcal{H}$ of functions on a set $\mathcal{X}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is defined as a complete inner product linear space. For each $x \in \mathcal{X}$, let $[x]$ map $f \in \mathcal{H}$ to $f(x) \in \mathbb{R}$, which is known as the evaluation functional at x. A Hilbert space is called a reproducing kernel Hilbert space (RKHS) if the evaluation functional $[x] f = f(x)$ is continuous in $\mathcal{H}$ for all $x \in \mathcal{X}$. The Reisz Representation Theorem gives that there exists $K_x \in \mathcal{H}$, the representer of the evaluation functional $[x](\cdot)$, such that $\langle K_x, f \rangle_{\mathcal{H}} = f(x)$ for all $f \in \mathcal{H}$. See Theorem 2.2 in Gu (2013).

The symmetric, bivariate function $K(x_1, x_2) \equiv K_{x_1}(x_2) = \langle K_{x_1}, K_{x_2} \rangle_{\mathcal{H}}$ is called the reproducing kernel (RK) of $\mathcal{H}$. The RK satisfies that for every $x \in \mathcal{X}$ and $f \in \mathcal{H}$, $K(x, \cdot) \in \mathcal{H}$, and $f(x) = \langle f, K(x, \cdot) \rangle_{\mathcal{H}}$. The second property is called the reproducing property of K. Every reproducing kernel uniquely determines the RKHS, and in turn, every RKHS has a unique reproducing kernel. See Theorem 2.3 in Gu (2013). The representer of any bounded linear functional can be obtained from the reproducing kernel K. Further, if a reproducing kernel Hilbert space $\mathcal{H}$ is a direct sum of two orthogonal subspaces $\mathcal{H}_0$ and $\mathcal{H}_1$ with RKs K_0 and K_1, that is, $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, then the reproducing kernel for $\mathcal{H}$ is $K(x_1, x_2) = K_0(x_1, x_2) + K_1(x_1, x_2)$. See Aronszajn (1950) for other RKHS properties.

Estimation of the Generalized Varying Coefficient Function via Bivariate Smoothing

For estimation of the generalized coefficient function ϕ with data, we transform the observed time points to lags and averages in the additive direction. Given subject i and a pair of indices $j > k$, define $\boldsymbol{v}_{ijk} = \left(t_{ij} - t_{ik}, \frac{1}{2}\left(t_{ij} + t_{ik}\right)\right) = (l_{ijk}, m_{ijk}) \in \mathcal{V} = [0,1]^2$ as the tuple corresponding to the transformed pair of j^{th} and k^{th} observation times on the i^{th} subject. Let $V = \bigcup_{i,j,k} \{\boldsymbol{v}_{ijk}\} \equiv \left\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{|V|}\right\}$ denote the set of unique within-subject pairs of observation times when pooled across N subjects.

We let the autoregressive coefficient function ϕ belong to a reproducing kernel Hilbert space $\mathcal{H}$ with reproducing kernel K, which is structured as a tensor sum of the null space $\mathcal{H}_0$ and penalized space $\mathcal{H}_1$ with reproducing kernels K_0 and K_1, respectively. Let the penalty functional $J(\phi)$ measuring the complexity of ϕ, be $\|P_1\phi\|^2$, the squared norm of the projection of ϕ onto the subspace $\mathcal{H}_1$.

For example, consider $\mathcal{H} = \mathcal{H}^{[l]} \otimes \mathcal{H}^{[m]}$, where $\mathcal{H}^{[l]} = W_2[0,1] = \mathcal{H}_0^{[l]} \oplus \mathcal{H}_1^{[l]}$ with $\mathcal{H}_0^{[l]} = \{1\} \oplus \{k_1(l)\}$ and $\mathcal{H}^{[m]} = W_2[0,1] = \mathcal{H}_0^{[m]} \oplus \mathcal{H}_1^{[m]}$ with $\mathcal{H}_0^{[m]} = \{1\}$. This choice results in the null space $\mathcal{H}_0$ comprised of linear functions of lag only and amounts to penalizing the main effect of m, $\phi_2(m)$, and interaction of l and m, $\phi_{12}(l,m)$, altogether in addition to the curvature of the main effect of l. It has the effect of pulling estimated ϕ towards smooth functions of lag only treated as one form of parsimony in covariance modeling.

Under model (4), the negative log likelihood satisfies

$$-2\ell\left(\phi, \sigma^2 | Y_1, \ldots, Y_N\right) = \sum_{i=1}^{N} \sum_{j=1}^{p_i} \left[\log \sigma^2(t_{ij}) + \frac{1}{\sigma^2(t_{ij})}\left(y(t_{ij}) - \sum_{k<j} \phi(t_{ij}, t_{ik}) y(t_{ik})\right)^2\right] \quad (6)$$

up to an additive constant.

Fixing the innovation variances $\sigma_{ij}^2 = \sigma^2(t_{ij})$, we take the estimator of ϕ to be the minimizer of the penalized negative log likelihood:

$$-2\ell\left(\phi | Y_1, \ldots, Y_N, \sigma^2\right) + \lambda J(\phi) = \sum_{i=1}^{N} \sum_{j=2}^{p_i} \frac{1}{\sigma_{ij}^2}\left(y(t_{ij}) - \sum_{k<j} \phi\left(\boldsymbol{v}_{ijk}\right) y(t_{ik})\right)^2 + \lambda J(\phi), \quad (7)$$

where $\lambda > 0$ is a smoothing parameter, and denote it by ϕ_λ. The smoothing parameter λ controls the tradeoff between the goodness of fit measure ℓ and the penalty $\|P_1\phi\|^2$.

The following theorem establishes the form of the minimizer of the penalized negative log likelihood (7) and that the solution belongs to a finite-dimensional subspace despite the minimization being carried out over an infinite-dimensional space.

Theorem 1. *Let* $\{\nu_1, \ldots, \nu_{\mathcal{N}_0}\}$ *span* $\mathcal{H}_0 = \{\phi \in \mathcal{H} : J(\phi) = 0\}$*, the null space of* $J(\phi) = \|P_1\phi\|^2$*. Then the minimizer* ϕ_λ *of* (7) *is of the form*

$$\phi_\lambda(\boldsymbol{v}) = \sum_{i=1}^{\mathcal{N}_0} d_i \nu_i(\boldsymbol{v}) + \sum_{j=1}^{|V|} c_j K_1(\boldsymbol{v}_j, \boldsymbol{v}), \tag{8}$$

where $K_1(\boldsymbol{v}_j, \boldsymbol{v})$ *denotes the reproducing kernel for* $\mathcal{H}_1$ *evaluated at* $\boldsymbol{v}_j$*, the* j^{th} *element of* V*, viewed as a function of* $\boldsymbol{v} = (l, m)$*,* $d_i \in \mathbb{R}$*, and* $c_j \in \mathbb{R}$*.*

This result is an example of the well-known representer theorem that holds for minimizers of regularized empirical risk functionals in a RKHS, and obtained by the standard argument with reproducing kernel properties. The proof is left to the Appendix. Using the representation of the minimizer, we discuss how to determine the coefficients d_i and c_j with data.

Model Fitting

Let $Y = \left(Y_1^{(-1)\prime}, Y_2^{(-1)\prime}, \ldots, Y_N^{(-1)\prime}\right)'$ denote the vector of length $n_Y = \sum_i p_i - N$, constructed by stacking the N observed response vectors, less their first element: $Y_i^{(-1)} = (y(t_{i2}), \ldots, y(t_{i,p_i}))'$. Define X_i to be the $(p_i - 1) \times |V|$ matrix containing the covariates for subject i necessary for regressing each measurement $y(t_{i2}), \ldots, y(t_{i,p_i})$ on its predecessors as in model (4), and let $X = \begin{bmatrix} X_1' & X_2' & \ldots & X_N' \end{bmatrix}'$. Define K_V to be the $|V| \times |V|$ matrix with (i, j) entry given by $K_1(\boldsymbol{v}_i, \boldsymbol{v}_j)$, and let B denote the $|V| \times \mathcal{N}_0$ matrix with (i, j) entry equal to $\nu_j(\boldsymbol{v}_i)$.

Assuming that σ_{ij}^2 are given for now, let D denote the $n_Y \times n_Y$ diagonal matrix of innovation variances σ_{ij}^2, and let $\tilde{Y} = D^{-1/2}Y$, $\tilde{B} = D^{-1/2}XB$, and $\tilde{K}_V = D^{-1/2}XK_V$. Using the representation of ϕ_λ in (8), and defining coefficient vectors $c = (c_1, \cdots, c_{|V|})'$ and $d = (d_1, \cdots, d_{\mathcal{N}_0})'$, the penalized negative log likelihood in (7) is given by

$$-2\ell\left(c, d | \tilde{Y}, \tilde{B}, \tilde{K}_V\right) + \lambda J(\phi_\lambda) = \left[\tilde{Y} - \tilde{B}d - \tilde{K}_V c\right]' \left[\tilde{Y} - \tilde{B}d - \tilde{K}_V c\right] + \lambda c' K_V c. \tag{9}$$

For fixed smoothing parameter, setting partial derivatives with respect to d and c equal to zero, the solution ϕ_λ is obtained by finding c and d which satisfy:

$$\begin{bmatrix} \tilde{B}'\tilde{B} & \tilde{B}'\tilde{K}_{V} \\ \tilde{K}_{V}'\tilde{B} & \tilde{K}_{V}'\tilde{K}_{V} + \lambda K_{V} \end{bmatrix} \begin{bmatrix} d \\ c \end{bmatrix} = \begin{bmatrix} \tilde{B}'\tilde{Y} \\ \tilde{K}_{V}'\tilde{Y} \end{bmatrix}. \tag{10}$$

When $\tilde{K}_{V}$ is full column rank, the solution can be obtained through the Cholesky decomposition of the symmetric matrix on the left side of the equality in (10). Writing

$$\begin{bmatrix} \tilde{B}'\tilde{B} & \tilde{B}'\tilde{K}_{V} \\ \tilde{K}_{V}'\tilde{B} & \tilde{K}_{V}'\tilde{K}_{V} + \lambda K_{V} \end{bmatrix} = CC',$$

the solution is given by $\begin{bmatrix} \hat{d}' & \hat{c}' \end{bmatrix}' = C^{-1}(C')^{-1} \begin{bmatrix} \tilde{B} & \tilde{K}_{V} \end{bmatrix}' \tilde{Y}$. Singularity of $\tilde{K}_{V}$ demands special computational consideration to solve (10). For detailed examination, we refer the reader to Gu and Wahba (1991).

The appropriate choice of smoothing parameter λ is crucial for effectively recovering the true ϕ. In practice, a number of data-driven methods are available for model selection such as the Akaike or Bayesian information criterion (Eilers & Marx, 1996) or cross validation-based procedures (Wahba, 1990; Gu & Wahba, 1991) including the leave-one-subject-out cross validation (losoCV) criterion for repeated measures data (Xu & Huang, 2012).

ESTIMATION OF THE INNOVATION VARIANCE FUNCTION VIA SMOOTHING SPLINES FOR EXPONENTIAL FAMILIES

Given an estimate of ϕ, we can estimate the innovation variance function $\sigma^2(t)$, using the corresponding residuals as working innovation errors. If the true innovations $\epsilon(t_{ij})$ were given, then the joint likelihood in (6) would reduce to

$$-2\ell\left(\sigma^2|Y_1,\ldots,Y_N,\phi\right) = \sum_{i=1}^{N}\sum_{j=1}^{p_i}\left(\log\sigma^2(t_{ij}) + \frac{\epsilon^2(t_{ij})}{\sigma^2(t_{ij})}\right) \tag{11}$$

for estimation of $\sigma^2(t)$. The fact that $\epsilon^2(t_{ij})$ is a scaled chi-square random variable and the form of the likelihood above motivate a variance model for $\sigma^2(t)$ using Gamma distributions

with the $\epsilon^2(t_{ij})$ serving as the response. When a Gamma distribution with shape parameter α and scale parameter β is reparameterized with mean parameter $\mu = \alpha\beta$ in place of β, a negative log likelihood of μ based on a single observation z from the distribution is shown to be proportional to $\alpha\left(\log\mu + \frac{z}{\mu}\right)$ with α^{-1} treated as a fixed dispersion parameter. Recognizing the connection between the Gamma likelihood and (11), we cast estimation of the innovation variance function in a generalized linear model framework with Gamma errors and fixed shape parameter. Further, to remove the constraint that $\mu > 0$, we transform μ to $\eta = \log\mu$ and reparameterize the Gamma likelihood as $\alpha\left[\eta + z\exp(-\eta)\right]$.

Defining $\eta(t) = \log\sigma^2(t)$ and assuming a smooth log innovation variance function, we use the smoothing spline method for regression relating squared innovations, $\epsilon^2(t_{ij})$, as Gamma responses to time points t_{ij}. Generalized smoothing spline models that relate the canonical parameter of an exponential family to a set of covariates have been studied extensively. See Wahba, Wang, Gu, Klein, and Klein (1995), Wang (1997), and Gu (2013).

As with the estimation of the functional varying coefficient, estimation is carried out by minimizing the penalized negative log likelihood with the working innovation errors. Given ϕ^*, an estimate of ϕ, define the working innovation errors, $\hat{\epsilon}(t_{ij}) = y(t_{ij}) - \sum_{k<j}\phi^*\left(\boldsymbol{v}_{ijk}\right)y(t_{ik})$, and the corresponding squared innovations, $z_{ij} \equiv z(t_{ij}) = \hat{\epsilon}^2(t_{ij})$. Let $Z_i = \left(z(t_{i1}), \ldots, z(t_{i,p_i})\right)'$ denote the vector of squared innovations for the i^{th} observed trajectory. With $Z_1, \ldots, Z_N$, the negative log likelihood of $\eta(t)$ becomes

$$-2\ell\left(\eta|Z_1, \ldots, Z_N\right) = \sum_{i=1}^{N}\sum_{j=1}^{p_i}\left(\eta(t_{ij}) + z_{ij}e^{-\eta(t_{ij})}\right). \tag{12}$$

Similar to the estimation of ϕ, we consider a function space $\mathcal{H}$ for $\eta(t)$ on $[0,1]$ with an orthogonal decomposition of $\mathcal{H}_0 \oplus \mathcal{H}_1$ and define a roughness penalty $J(\eta)$ that can be written as the squared norm of the projection of η to $\mathcal{H}_1$. For instance, take $\mathcal{H} = W_2[0,1]$ with $J(\eta) = \int_0^1(\eta'(t))^2dt$ which corresponds to $\mathcal{H}_0 = \{1\}$. Combining the likelihood with a penalty, we define our estimator of $\eta(t)$ to be the minimizer of the penalized negative log likelihood:

$$-2\ell\left(\eta|Z_1, \ldots, Z_N\right) + \lambda J\left(\eta\right) = \sum_{i=1}^{N}\sum_{j=1}^{p_i}\left(\eta(t_{ij}) + z_{ij}e^{-\eta(t_{ij})}\right) + \lambda J\left(\eta\right). \tag{13}$$

The first term in (13) serves as a measure of the goodness of fit of η to the data, and only depends on η through the evaluation of η at observed time points. Thus, the argument justifying the form of the minimizer in (8) applies to η. Let $\mathcal{T}_{obs} = \bigcup_{i,j} \{t_{ij}\}$ denote the unique values of the observations times pooled across subjects. The minimizer of the penalized likelihood (13) has the form

$$\eta_\lambda(t) = \sum_{i=1}^{\mathcal{N}_0} d_i \nu_i(t) + \sum_{j=1}^{|\mathcal{T}_{obs}|} c_j K_1(t_j, t), \tag{14}$$

where $\{\nu_i\}$ form a basis for the null space $\mathcal{H}_0$ and $K_1(t_j, t)$ is the reproducing kernel for $\mathcal{H}_1$ evalutated at t_j, the j^{th} element of $\mathcal{T}_{obs}$, viewed as a function of t.

To jointly estimate the autoregressive coefficient function and the innovation variance function, we adopt an iterative approach in the spirit of Pourahmadi (2000), Huang et al. (2006), and Huang, Liu, and Liu (2007). A procedure for minimizing

$$-2\ell(\phi, \eta | Y_1, \ldots, Y_N) + \lambda_\phi J_\phi(\phi) + \lambda_\eta J_\eta(\eta)$$

starts with initializing $\eta(t_{ij}) = 0$ or $\sigma_{ij}^2 = \exp(\eta(t_{ij})) = 1$ for $i = 1, \ldots, N$, $j = 1, \ldots, p_i$. For fixed η, we find ϕ^* minimizing the penalized negative log likelihood

$$-2\ell(\phi | Y_1, \ldots, Y_N, \eta) + \lambda_\phi J_\phi(\phi).$$

Given ϕ^*, we update our estimate of η by taking η^* that minimizes the penalized negative log likelihood with the working squared residuals

$$-2\ell(\eta | Z_1, \ldots, Z_N, \phi^*) + \lambda_\eta J_\eta(\eta).$$

This process of iteratively updating ϕ^* and η^* is repeated until convergence.

SIMULATION STUDIES

In this section we compare our bivariate spline estimator to other methods of covariance estimation through simulation studies with generative models. Our primary comparisons are that with the polynomial estimator for ϕ and σ^2 proposed by Pan and Mackenzie (2003). Their approach, which is also based on the Cholesky decomposition, permits unbalanced data

without requiring missing data imputation. However, the polynomial estimator assumes that $\phi(t,s)$ can be parameterized as a (univariate) polynomial in $l = t - s$ only. Thus, discrepancies in the performance of the estimators may be indicative of situations in which our parameterization (5) is advantageous. We also consider the performance of the oracle estimator under each of the generating models, the sample covariance matrix and two of its regularized variants: the tapered sample covariance matrix (Cai et al., 2010) and the soft thresholding estimator (Rothman et al., 2009), neither of which rely on a natural ordering among the variables.

We consider the following five covariance structures for the data generating distribution. The covariance functions as two-dimensional surfaces corresponding to each generating model are shown left to right in Figure 2. The first row displays the surface coinciding with the appropriate discrete covariance matrix on a 10×10 grid, and the second row displays the surfaces of the corresponding Cholesky factors (the lower triangle of $-T$). The precise model definitions are in Table 1. When Σ is not directly specified in the table, the covariance matrices in Figure 2 are obtained by either evaluating the covariance function $\sigma(t,s)$ at 10 equally spaced points, $\{t_1, \cdots, t_{10}\}$, from $[0,1]$ or numerically constructing $\Sigma = T^{-1}DT'^{-1}$ after forming T and D from the specified autoregressive coefficient and innovation variance functions $\phi(t,s)$ and $\sigma^2(t)$.

Under each of the five covariance models, we generated data from a mean zero p-variate Normal distribution with covariance matrix $\Sigma = T^{-1}DT'^{-1}$ and constructed an estimate of Σ for each combination of $p = 10,\ 20,\ 30$ and sample size $N = 50,\ 100$. Since construction of the sample covariance matrix S and regularized variants S^{ω} (tapered) and S^{λ} (soft-thresholded) requires an equal number of observations on each subject taken at a common set of observation times, simulations were conducted using complete data, with observation times $t = 1, \ldots, p$ mapped to the unit interval. The smoothing spline estimator $\hat{\Sigma}_{SS}$ was constructed by using a tensor product cubic smoothing spline for ϕ and univariate cubic smoothing spline for $\sigma^2(t)$.

Following Pan and Mackenzie (2003), the polynomial estimator $\hat{\Sigma}_{poly}$ was obtained by modeling the generalized autoregressive coefficients $\phi(t_{ij}, t_{ik})$ as a degree q polynomial of $(t_{ij} - t_{ik})$ and the log innovation variances $\log \sigma^2(t_{ij})$ as a degree d polynomial of t_{ij}. The

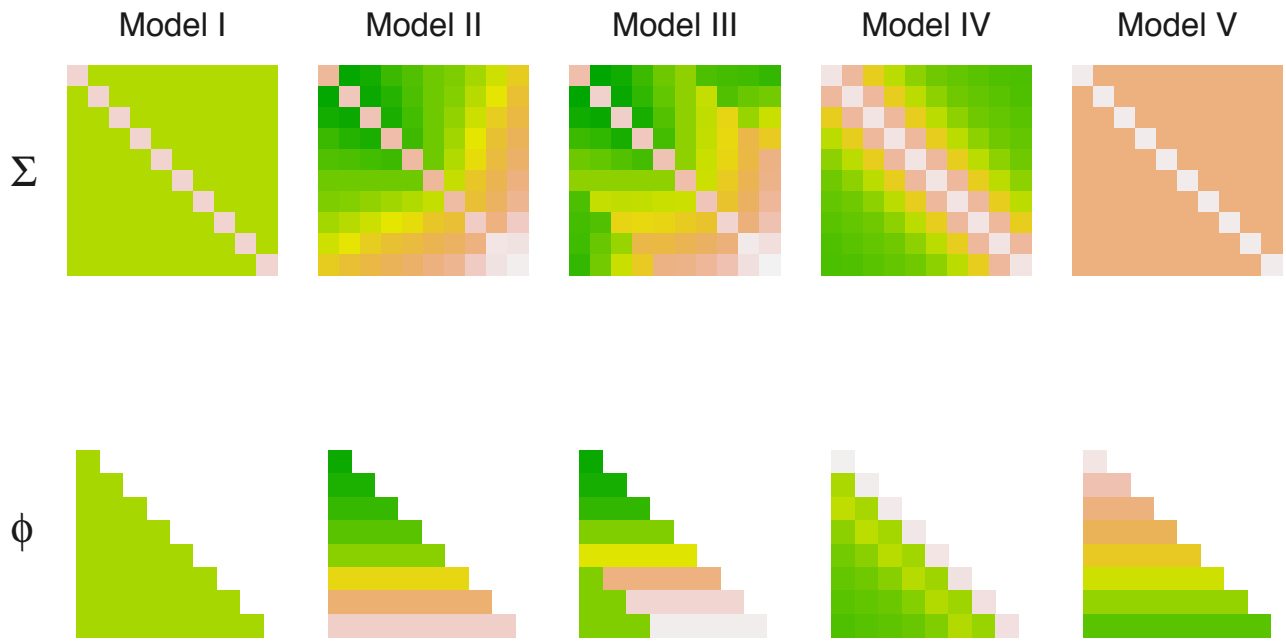


Figure 2: *Heatmaps of the true covariance matrices corresponding to Models I–V and ϕ defining the corresponding Cholesky factor T. The smallest elements of each matrix correspond to dark green pixels; the light pink (white) pixels correspond to the large (largest) elements of the matrix.*

regression parameters were estimated via maximum likelihood, and the optimal pair of polynomial orders (q, d) was selected using the Bayesian information criterion (BIC).

To assess performance of an estimator $\hat{\Sigma}$, we consider the entropy loss

$$\Delta\left(\Sigma, \hat{\Sigma}\right) = tr\left(\Sigma^{-1}\hat{\Sigma}\right) - \log|\Sigma^{-1}\hat{\Sigma}| - p,$$

which can be derived from the Wishart likelihood (Anderson, 1984). Given Σ, we prefer the estimator with the smallest risk, $R\left(\Sigma, \hat{\Sigma}\right) = E_{\Sigma}\left[\Delta\left(\Sigma, \hat{\Sigma}\right)\right]$. To evaluate the risk via Monte Carlo approximation, we generated 100 replicates of $\hat{\Sigma}$ and calculated the corresponding average loss.

Figure 3 provides a visual summary of the qualitative differences between the estimates resulting from each of the six methods of estimation for the five covariance structures used for simulation. The first row in the grid shows the surface plot of each of the true covariance structures, and each row thereafter corresponds to the five covariance estimates for the given

Model	Σ or $\sigma(t,s)$	$\phi(t,s)$ for $t > s$	$\sigma^2(t)$
I: Independence	I	0	1
II: Linear Coefficient	*	$t - 0.5$	0.1^2
III: Banded Linear	*	$\begin{cases} t-0.5 & \text{if } t-s \leq 0.5 \\ 0 & \text{if } t-s > 0.5 \end{cases}$	0.1^2
IV: Rational Quadratic	$\left(1 + \frac{(t-s)^2}{2k^2}\right)^{-1}$ with $k = 0.6$	*	*
V: Compound Symmetry	$(1-\rho)\,\mathrm{I} + \rho\mathrm{J}$ with $\rho = 0.7$	$\phi(t_j, t_k) = \frac{\rho}{1+(j-2)\rho}$ for $j > k$	$\sigma^2(t_j) = 1 - \frac{(j-1)\rho^2}{1+(j-2)\rho}$

Table 1: *Covariance models for data generation. The true covariance function $\sigma(t,s)$, varying coefficient function $\phi(t,s)$, and innovation variance function $\sigma^2(t)$ are defined with the domain $\mathcal{T} = [0,1]$. The asterisks indicate that the entries are determined numerically when discretized.*

estimation method.

Oracle estimators for each covariance model were constructed assuming that the structure of the underlying generating model is known. For example, the oracle estimator of the covariance matrix corresponding to mutual independence with constant variance is a diagonal matrix with the diagonal elements given by $\hat{\sigma}^2$, which is an estimate of the variance based on all of the data, $\{y_{t_{ij}}\}$. For Model II, the oracle estimator was obtained by fitting a linear function of t for the varying coefficient function and assuming constant innovation variance. For compound symmetry, a random effects model with subject-specific effects that yield the same covariance structure was considered and its variance parameters were estimated using the restricted maximum likelihood method to produce the oracle estimator. For each simulation setting, the risk of the oracle estimator serves as a lower bound on the risk for the given covariance structure.

A summary of the estimated entropy risk for the covariance estimators is presented in Table 2. Smoothing parameters for $\hat{\Sigma}_{SS}$ were chosen using the unbiased risk estimate (Gu, 2013, Chapter 3.22) and leave-one-subject-out cross validation. Performance is similar under

both criteria; for brevity, results under losoCV are omitted. Tuning parameter selection for the regularized versions of the sample covariance matrix was performed using cross validation. For detailed discussion, see Fang, Wang, and Feng (2016).

In general, our estimator outperforms the alternative estimators, particularly when the underlying true covariance matrix does not satisfy the implicit structural assumptions motivating their construction. For example, the risk for the polynomial estimator is much higher than our estimator under Models II and III due to model misspecification; the underlying ϕ is not a function of lag only. While the sample covariance matrix is an unbiased estimator of the unstructured covariance matrix, the smoothing spline estimator is better for every simulation model, and the difference is larger as p increases. Our estimator effectively makes use of the functional nature of the generating covariance models and their smoothness. It performs most poorly on Model III, where ϕ does not belong to the prescribed model space for smooth functions due to its discontinuous first derivative. Overall, the results indicate that the smoothing spline estimator achieves what it was designed to do. It provides a more stable and accurate estimate than the sample covariance matrix, but is guaranteed to be positive-definite unlike the soft thresholding estimator and the tapering estimator. It achieves this stability in performance across different scenarios with added flexibility over the polynomial estimator and exhibits better performance than the polynomial estimator under Model IV where parametric models of lag only may be appropriate.

To see how performance of our estimator changes when data are irregularly sampled, we carried out an additional experiment where data are subsampled from the complete trajectories in the first experiment by randomly omitting each observation with fixed probability in the range of 10 to 30%. Performance degradation of the estimator in the presence of missing data is highly dependent on the underlying structure of the Cholesky factor. The performance remains fairly stable across increasing proportions of missing data under Models I and IV. See Blake (2018) for details.

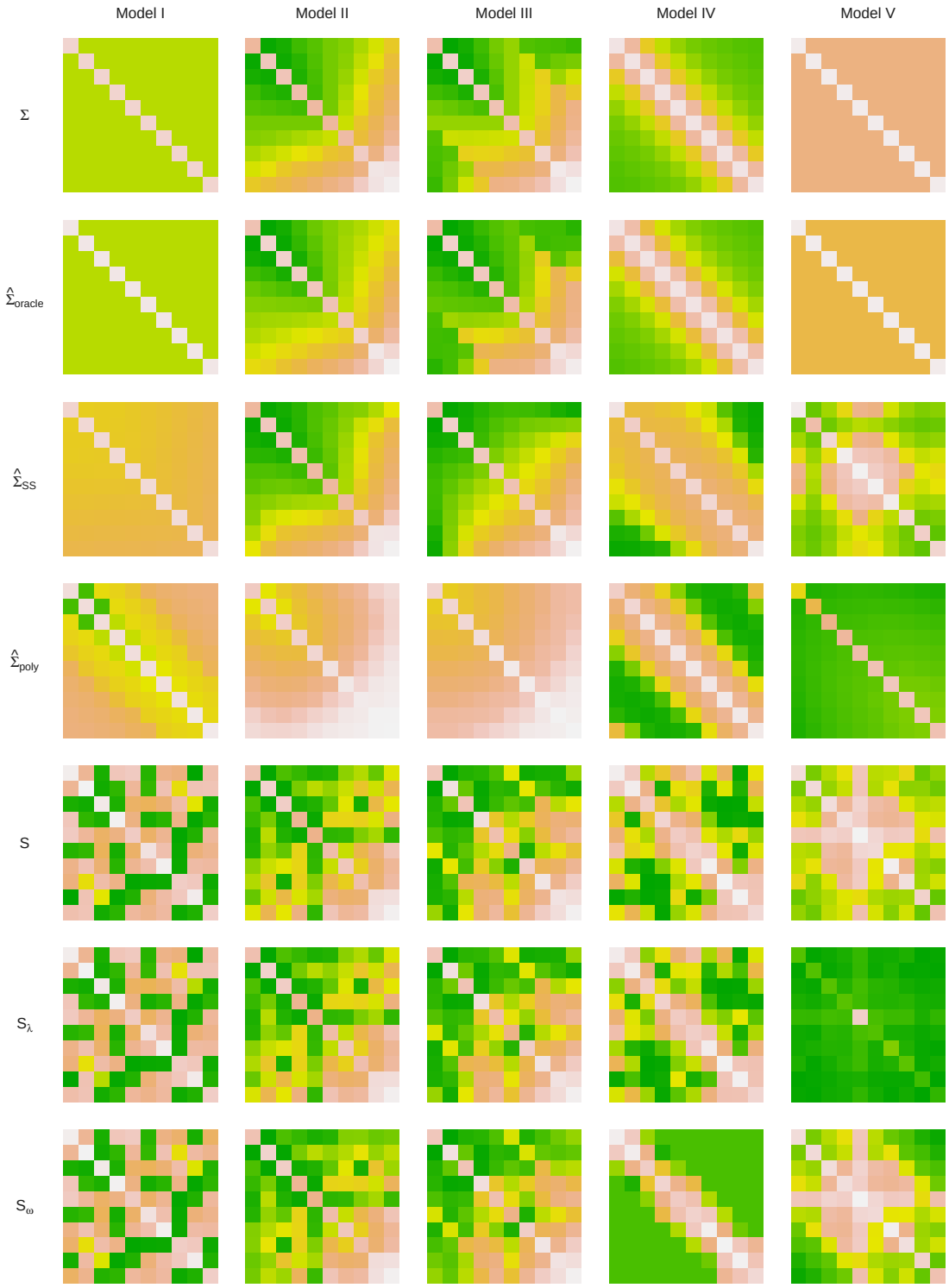


Figure 3: *Covariance Models I–V used for simulation and corresponding estimates with various methods. True covariance structures are shown in the first row followed by their estimates from the oracle estimator, smoothing spline ANOVA estimator, parametric polynomial estimator, the sample covariance matrix, the tapered sample covariance matrix, and the soft thresholding estimator.*

		p	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
Model I	$N = 50$	10	0.0135	0.0685	0.1102	1.2047	0.5369	1.1742
		20	0.0229	0.0834	0.1096	4.9850	1.3957	4.7796
		30	0.0196	0.1102	0.1127	12.5517	2.8019	11.3175
	$N = 100$	10	0.0105	0.0451	0.0531	0.5685	0.2045	0.5236
		20	0.0105	0.0425	0.0512	2.2831	0.5724	2.1358
		30	0.0139	0.0431	0.0472	5.2770	1.2430	4.9126
Model II	$N = 50$	10	0.0581	0.0689	4.7673	1.2832	1.4644	1.1770
		20	0.0439	0.0581	97.2334	5.1665	21.6407	39.3522
		30	0.0627	0.0811	153.9665	12.3582	55.3674	133.9980
	$N = 100$	10	0.0386	0.0457	4.7911	0.5812	0.8335	0.5628
		20	0.0269	0.0416	98.1989	2.3364	10.1841	10.0864
		30	0.0288	0.0367	158.2480	5.2389	33.5207	62.5030
Model III	$N = 50$	10	0.0619	0.3296	3.0108	1.2030	1.1460	1.1467
		20	0.0695	1.1100	62.7522	4.9824	17.2244	14.9189
		30	0.0576	2.3215	1091.1933	12.4792	49.9135	121.7795
	$N = 100$	10	0.0268	0.2904	3.0383	0.5699	0.5545	0.5371
		20	0.0275	1.1963	62.8960	2.2700	11.8274	9.5217
		30	0.0221	2.2811	1105.0449	5.2234	29.1693	60.3529
Model IV	$N = 50$	10	0.0217	0.3348	0.7144	1.2218	0.7397	1.1921
		20	0.0286	0.9177	1.4588	4.9091	1.9786	4.9206
		30	0.0283	1.5992	2.2173	12.6114	3.7440	12.1489
	$N = 100$	10	0.0125	0.3047	0.6958	0.5570	0.3168	0.5515
		20	0.0105	0.8911	1.4813	2.2659	0.9365	2.2474
		30	0.0134	1.5213	2.2228	5.2106	1.9312	5.2111
Model V	$N = 50$	10	0.0986	0.2769	1.2420	1.2023	18.5222	2.9824
		20	0.2512	0.7514	2.8557	5.0195	34.6618	13.8690
		30	0.2641	1.1776	4.5791	12.3460	46.5437	26.1364
	$N = 100$	10	0.0520	0.2416	1.1491	0.5821	16.4081	1.7397
		20	0.0827	0.7286	2.9080	2.2918	32.5295	5.4649
		30	0.1799	1.1813	4.4402	5.2197	39.2914	15.4295

Table 2: *Multivariate normal simulations for Models I–V. Estimated entropy risk is reported for the oracle estimator, our smoothing spline ANOVA estimator, the parametric polynomial estimator of Pan and MacKenzie (2003), the sample covariance matrix, the tapered sample covariance matrix, and the soft thresholding estimator.*

DATA ANALYSIS

Kenward (1987) reported an experiment designed to investigate the impact of the control of intestinal parasites in cattle. To compare two methods for controlling the disease, say treatment A and treatment B, each of 60 cattle was assigned randomly to two groups, each of size 30. Animal subjects were put out to pasture at the start of grazing season, with each member of the groups receiving one of the two treatments. Animals were weighed 11 times ($p = 11$) over a 133-day period; the first 10 measurements on each animal were made at two-week intervals and the final measurement was made one week later. The longitudinal dataset is balanced, as there were no missing observations for any of the experimental units. Observed weights are shown in Figure 4.

The analysis of the same dataset provided by Zimmerman and Núñez-Antón (1997) rejected equality of the two covariance matrices corresponding to treatment group using the classical likelihood ratio test, making it reasonable to study each treatment group's covariance matrix separately. Following Pourahmadi (1999), Zhang, Leng, and Tang (2015), and Pan and Pan (2017), we analyze the data from the cattle assigned to treatment group A ($N = 30$). Given that the animals belong to the same treatment group and share a common set of observation times, we posit common covariance matrix Σ for each subject. The left profile plot in Figure 4 of the weights for units in treatment group A shows a clear upward trend in weights. Variances appear to increase over time, suggesting that the covariance structure is nonstationary.

The nonstationarity suggested in Figure 4 is also supported by the sample correlations given in Table 3; correlations within the subdiagonals are not constant and increase over time, a secondary indication that a stationary covariance is not appropriate for the data.

As evident in Figure 4 with a trend in the observed weight trajectories, covariance estimation generally involves simultaneous modeling of mean trajectories. We model the observed

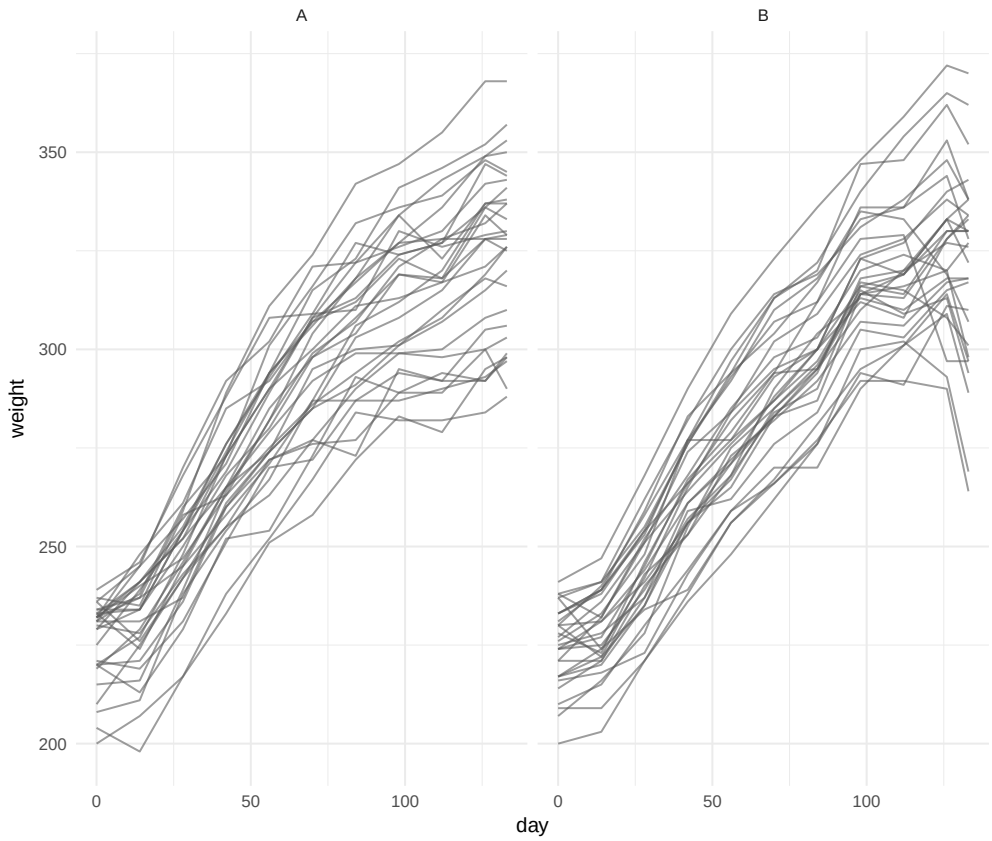


Figure 4: *Subject-specific weight curves over time for treatment groups A and B.*

trajectory for the i^{th} subject, Y_i, as

$$Y_i = f\left(\mathcal{T}_i\right) + \epsilon_i^*, \quad i = 1, \ldots, N,$$

where $f\left(\mathcal{T}_i\right) = \left(f(t_{i1}), \cdots, f(t_{i,p_i})\right)'$ is a vector of evaluation of a smooth function $f(t)$ that is common across the subjects at observed time points. For the cattle data, $\mathcal{T}_i = \{t_1 = 0, t_2 = 14, \cdots, t_{11} = 133\}$ same across the subjects. We assume that the measurement errors, ϵ_i^* follow $N\left(0, \Sigma\right)$. The mean trajectory was estimated by the sample mean of Y_i. Akin to conditionally linear mixed models, more refined mean modeling is possible for the data by allowing individual mean trajectories with a random intercept. For comparison with previous analyses, however, we assume the simple model for mean trajectories and focus on covariance modeling.

Analyzing the sample regressogram and sample innovation variogram, Pourahmadi (1999) suggested that both sample generalized autoregressive parameters and the logarithms of the

	Day										
	0	14	28	42	56	70	84	98	112	126	133
0	1.00										
14	0.82	1.00									
28	0.76	0.91	1.00								
42	0.65	0.86	0.93	1.00							
56	0.63	0.83	0.89	0.93	1.00						
70	0.58	0.75	0.85	0.90	0.94	1.00					
84	0.51	0.64	0.75	0.80	0.85	0.92	1.00				
98	0.52	0.68	0.77	0.82	0.88	0.93	0.92	1.00			
112	0.51	0.61	0.71	0.74	0.81	0.89	0.92	0.96	1.00		
120	0.46	0.59	0.69	0.70	0.77	0.85	0.86	0.94	0.96	1.00	
133	0.46	0.56	0.67	0.67	0.74	0.81	0.84	0.91	0.95	0.98	1.00

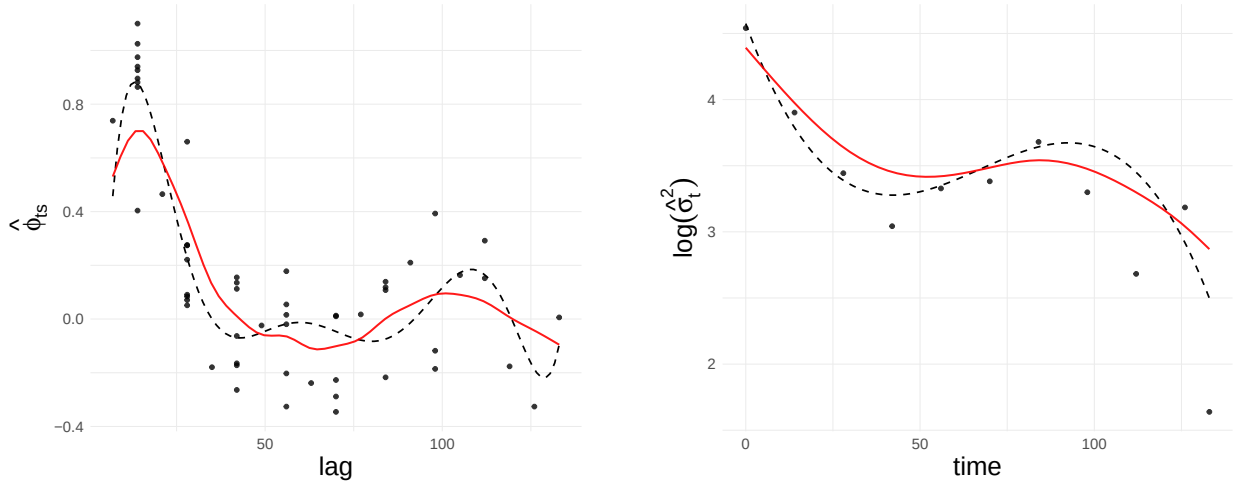
Table 3: *Cattle data: treatment group A sample correlations.*

innovation variances can be characterized in terms of low degree polynomials of the lag only and time, respectively. Pan and Pan (2017) had the same observation that the regressogram of empirical estimates of $\phi_{t,s}$ show consistent behaviour over $l = t - s$ for each value of t, indicating a lack of a strong functional component of m. They used the Bayesian information criterion (BIC) to choose the order of the polynomials for the generalized autoregressive parameters and innovation variances. The dashed lines in Figure 5 show the estimated polynomials according to the suggested model using the detrended trajectories with estimated means; polynomials of degree five and three were selected for $\phi_{t,s}$ and $\log(\sigma_t^2)$, respectively.

To balance the consideration of previous analyses with the interest of entirely data-driven model specification, we take our approach to estimation of the autoregressive coefficient function ϕ using a two-way ANOVA model in a tensor product space $\mathcal{H} = \mathcal{H}^{[l]} \otimes \mathcal{H}^{[m]}$, where penalties for $\mathcal{H}^{[l]}$ and $\mathcal{H}^{[m]}$ are specified to induce cubic splines for both of the marginal subspaces corresponding to l and m. We refine the approach with a two-way ANOVA model by introducing rescaling parameters for the nonparametric components of ϕ_1, ϕ_2 and ϕ_{12} and tuning the scale of ϕ_2 and ϕ_{12} components relative to the lag component ϕ_1. For the innovation variance function, we consider the same model space as the marginal function spaces for the coefficient function. For selection of smoothing parameters, we used cross validation: losoCV for estimation of the coefficient function and GCV for the innovation

variance function. Figures 5 and 6 show the estimated autoregressive coefficient function and log innovation variance function using our approach.

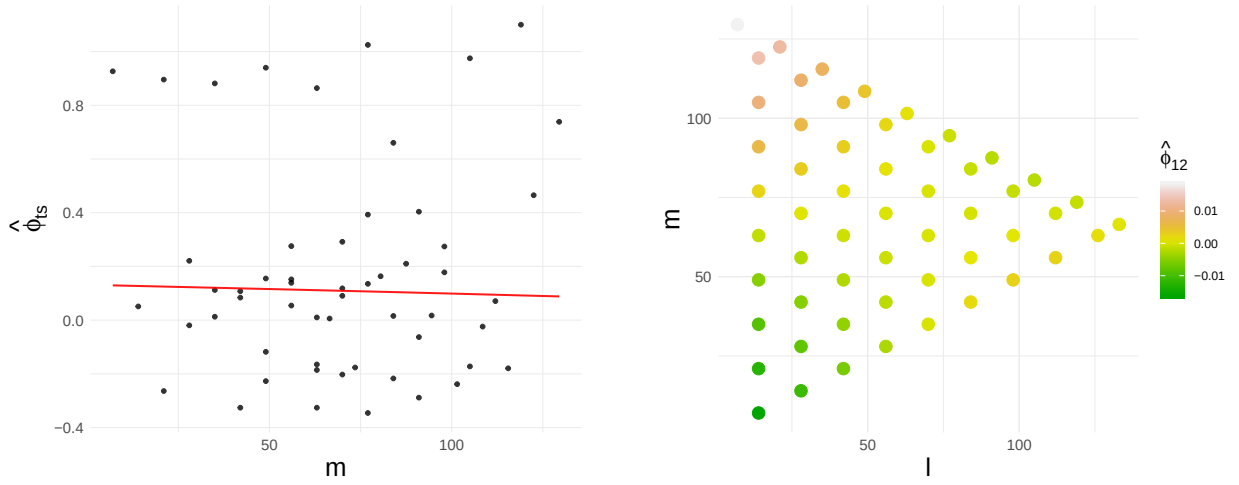
The solid lines in Figure 5 are the estimated main effect of lag, $\phi_1(l)$, including the overall mean $\hat{\mu}$ and log innovation variance function $\log(\sigma^2(t))$. Figure 6a shows the estimated main effect, $\phi_2(m)$, in the additive direction including $\hat{\mu}$. The result confirms that the m component is negligible as in the previous analyses. Further, Figure 6b displays the estimated two-way interaction between lag and additive direction at the sample points (l, m) defined by the observed times. The estimated interaction captures the pattern that given the same lag, generalized autoregressive parameters tend to be larger with large m than small m, which is more visible in the range of small values of lag. However, the size of interaction is minuscule. The estimated functions in Figure 6 are largely parametric resulting from rescaling of the nonparametric components of ϕ_2 and ϕ_{12}.



(a) *Sample generalized autoregressive parameters* $\hat{\phi}_{t,s}$

(b) *Sample log innovation variances* $\log(\hat{\sigma}_t^2)$

Figure 5: *Sample regressogram and log innovation variances for the cattle data from treatment group A. The dashed line in (a) is the polynomial fit (degree 5) and the solid line is the estimated main effect of* $l = t - s$ *of the cubic smoothing spline fit. The dashed line superimposed over the log innovation variances in (b) is the polynomial fit (degree 3), and the solid line is the cubic smoothing spline fit.*



(a) *Estimate of* $\phi_2(m)$ (b) *Estimate of* $\phi_{12}(l, m)$

Figure 6: *The estimated main effect of additive direction* $m = (t + s)/2$ *superimposed over the sample regressogram for the cattle data from treatment group A and the estimated interaction between lag (l) and additive direction (m) evaluated on the grid defined by the observed time points.*

The estimated generalized autoregressive coefficient function was evaluated at pairs of observation times, and the size of the functional components was measured in terms of the squared vector norm of each component evaluated at the sample points. The squared norm of the main effect of lag (4.826) was much larger than that of the main effect of additive direction (0.005) or the interaction term (0.001), which is clearly indicated by Figures 5 and 6. The size of the functional components indicates a certain degree of concordance with the models proposed by Pourahmadi (1999). This suggests that parameterizing ϕ as a univariate function of lag only is a reasonable modeling choice.

CONCLUSIONS

We have proposed a general nonparametric framework for covariance estimation with longitudinal data. The Cholesky decomposition supplies a reparameterization of the covariance matrix allowing for unconstrained estimation. The elements of the reparameterization can be interpreted as parameters for an autoregressive model. We allow for irregular, subject-specific

time points by extending this regression model to a functional varying coefficient model. By reframing covariance estimation as the estimation of the functional varying coefficient function and the error variance function, our approach leverages regularization techniques that are typically reserved for function estimation. A functional ANOVA model leads to an interpretable decomposition of the varying coefficient into its stationary and non-stationary functional components. This parameterization naturally allows for shrinkage of estimated covariances toward those corresponding to stationary models.

Coupling the form of penalty functional with desired simplicity in the dependence structure is the key to successful applications of the proposed framework for covariance estimation. In addition to the notion of stationarity this paper has focused on, the proposed approach can be applied to other forms of parsimony such as independence, short-term dependence, and diminishing dependence with lags. While appropriate forms of penalty functionals for modeling short-term dependence or diminishing dependence are not obvious, they may be dealt with using alternative representations for the lag component of the autoregressive coefficient function tailored to right-truncated functions or monotone functions. In practice, we suggest to choose the appropriate class of null models for the varying coefficient function in a data-driven manner following a careful examination of the observed dependence through sample regressograms.

FUNDING INFORMATION

This research was supported in part by the National Science Foundation under grant DMS-15-13566.

ACKNOWLEDGEMENTS

We thank Grace Wahba for introducing smoothing spline ANOVA models to statistics, an inspiration for this work, and congratulate Grace on her retirement in 2018. We also thank Chong Gu for creating the `gss` R package that implements smoothing spline ANOVA models. Our implementation is primarily based on functions in the package.

APPENDIX

Proof of Theorem 1. The function space $\mathcal{H}$ is decomposed into $\mathcal{H}_0$ and $\mathcal{H}_1$ with the penalty functional $J(\phi)$. $\mathcal{H}_1$ can be further decomposed into the finite dimensional subspace spanned by $\{K_1(\boldsymbol{v}_j, \boldsymbol{v})\}$, $j = 1, \ldots, |V|$ and its orthogonal complement in $\mathcal{H}_1$. Considering the three subspaces, any $\phi \in \mathcal{H}$ can be written as

$$\phi(\boldsymbol{v}) = \sum_{i=1}^{\mathcal{N}_0} d_i \nu_i(\boldsymbol{v}) + \sum_{j=1}^{|V|} c_j K_1(\boldsymbol{v}_j, \boldsymbol{v}) + \rho(\boldsymbol{v}), \tag{15}$$

where $\rho \in \mathcal{H}_1$ is perpendicular to $\nu_1, \ldots, \nu_{\mathcal{N}_0}$ and $K_1(\boldsymbol{v}_j, \boldsymbol{v})$ for each $\boldsymbol{v}_j \in V$.

Using the properties of the reproducing kernel $K = K_0 + K_1$, we can show that evaluation of any $\phi \in \mathcal{H}$ at $\boldsymbol{v}_\ell \in V$ does not depend on ρ:

$$\begin{aligned}
\phi(\boldsymbol{v}_\ell) &= \langle \phi(\cdot), K(\boldsymbol{v}_\ell, \cdot) \rangle_{\mathcal{H}} \\
&= \langle \sum_{i=1}^{\mathcal{N}_0} d_i \nu_i(\cdot) + \sum_{j=1}^{|V|} c_j K_1(\boldsymbol{v}_j, \cdot) + \rho(\cdot), K(\boldsymbol{v}_\ell, \cdot) \rangle_{\mathcal{H}} \\
&= \langle \sum_{i=1}^{\mathcal{N}_0} d_i \nu_i(\cdot) + \sum_{j=1}^{|V|} c_j K_1(\boldsymbol{v}_j, \cdot) + \rho(\cdot), K_0(\boldsymbol{v}_\ell, \cdot) \rangle_{\mathcal{H}} \\
&\qquad + \langle \sum_{i=1}^{\mathcal{N}_0} d_i \nu_i(\cdot) + \sum_{j=1}^{|V|} c_j K_1(\boldsymbol{v}_j, \cdot) + \rho(\cdot), K_1(\boldsymbol{v}_\ell, \cdot) \rangle_{\mathcal{H}} \\
&= \langle \sum_{i=1}^{\mathcal{N}_0} d_i \nu_i(\cdot), K_0(\boldsymbol{v}_\ell, \cdot) \rangle_{\mathcal{H}} + \langle \sum_{j=1}^{|V|} c_j K_1(\boldsymbol{v}_j, \cdot), K_1(\boldsymbol{v}_\ell, \cdot) \rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{\mathcal{N}_0} d_i \nu_i(\boldsymbol{v}_\ell) + \sum_{j=1}^{|V|} c_j K_1(\boldsymbol{v}_j, \boldsymbol{v}_\ell).
\end{aligned}$$

The last two equalities result from the orthogonality of $\mathcal{H}_0$, $\{K_1(\boldsymbol{v}_j, \boldsymbol{v})\}$, and ρ, and the reproducing property of K. Thus, the negative log likelihood in (7) depends only on $\sum_{i=1}^{\mathcal{N}_0} d_i \nu_i(\boldsymbol{v}) + \sum_{j=1}^{|V|} c_j K_1(\boldsymbol{v}_j, \boldsymbol{v})$. On the other hand, the penalty is given by

$$\|P_1 \phi\|^2 = \left\| \sum_{j=1}^{|V|} c_j K_1(\boldsymbol{v}_j, \cdot) + \rho(\cdot) \right\|^2 = \left\| \sum_{j=1}^{|V|} c_j K_1(\boldsymbol{v}_j, \cdot) \right\|^2 + \|\rho(\cdot)\|^2.$$

The penalized negative log likelihood is obviously minimized when $\|\rho\|^2 = 0$, or $\rho(\cdot) = 0$. This leads to the form of the minimizer for ϕ_λ as stated in Theorem 1.

□

References

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis*. Wiley.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, *68*(3), 337–404.

Berlinet, A., & Thomas-Agnan, C. (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

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

Blake, T. A. (2018). *Nonparametric covariance estimation for longitudinal data* (Unpublished doctoral dissertation). The Ohio State University.

Cai, T. T., Zhang, C.-H., & Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, *38*(4), 2118–2144.

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science*, 89–102.

Fang, Y., Wang, B., & Feng, Y. (2016). Tuning-parameter selection in regularized estimations of large covariance matrices. *Journal of Statistical Computation and Simulation*, *86*(3), 494–509.

Gu, C. (2013). *Smoothing spline ANOVA models* (Vol. 297). Springer Science & Business Media.

Gu, C., & Wahba, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM Journal on Scientific and Statistical Computing*, *12*(2), 383–398.

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

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

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 295–327.

Kenward, M. G. (1987). A method for comparing profiles of repeated measurements. *Applied Statistics*, 296–308.

Leng, C., Zhang, W., & Pan, J. (2010). Semiparametric mean–covariance regression analysis for longitudinal data. *Journal of the American Statistical Association*, *105*(489), 181–193.

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

Lin, S. P. (1985). A Monte Carlo comparison of four estimators for a covariance matrix. *Multivariate Analysis*, *6*, 411–429.

Liu, A., & Wang, Y. (2004). Hypothesis testing in smoothing spline models. *Journal of Statistical Computation and Simulation*, *74*(8), 581–597.

Pan, J., & Mackenzie, G. (2003). On modelling mean-covariance structures in longitudinal studies. *Biometrika*, *90*(1), 239–244.

Pan, J., & Pan, Y. (2017). jmcm: An R package for joint mean-covariance modeling of longitudinal data. *Journal of Statistical Software*, *82*(1), 1–29.

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

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, 425–435.

Pourahmadi, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statistical Science*, 369–387.

Pourahmadi, M., & Daniels, M. (2002). Dynamic conditionally linear mixed models for longitudinal data. *Biometrics*, *58*(1), 225–231.

Rothman, A. J., Levina, E., & Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, *104*(485), 177–186.

Wahba, G. (1990). *Spline models for observational data* (Vol. 59). SIAM.

Wahba, G., Wang, Y., Gu, C., Klein, R., & Klein, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. *The Annals of Statistics*, 1865–1895.

Wang, Y. (1997). GRKPACK: fitting smoothing spline ANOVA models for exponential

families. *Communications in Statistics - Simulation and Computation*, *26*(2), 765–782.

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

Xu, G., & Huang, J. Z. (2012). Asymptotic optimality and efficient computation of the leave-subject-out cross-validation. *The Annals of Statistics*, *40*(6), 3003–3030.

Yao, F., Müller, H.-G., & Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, *100*(470), 577–590.

Zhang, W., Leng, C., & Tang, C. Y. (2015). A joint modelling approach for longitudinal studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *77*(1), 219–238.

Zimmerman, D. L., & Núñez-Antón, V. (1997). Structured antedependence models for longitudinal data. In *Modelling longitudinal and spatially correlated data* (pp. 63–76). Springer.