Nonparametric covariance estimation with shrinkage toward stationary models

Tayler A. Blake* Yoonkyung Lee*

Article Type:

Focus Article

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.

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

GRAPHICAL TABLE OF CONTENTS

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 the 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); Cai, Zhang, Zhou, et al. (2010); Yao, Müller, and Wang (2005); Rothman, Levina, and Zhu (2009). 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.

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 Section 2, we review the role of the modified Cholesky decomposition in the unconstrained reparametrization of a covariance matrix. In Section 3, we present a functional varying coefficient model for the elements of the reparameterized covariance matrix and propose reproducing kernel Hilbert space framework for estimation of the varying coefficient function. Section 4 outlines the estimation of the innovation variances . Section 4 [Comment: Need an introdcution to what this paper is about and the outline of the paper here]

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 co-

efficients 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 $Var(\epsilon_t) = \sigma_t^2$ 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 $D = \text{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_p)'$, then the sequence of regression models in (2) can be written in matrix form as

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

The (t, j) entry of T is $-\phi_{tj}$, and $\sigma_t^2 = Var(\epsilon_t)$. 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 variances $\log \sigma_t^2$ are unconstrained. The regression coefficients of the model in (2) 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.

[Comment: need a sentence for transition from this section to next]

Llarge 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.

To estimate $\phi(t, s)$ and $\sigma^2(t)$, we employ the smoothing spline framework of Kimeldorf and Wahba (1971) and adopt the smoothing spline analogue of the classical analysis of variance (ANOVA) model proposed by Gu (2013). Smoothing spline ANOVA models are rooted in the theory of reproducing kernel Hilbert spaces (Aronszajn, 1950; Wahba, 1990; Berlinet & Thomas-Agnan, 2011). 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).

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

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

where the prediction errors $\epsilon(t)$ follow a mean zero Gaussian process with variance function $\sigma^2(t)$.

[Comment: I think the rest of this section needs to be reorganized and rewritten. Motivation for SS framework should come before the following paragraph.]

Rather than modelling $\tilde{\phi}$ directly, we reparameterize the varying coefficient function so that the fitted function can easily be used for suggesting parsimonious or structured models for the Cholesky decomposition. Specifically, we take stationarity as a form of parsimony in covariance models, including those parameterizing the elements of T as a function of the

lag between observations (Leng, Zhang, & Pan, 2010; Pan & Mackenzie, 2003; Pourahmadi, 1999; Pourahmadi & Daniels, 2002). To facilitate such model specification, we transform inputs from a pair of time points (t, s) with $t > s$ to the lag, $l = t - s$, and additive direction, $m = \frac{t+s}{2}$, and model

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

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 model simplicity, 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.

REPRODUCING KERNEL HILBERT SPACES

A Hilbert space $\mathcal{H}$ of functions on a set χ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is defined as a complete inner product linear space. For each $x \in \chi$, let $[x]$ map $f \in \mathcal{H}$ to $f(x) \in \Re$, which is known as the evaluation functional at x. A Hilbert space is called a reproducing kernel Hilbert space if the evaluation functional $[x] f = f(x)$ is continuous in $\mathcal{H}$ for all $x \in \chi$. 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 $K(x_1, x_2) = K_{x_2}(x_1) = \langle K_{x_1}, K_{x_2} \rangle_{\mathcal{H}}$ is called the reproducing kernel (RK) of $\mathcal{H}$. The RK satisfies that for every $x \in \chi$ and $f \in \mathcal{H}$, $K(\cdot, x) \in \mathcal{H}$, and $f(x) = \langle f, K(\cdot, x) \rangle_{\mathcal{H}}$. 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 kernel satisfies that for any $\{x_1, \ldots, x_{n_1}\}$, $\{u_1, \ldots, u_{n_2}\} \in \chi$ and $\{a_1, \ldots, a_{n_1}\}$, $\{b_1, \ldots, b_{n_2}\} \in \Re$,

$$\langle \sum_{i=1}^{n_1} a_i K(\cdot, x_i), \sum_{j=1}^{n_2} b_j K(\cdot, u_j) \rangle_{\mathcal{H}} = \sum_i \sum_j a_i b_j K(x_i, u_j). \tag{6}$$

The representer of any bounded linear functional can be obtained from the reproducing kernel K.

ESTIMATION OF THE GENERALIZED VARYING COEFFICIENT FUNCTION VIA BIVARIATE SMOOTHING

We let ϕ belong to a reproducing kernel Hilbert space $\mathcal{H}$ with reproducing kernel K. Define $\boldsymbol{v}_{ijk} = \left(t_{ij} - t_{ik}, \frac{1}{2}\left(t_{ij} + t_{ik}\right)\right) = \left(l_{ijk}, m_{ijk}\right)$, $\boldsymbol{v}_{ijk} \in \mathcal{V} = [0,1]^2$ as the tuple corresponding to the transformed pair of j^{th} and k^{th} observation times on the i^{th} subject. Fixing the innovation variances $\sigma_{ij}^2 = \sigma^2\left(t_{ij}\right)$, we take the estimator of ϕ to be the minimizer of the penalized negative log likelihood:

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

The penalty functional $J\left(\phi\right) = ||P_1\phi||^2$, which measures the complexity of ϕ, can be written as the squared norm of $P_1\phi$, the projection of ϕ onto a subspace $\mathcal{H}_1$. $||\cdot||$ denotes the norm in $\mathcal{H}$, and the smoothing parameter λ controls the tradeoff between the goodness of fit measure ℓ and the penalty $||P_1\phi||^2$.

Let $V = \bigcup_{i,j,k} \left\{\boldsymbol{v}_{ijk}\right\} \equiv \left\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{|V|}\right\}$ denote the set of unique within-subject pairs of observation times when pooled across N subjects. 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 0.1 *Let* $\{\nu_1, \ldots, \nu_{\mathcal{N}_0}\}$ *span* $\mathcal{H}_0 = \{\phi \in \mathcal{H} : J\left(\phi\right) = 0\}$, *the null space of* $J\left(\phi\right) = ||P_1\phi||^2$. *Then the minimizer* ϕ_λ *of* (7) *is given by*

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

where $K_1\left(\boldsymbol{v}_j, \boldsymbol{v}\right)$ *denotes the reproducing kernel for* $\mathcal{H}_1$ *evaluated at* $\boldsymbol{v}_j$, *the* j^{th} *element of* V, *viewed as a function of* $\boldsymbol{v}$.

The proof is left to the Appendix.

MODEL FITTING

Let $Y = \left(Y_1^{(-1)'}, Y_2^{(-1)'}, \ldots, Y_N^{(-1)'}\right)'$ denote the vector of length $n_Y = \sum_i p_i - N$, constructed by stacking the N observed response vectors, less their first element: $Y_i^{(-1)'} = (y_{i2}, \ldots, y_{in_i})'$. Define X_i to be the $(p_i - 1) \times |V|$ matrix containing the covariates for subject i necessary for regressing each measurement $y_{i2}, \ldots, y_{i,p_i}$ 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) element equal to $\nu_j(\boldsymbol{v}_i)$. 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 Representer Theorem (8), the penalized negative log likelihood in (7) is given by

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

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

$$\begin{bmatrix} \tilde{B}'\tilde{B} & \tilde{B}'\tilde{K}_{_V} \\ \tilde{K}_{_V}'\tilde{B} & \tilde{K}_{_V}'\tilde{K}_{_V} + \lambda K_{_V} \end{bmatrix} \begin{bmatrix} d \\ c \end{bmatrix} = \begin{bmatrix} \tilde{B}'\tilde{Y} \\ \tilde{K}_{_V}'\tilde{Y} \end{bmatrix}. \quad (10)$$

When $\tilde{K}_{_V}$ is full column rank, the solution to (10) is given by $\begin{bmatrix} \hat{d}' & \hat{c}' \end{bmatrix}' = C^{-1}(C')^{-1}\begin{bmatrix} \tilde{B} & \tilde{K}_{_V} \end{bmatrix}'\tilde{Y}$, where

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

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, et al., 2012).

SIMULATION STUDIES

In this section we compare our bivariate spline estimator to other methods of covariance estimation. 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 $\tilde{\phi}$ can be parameterized as a (univariate) polynomial in $l = t - s$. 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:

Model I: $\Sigma = I$ (*the identity matrix.*)

Model II: $\Sigma^{-1} = T'D^{-1}T$, where $D = 0.01 \times I$, and $T = \left[-\tilde{\phi}\left(t_i, t_j\right)\right]$ with $\tilde{\phi}\left(t_i, t_i\right) = 1$, $\tilde{\phi}\left(t_i, t_j\right) = t_i - \frac{1}{2}$ for $0 \leq t_j < t_i \leq 1$, $\tilde{\phi}\left(t_i, t_j\right) = 0$ otherwise.

Model III: $\Sigma^{-1} = T'D^{-1}T$, where $D = 0.01 \times I$, and $T = \left[-\tilde{\phi}\left(t_i, t_j\right)\right]$ with $\tilde{\phi}\left(t_i, t_i\right) = 1$, $\tilde{\phi}\left(t_i, t_j\right) = t_i - \frac{1}{2}$ for $0 \leq t_i - t_j \leq \frac{1}{2}$, $\tilde{\phi}\left(t_i, t_j\right) = 0$ otherwise.

Model IV: $\Sigma = [\sigma_{ij}]$, $\sigma_{ij} = \left(1 + \frac{(t_i - t_j)^2}{2k^2}\right)$ where $k = 0.6$, $0 < t_i, t_j < 1$ (*the rational quadratic model.*)

Model V: $\Sigma = T^{-1'}DT^{-1} = \rho \mathrm{J} + (1 - \rho)\,\mathrm{I}$, $\rho = 0.7$, where $D = \mathrm{diag}\left(\sigma_t^2, \ldots, \sigma_p^2\right)$ with $\sigma_t^2 = 1 - \frac{(t-2)\rho^2}{1+(t-2)\rho}$, $t = 2, \ldots, p$, $T = \left[-\tilde{\phi}\left(t, s\right)\right]$ with $\tilde{\phi}\left(t, t\right) = 1$, $\tilde{\phi}\left(t, s\right) = 1 - \frac{(t-2)\rho^2}{1+(t-2)\rho}$ for $t = 2, \ldots, p$, $\tilde{\phi}\left(t, s\right) = 0$ otherwise (*the compound symmetry model.*)

The two-dimensional surfaces corresponding to each generating model are shown left to right in Figure 1. The first row displays the surface coinciding with the appropriate discrete covariance matrix, and the second row displays the surfaces of the corresponding Cholesky factors.

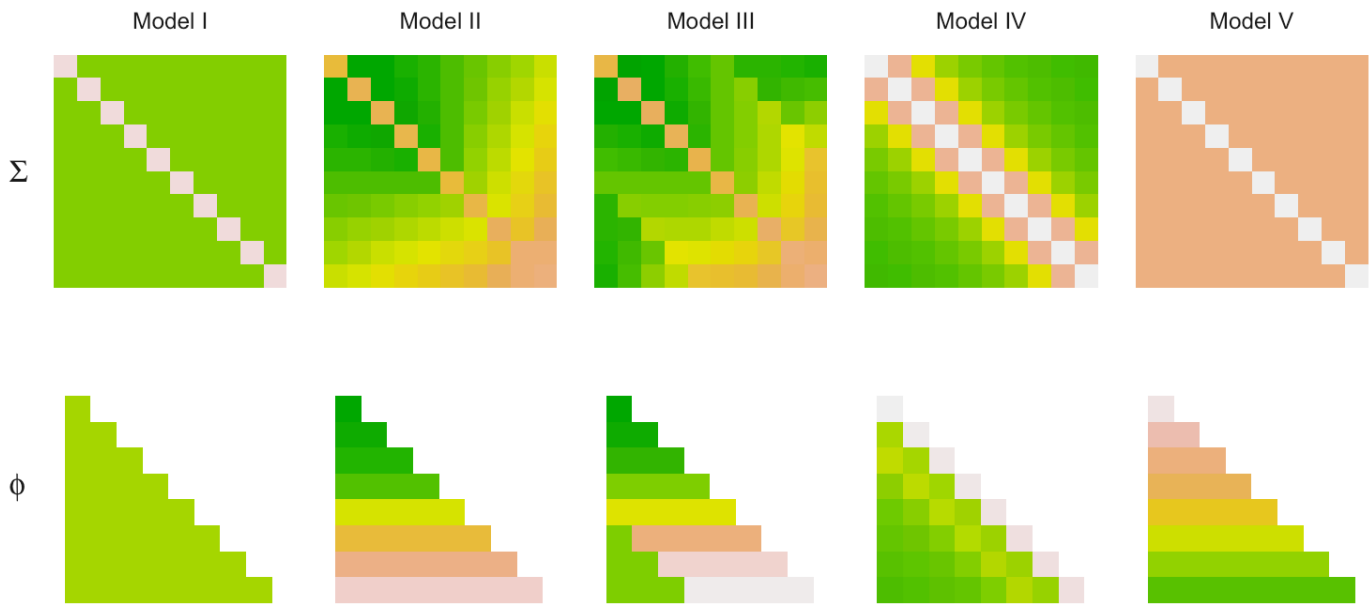


Figure 1: *Heatmaps of the true covariance matrices corresponding to Model I - Model V and ϕ defining the corresponding Cholesky factor T. The smallest elements of each matrix correspond to dark green pixels; the light pink (white) pixels correspond to the large (largest) elements of the matrix.*

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

$$R\left(\Sigma, \hat{\Sigma}\right) = E_{\Sigma}\left[\Delta\left(\Sigma, \hat{\Sigma}\right)\right],$$

which we approximate via Monte Carlo simulation. For each combination of $p = 10, 20, 30$ and sample size $N = 50, 100$, we construct an estimate from each of 100 samples from a mean

zero p-dimensional multivariate Normal distribution with covariance matrix $\Sigma = T^{-1}DT'^{-1}$ and calculate the corresponding loss. Construction of the sample covariance matrix S and regularized variants S^{ω} and S^{λ} requires an equal number of observations on each subject taken at a common set of observation times, so 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 $\sigma^2(t)$.

Results are presented in Table 1. 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 can be found in the Appendix. For each simulation setting, the risk of the oracle estimator serves as a lower bound on the risk for the given covariance structure. 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. 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. The smoothing spline estimator performs the most poorly on Model III, where ϕ does not belong to the tensor product smoothing spline model space 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 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 with added flexibility over the polynomial 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

		p	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
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 1: *Multivariate normal simulations for Model I - Model 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 were 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 $p = 11$ times 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 2.

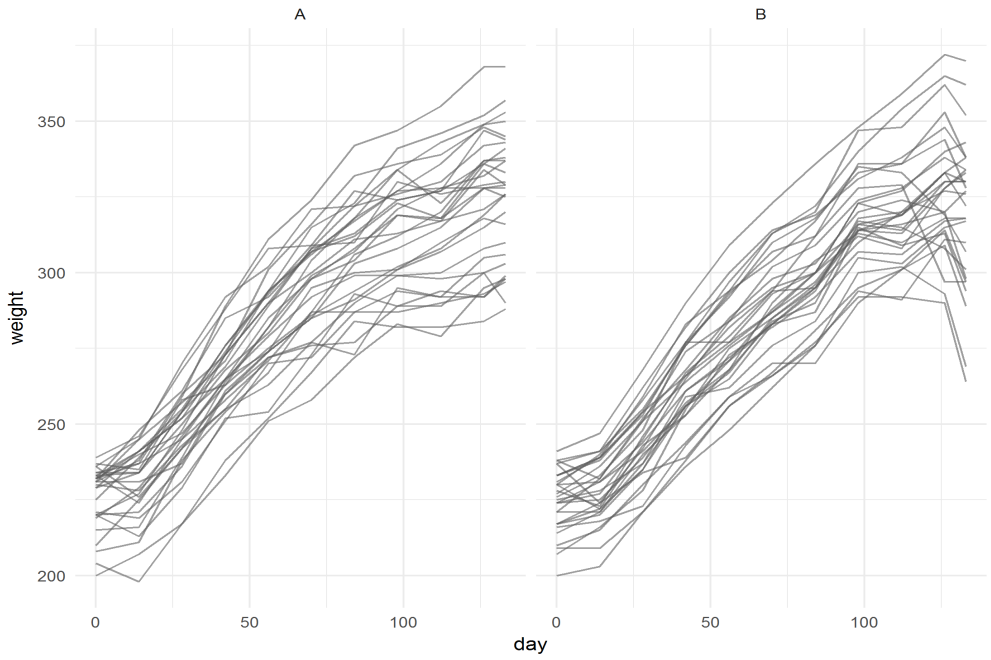


Figure 2: *Subject-specific weight curves over time for treatment groups A and B.*

The analysis of the same dataset provided by (Zimmerman & 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 (Pan & Pan, 2017), (Zhang, Leng, & Tang, 2015), and (Pourahmadi, 1999), we analyze the data from the $N = 30$ cattle assigned to treatment group A, which

we assume share a common 11×11 covariance matrix Σ. The left profile plot in Figure 2 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 2 is also supported by the sample correlations given in Table 2; correlations within the subdiagonals are not constant and increase over time, a secondary indication that a stationary covariance is not appropriate for the data.

	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 2: *Cattle data: treatment group A sample correlations.*

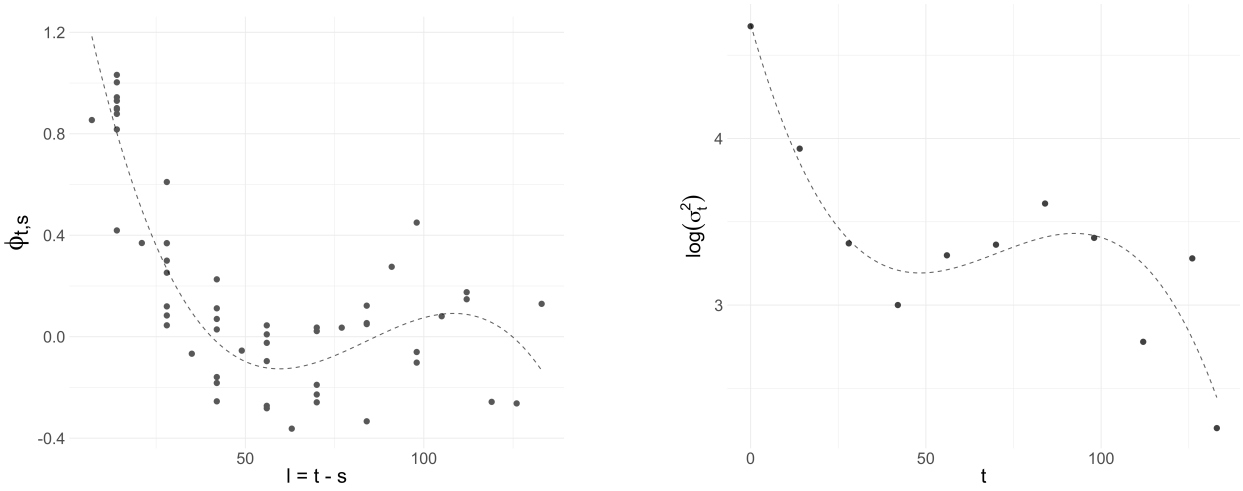
Analyzing the sample regressogram and sample innovation variogram, (Pourahmadi, 1999) suggested that both sample generalized autoregressive parameters and the logarithms of the innovation variances can be characterized in terms of cubic functions of the lag only. They model

$$\begin{aligned} \phi_{ts} &= x_{ts}'\beta, \\ \log\left(\sigma_t^2\right) &= z_t'\gamma, \end{aligned} \tag{11}$$

for $t = t_2, \ldots, t_{11}$ where

$$x_{ts}' = \begin{bmatrix} 1 & t-s & (t-s)^2 & (t-s)^3 \end{bmatrix}, \text{ and } z_t' = \begin{bmatrix} 1 & t & t^2 & t^3 \end{bmatrix}.$$

They estimate β and γ via maximum likelihood. Figure 3 shows the estimated cubic polynomials corresponding to Model (11).



(a) *Sample generalized autoregressive parameters* $\hat{\phi}_{ts}$.

(b) *Sample log innovation variances.*

Figure 3: *Cubic polynomomials fitted to the sample regressogram and log innovation variances for the cattle data from treatment group A.*

To model the mean weight trajectories, we adopt an approach akin to the dynamical conditionally linear mixed model presented in (Pourahmadi & Daniels, 2002):

$$Y_i = f(t_i) + Z_i b_i + \epsilon_i^*, \quad i = 1, \ldots, N, \tag{12}$$

where Y_i is the measurement corresponding response vector for the i^{th} subject, b_i is a $q \times 1$ vector of unknown random effects parameters, and Z_i is a known $p_i \times q$ design matrix. f is the smooth function of t, and $t_i = (t_{i1}, \ldots, t_{i,p_i})'$ is the $p_i \times 1$ vector of measurement times for subject i. We take $Z_i = (1, \ldots, 1)'$ so that the random effects α_i correspond to subject-specific shifts. We assume to that the random intercepts are independent and identically distributed $N(0, \sigma_\alpha^2)$. We assume that the $p_i \times 1$ vector of residuals $\epsilon_i^* \sim N(0, \Sigma_i)$ are mutually independent of the random intercepts α_i. Given that the animals belong to the same treatment group and share a common set of observation times, we assume each subject shares common covariance matrix $\Sigma_i = \Sigma$. We take the cubic smoothing spline

$f \in \mathcal{H} = \mathcal{C}^2 = \left\{f : \ f, \ f' \text{ absolutely continuous}, \ \int \left(f''(x)\right)^2 \ dx < \infty\right\}$, equipped with the inner product which corresponds to $J(f) = \int \left(f''(x)\right)^2 \ dx$. We take the estimators of f, $\alpha = (\alpha_1, \ldots, \alpha_N)'$ to minimize the penalized joint log likelihood

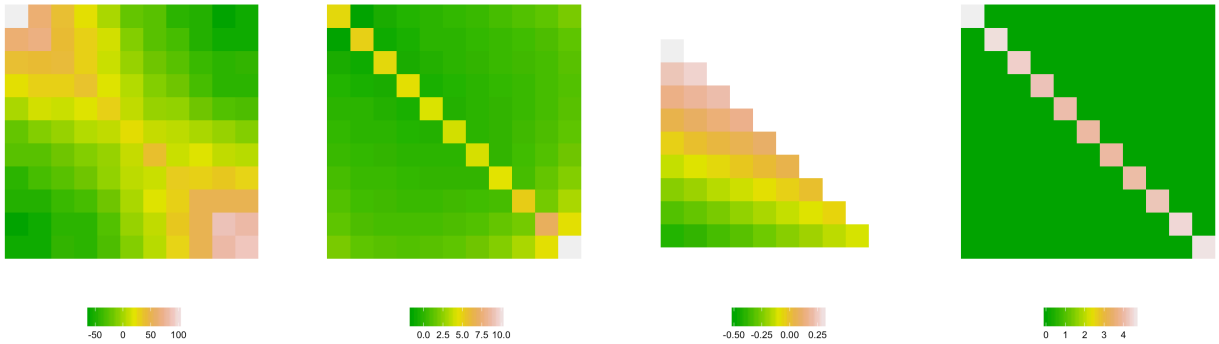
$$\sum_{i=1}^{N} \sum_{i=1}^{p_i} \left(y_{ij} - f(t_{ij}) - \alpha_i\right)^2 + \alpha' \Sigma_\alpha^{-1} \alpha + \lambda J(f), \tag{13}$$

where $\Sigma_\alpha = \sigma_\alpha^2 \mathrm{I}$.

(Pan & Pan, 2017) concluded that the regressogram of empirical estimates of $\tilde{\phi}_{t,s}$ show consistent behaviour over $l = t - s$ for each value of t, indicating a lack of a strong functional component of m. This is consistent Pourahmadi's choice (see ?, ?) in the specification of model (11) in terms of lag only. To balance the consideration of previous analyses with the interest of entirely data-driven model specification, we let $\phi \in \mathcal{H} = \mathcal{H}_{[l]} \otimes \mathcal{H}_{[m]}$, where

$$\mathcal{H}_{[l]} = \left\{\phi : \phi' = 0\right\} \oplus \left\{\phi : \phi(0) = \phi'(0) = 0; \ \int \phi''(l) \ dl < \infty\right\}$$
$$\mathcal{H}_{[m]} = \left\{\phi : \phi \propto 1\right\} \oplus \left\{\phi : \int_0^1 \phi(m) \ dm = 0, \ \int \phi''(m) \ dm < \infty\right\}$$

Figure 4 shows the estimated Cholesky surface and innovation variance function evaluated at $t = 0, 14, 28, \ldots, 112, 126, 133$ and the corresponding pairs of observation times (t, s), $0 \leq s < t \leq 133$. Figure 5 shows $\hat{\phi}$ decomposed into the functional components of its ANOVA decomposition.



(a) S (b) $\hat{\Sigma} = \hat{T}^{-1}\hat{D}\hat{T}'^{-1}$ (c) $\hat{\phi}(t,s)$ (d) $\log\hat{\sigma}^2(t)$

Figure 4: *The sample covariance matrix S, the estimated covariance matrix for the cattle weight data from treatment group A and the estimated Cholesky decomposition of the covariance matrix. The generalized autoregressive coefficient function $\phi(t,s)$ and the log innovation variances $\log\sigma^2(t)$ were estimated using a tensor product cubic spline and cubic spline, respectively. The fitted functions define the components of the Cholesky factor $\hat{T}$ and diagonal matrix $\hat{D}$.*

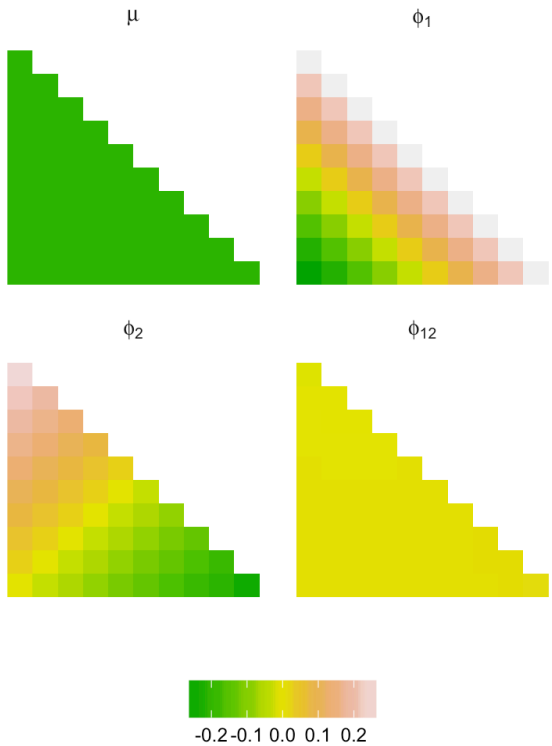


Figure 5: *Components of the SSANOVA decomposition of the estimated generalized autoregressive coefficient function ϕ evaluated on the grid defined by the observed time points.*

The size of the functional components (in terms of the squared functional norm) indicate a certain degree of concordance with the models proposed by (Pourahmadi, 1999). The squared norm of the main effect of l (1.914) is over twice that of the main effect of m (0.790). The squared norm of the interaction term, as clearly indicated by Figure 5, is negligible in comparison to the main effects, which suggests that parameterizing ϕ as a univariate function of lag l is a reasonable modeling choice.

CONCLUSIONS

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.

Cai, T. T., Zhang, C.-H., Zhou, H. H., et al. (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.

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.

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.

Kimeldorf, G., & Wahba, G. (1971). Some results on Tchebycheffian spline functions. Journal of Mathematical Analysis and Applications, 33(1), 82–95.

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.

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.

Xu, G., Huang, J. Z., et al. (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.