Nonparametric Covariance Estimation for Longitudinal Data

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Tayler A. Blake

Graduate Program in Department of Statistics

The Ohio State University

2018

Dissertation Committee:

Yoonkyung Lee, Advisor

Catherine A. Calder

Sebastian Kurtek

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 propose a general framework for nonparametric covariance estimation for longitudinal data where the variables have a natural ordering. Modeling the Cholesky decomposition of the covariance matrix removes constraints from estimation, including those posed by positive definiteness. In addition, the Cholesky decomposition enjoys the added advantage over alternative matrix decompositions by supplying a meaningful statistical interpretation of the corresponding estimated parameters. We illustrate the equivalence of covariance estimation and the estimation of a varying coefficient autoregressive model. By defining the varying coefficient as a bivariate function, we naturally accommodate sparsely or irregularly sampled longitudinal data without the need for imputation.

This framework extends the set of tools available for covariance estimation to any of those employed in the typical function estimation setting. Viewing stationarity as a form of simplicity or parsimony in covariance models, we specify the varying coefficient as a function so that we can conveniently penalize the components capturing the nonstationarity in the fitted function. Casting covariance estimation as bivariate smoothing problem, we demonstrate construction of a covariance estimator using the smoothing spline framework and a penalized B-spline expansion. A simulation study

establishes the advantage of our estimator over alternative estimators proposed in this setting. We analyze a longitudinal dataset to illustrate application of the methodology and compare our estimates to those resulting from alternative models proposed for the covariance for longitudinal data.

This one is for all the ones I love: my family, Todd and my dearest friends. You make me the best version of myself.

Acknowledgments

Far more people have had a contributing impact on this dissertation than I could possibly acknowledge in a few pages. The research that follows is as much the culmination of the oblique but persistent influence of the supporting cast of characters as it is the culmination of the core protagonists' focused efforts. Despite the incomplete list to follow, I must express my deepest gratitude for the village that has shaped me as a person and as a thinker over the years during which this work was completed.

I want to thank my family and Todd for their patience and support throughout the duration of this excursion. Everyone should be as lucky as I am to have people in her life who empower her to forge her own path, and to be different if different is what she needs to be.

I thank my advisor, Yoon Lee, for her both exceptional and patient academic guidance. Her ability to balance succinct, economical expression of thought with acute attention to detail holds my true admiration. I owe much of the clarity in the exposition of this work to her. I also want to extend my gratitude to the other members of my committee, Kate Calder and Sebastian Kurtek for their time and thoughtful feedback. I owe much of the conclusion of this work to Kate, whose transparent and honest advice provided perspective at a time when I most needed it.

I have learned as much from my fellow graduate students as anyone else at Ohio State. I feel incredibly lucky to have been a part of a program in which the sense of competition between the students is replaced with a sense of community and teamwork. There is a warm, fuzzy spot in my

heart for you guys and the slurry of Cockins Hall memories, wonderful and horrible and everything in between.

My most sincere appreciation goes to Chris Holloman and my team at ICC. It is a rare opportunity to work with a group of individuals who make work as much of an opportunity to have fun it is an opportunity for growth. They have made simultaneously working while finishing this paper as easy as it could possibly be. You guys are the best.

This research is supported in part by the National Science Foundation under grants DMS-15-13566 and DMS-16-13110.

Vita

May 2007	B.S. Mathematics and Computer Science, Capital University, Columbus, OH
May 2007	Winner, James L. and E. Marlene Bruning Award for Undergraduate Research, Capital University, Columbus, OH
December 2009	M.S. Statistics, The Ohio State University, Columbus, OH
May 2014 - December 2015	Data Scientist, Starbucks Coffee Co. Seattle, WA
January 2016 - December 2016	Machine Learning Engineer, Pillar Technology Inc. Columbus, OH
January 2017 - present	Senior Data Scientist, Information Control Company, Columbus, OH

Fields of Study

Major Field: Statistics

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The covariance matrix is the simplest summary statistic characterizing the dependence among a set of variables. An estimate of the covariance matrix or its inverse is required for nearly all statistical procedures in classical multivariate data analysis, time series analysis, spatial statistics and, more recently, the growing field of statistical learning. Covariance estimates play a critical role in the performance of techniques for clustering and classification such as linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), factor analysis, and principal components analysis (PCA), analysis of conditional independence through graphical models, classical multivariate regression, prediction, and Kriging. Covariance estimation for high dimensional data has recently gained growing interest, with less focus on inference and more attention on establishing consistent estimators when the sample size and number of parameters tend to infinity. While there is a bit of a gap between theory and practice in this area, much of the work currently being done is in an effort to fill it.

It is well recognized that there are two primary difficulties in modeling covariance matrices: high dimensionality and the positive-definiteness constraint. Prevalent technological advances in industry and many areas of science make high dimensional longitudinal and functional data a common occurrence, arising in numerous areas including medicine, public health, biology, and

environmental science with specific applications including fMRI, spectroscopic imaging, gene microarrays among many others. This influx of data presents a need for effective covariance estimation in high dimensions. However, high dimensional data can fall into two categories of sorts: the first is the case of functional data or times series data that one typically associates with "big p, small n", with each observation corresponding to a curve sampled densely at a fine grid of time points. On the other hand, in longitudinal studies, the measurement schedule could consist of targeted time points or could consist of completely arbitrary (random) time points. If the measurement schedule has targeted time points which are not necessarily equally spaced or if there is missing data, then we have what is considered incomplete and unbalanced data. If the measurement schedule has arbitrary or almost unique time points for every individual so that at a given time point there could be very few or even only a single measurement, we must consider how to handle what we consider as sparse longitudinal data. Thus, high dimensionality may be a consequence of irregular sampling schemes rather than having more measurements per subject than the number of subjects themselves.

A common way of reducing parameter dimensionality is to choose a simple parametric model to characterize the covariance structure. Particularly in the applied statistics literature, there is a propensity to characterize the dependency structure of the data by choosing a structured covariance matrix from a number of models on the menu offered from readily available software. Alternatively, the sample covariance matrix S is an unbiased estimator, but is known to be unstable in high dimensions. An extensive catalogue of methods have been developed to stabilize the naive estimator. Several have proposed shrinking the eigenvalues toward a central value; Stein's family of estimators which shrink the eigenvalues of S, but leave the eigenvectors untouched. Recent pursuit of sparsity, however, has lead to estimators that shrink also shrink the eigenvectors, or the sample covariance matrix itself toward sparse target structures such as diagonal or banded structures.

There has been a recent shift in covariance estimation toward regression-based approaches to eliminate the positive definite constraint from estimation procedures altogether. Principle components analysis and Gaussian graphical models among many others can be fit using regression models, the parameters of which are not constrained to maintain the positive definiteness of the final estimator. Germane to this idea is the approach of modeling various matrix decompositions directly, rather than the covariance matrix itself. The variance-correlation decomposition, spectral decomposition, and Cholesky decomposition are just a few examples of reparameterizations that dissolve the optimization constraints imposed by the positive definiteness requirement. The Cholesky decomposition in particular has recently received much attention because of its qualities that make it particularly attractive for its use in covariance estimation. The entries of the lower triangular matrix and the diagonal matrix from the modified Cholesky decomposition have interpretations as autoregressive coefficients and prediction variances when regressing a measurement on its predecessors. The unconstrained reparameterization and its statistical interpretability makes it easy to incorporate covariates in covariance modeling and to cast covariance modeling into the generalized linear model framework while guaranteeing that the resulting estimates are positive definite. This formulation sets one up to incorporate the arsenal of techniques for penalized regression for the challenging task of characterizing the dependency among a set of variables.

However, caution must be exercised when using generalizing linear models for the covariance of unbalanced data; direct application of much of the previous work in this particular area requires complete, balanced longitudinal data. When using covariates to model the covariance when the data are unbalanced, one encounters the issue of incoherence of the autoregressive coefficients and prediction variances. Much of the existing literature leveraging this framework fails to point this out or explicitly address the problem. Through its Cholesky decomposition, we propose an approach to covariance estimation that naturally permits missing observations in the longitudinal

dataset. Viewing vectors of repeated measurements as the observation of a continuous process at a sequence of time points, we accommodate unbalanced data by extending the regression model associated with the reparameterization to a bivariate functional varying coefficient model.

The remainder of this dissertation is structured as follows. Chapter 2 serves as a brief survey of developments in covariance estimation, concluding with presentation of the Cholesky decomposition and the estimation of the associated regression model. We extend this regression model to a functional varying coefficient model for (potentially) unbalanced data; in Chapters 3 and 4 we demonstrate covariance estimation with two approaches to bivariate smoothing. Chapter 5 examines various aspects of the performance of our proposed estimator through simulation studies, and in Chapter 6 we apply our procedure to a real dataset. We conclude with a summary of our work and remark on remaining open questions and future endeavors in Chapter 7.

Chapter 2: Covariance Estimation: A Review

Estimation of a covariance matrix Σ is fundamental to the analysis of multivariate data. The two primary challenges in fulfilling this prerequisite are due to the total number of parameters to be estimated in relation to the dimension, and the structural constraints that the elements of a covariance matrix should satisfy. The number of parameters grows quadratically in the dimension, and these parameters must satisfy the positive-definiteness constraint. That is, the elements of a $p \times p$ covariance matrix $\Sigma = [\sigma_{ij}]$ for p variables, satisfy the constraint that

$$c'\Sigma c = \sum_{i,j=1}^{p} c_i c_j \sigma_{ij} \geq 0 \tag{2.1}$$

for all $c = (c_1, \ldots, c_p)' \in \Re^p$. These challenges have motivated a growing body of research aimed at effectively estimating covariance matrices. Given a sample of random vectors $Y_1, \ldots, Y_N$ from a distribution with covariance matrix Σ, a common starting point in the pursuit of an estimate of this matrix is the sample covariance matrix S:

$$S = (N-1)^{-1} \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)\left(Y_i - \bar{Y}\right)' \tag{2.2}$$

where $\bar{Y} = N^{-1} \sum_{i=1}^{N} Y_i$ denotes the sample mean vector. The sample covariance matrix is both a straightforward and flexible estimator of the $\frac{p(p+1)}{2}$ parameters of the unstructured covariance matrix Σ, and it is unbiased for Σ. Its construction produces a positive definite estimate so that the constraint in (2.1) is satisfied.

Despite these merits, it has been well established that the empirical covariance matrix is unstable in high dimensions; see Lin (1985) or Johnstone (2001), for example. The sample covariance is not parsimonious, making it unsatisfactory when it is suspected that the true underlying covariance matrix is sparse, or has many of its elements equal to zero. Moreover, it is not uncommon to encounter practical situations in which the data do not permit the straightforward construction in (2.2). Specifically, we are interested in estimating the covariance matrix associated with a vector of repeated measurements generated from longitudinal studies in which the measurements on the i^{th} subject $Y_i = (y_{i1}, y_{i2}, \ldots, y_{ip})'$ are associated with measurement times $t_i = (t_{i1}, t_{i2}, \ldots, t_{ip})'$. In this setting, the sample covariance matrix is not necessarily an optimal estimator of the covariance matrix because it does not naturally incorporate the temporal structure of the data. Moreover, construction of the sample covariance matrix requires rectangular, or balanced, data. Table 2.1 shows the ideal form of a (rectangular) longitudinal data set. Unfortunately, longitudinal studies frequently produce non-rectangular data, where trajectories are potentially sparsely observed at times which are not common across all subjects. In the case, construction of the sample covariance matrix as defined in (2.2) is infeasible.

Table 2.1: *Ideal form of repeated measurements.*

		Time					
		1	2	...	t	...	p
Unit	1	y_{11}	y_{12}	...	y_{1t}	...	y_{1p}
	2	y_{21}	y_{22}	...	y_{2t}	...	y_{2p}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	i	y_{i1}	y_{i2}	...	y_{it}	...	y_{ip}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	N	y_{N1}	y_{N2}	...	y_{Nt}	...	y_{Np}

These drawbacks have accelerated numerous initiatives detouring the pitfalls on the most obvious route to a covariance estimate toward deliberate modeling of structured covariance matrices for longitudinal data. These methods employ a number of approaches to reducing the dimension of the parameter space to balance flexibility and stability of estimators. In this chapter, we present a review of existing methods for covariance estimation, focusing on those developed specifically for the application to longitudinal data. Our review is by no means exhaustive and focuses on developments made in covariance estimation from two connected perspectives: regularized covariance matrices, and parsimonious models, including the use of covariates in low dimensions through generalized linear models (GLM) for covariance. We examine three general classes of estimators: structured covariance models, the sample covariance matrix and its regularized variants, and models for reparameterizations of the covariance matrix. To promote clarity in the discussion of covariance estimation, for the remainder of this dissertation, we assume that a random vector $Y = (y_1, \ldots, y_p)'$ is centered to have mean zero, unless explicitly indicated.

2.1 Structured Parametric Covariances

In the applied statistics literature, particularly for repeated measures data, it is quite common to pick a stationary covariance matrix for the covariance structure. These parametric structures were an attractive alternative to the common approaches that had, at the time, historically been used to estimate the covariance of a multivariate random vector. These approaches included the univariate ANOVA and repeated measures ANOVA. In the univariate ANOVA setting, a separate conventional analysis of variance is performed on the data from each distinct measurement time. Repeated measures ANOVA entails performing ANOVA as if the data were from a split-plot experiment with time of measurement being factor defining the split-plots. The ensuing parametric covariance models sought to exploit the temporal structure in longitudinal data, while the ANOVA-based

approaches fail to explicitly make sure of this information. The parsimony of these parametric structures make their computational requirements modest, and software packages implementing fitting procedures for a growing number of simple models are readily accessible. In this section, we discuss some of of the parametric models most commonly encountered in the covariance estimation literature. For comprehensive discussion of parametric models for repeated measures data, see Jennrich and Schluchter (1986), for example.

The Compound Symmetric Model

At one time, the compound symmetric model was a very popular choice for parametric covariance structure. It specifies constant variance and constant correlation between all pairs of variables, where the elements of the variance-correlation matrix are given by

$$\sigma_{ii} = \sigma^2 \text{ for } i = 1, \ldots, p, \quad \rho_{ij} = \rho \text{ for } i \neq j, \tag{2.3}$$

where $\sigma_{ii} = Var\left(y_i\right)$ denotes the i^{th} diagonal element of Σ, and $\rho_{ij} = Cov\left(y_i, y_j\right) / \sqrt{\sigma_{ii}\sigma_{jj}}$. The parsimony of this model is a primary reason for its attractiveness, having only two parameters to be estimated. However, with the development of models allowing for heterogeneous variances and non-constant correlation, it has received less attention as of late, particularly in the longitudinal statistics literature.

Autoregressive Models

Low order autoregressive models are among the most frequently used models for time series and repeated measures data. The first order autoregressive model for response variable y_t associated with measurement time t specifies

$$y_t = \begin{cases} \epsilon_t, & t = 1, \\ \phi y_{t-1} + \epsilon_t, & t = 2, \ldots, p, \end{cases} \tag{2.4}$$

where $|\phi| < 1$, and the innovations $\{\epsilon_t\}$ are independently distributed according to $N(0, \sigma_t^2)$ with $\sigma_1^2 = \sigma^2/(1-\rho^2)$, and $\sigma_t^2 = \sigma^2$ for $t = 2, \ldots, p$. The correlations are given by $\rho_{ij} = \phi^{|i-j|}$ for $i \neq j$, which are monotonically decreasing in $l = |i-j|$.

The AR(1) model generalizes to any arbitrary order k by simply adding additional predecessors to the covariates in the linear model for y_t:

$$y_t = \begin{cases} \epsilon_t, & t = 1, \\ \sum_{j=1}^{k} \phi_j y_{t-j} + \epsilon_t, & t = 2, \ldots, p, \end{cases}$$

where $k = \min(p, t-1)$, and the $\{\epsilon_t\}$ are independent mean zero Normal random variables. The variance of $\{\epsilon_t\}$ is constant for $t > k$, and for $t \leq k$, the variance is specified so as to ensure that the variance is constant across all responses y_t and the covariance between y_i and y_j depends only on $|i-j|$.

Moving Average Models

Equally as common as the autoregressive model is the moving average model. The response specification for q^{th} order moving average model is given by

$$y_t = \sum_{j=0}^{q} \theta_j \epsilon_{t-j}, \tag{2.5}$$

where the $\{\epsilon_t\}$ are independently and identically distributed mean zero Normal random variables with variance σ^2. This model corresponds to correlation matrix with non-diagonal elements

$$\rho_{ij} = \begin{cases} \left(\theta_{i-j} + \theta_1\theta_{i-j+1} + \cdots + \theta_{q-i+j}\theta_q\right) / \left(1 + \sum_{j=1}^{q} \theta_j^2\right), & 0 < |i-j| \leq q, \\ 0, & |i-j| > q \end{cases}$$

The variances defining the diagonal elements of the covariance matrix are given by

$$\sigma_{ii} = \sigma^2 \sum_{j=0}^{q} \theta_j^2, \quad i = 1, \ldots, p.$$

The variances are constant, and the correlations between y_t and y_{t-l} vanish beyond a finite, constant lag l. Here $\theta_1, \ldots, \theta_q$ are arbitrary parameters subject only to positive definiteness constraints.

This model generalizes to a q^{th}-order Toeplitz model, with elements of the covariance matrix given by

$$\sigma_{ij} = \begin{cases} \sigma^2 \rho_{i-j} & |i-j| \leq q, \\ 0 & |i-j| > q, \\ \sigma^2 & i = j, \end{cases} \tag{2.6}$$

or covariance matrix of the form

$$\begin{bmatrix} m_0 & m_1 & m_2 & \ldots & m_{p-1} \\ m_1 & m_0 & m_1 & \ldots & m_{p-2} \\ m_2 & m_1 & m_0 & \ldots & m_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{p-1} & m_{p-2} & m_{p-3} & \ldots & m_0 \end{bmatrix}, \tag{2.7}$$

where $m_j = 0$ for all $j > q$.

The aforementioned models are stationary, specifying constant variance and with equal same-lag correlations among responses when the data are observed on a regular grid. Heterogeneous extensions of these models specify the same form of the correlation but allow time-dependent response variances. Completely general time dependence (subject to positive definiteness constraints) requires the covariance structure to be characterized by $O(p)$ parameters, while specifying linear or quadratic dependence on time leads to more parsimonious heterogeneous models.

ARIMA Models

An ARIMA(p, d, q) model generalizes a stationary autoregressive moving average (ARMA) model by postulating that not the observations themselves, but rather the d^{th}-order differences among consecutive measurements follow a stationary ARMA(p, q) model. A special case is the

ARIMA$(0, 1, 0)$ model - the random walk:

$$y_j = \sum_{k=1}^{j} \epsilon_k, \quad j = 1, \ldots, p, \tag{2.8}$$

where the ϵ_k are independent mean zero Normal random variables with variance σ^2. The variance of the process $Var\left(y_j\right) = j\sigma^2$ increases linearly in time. The correlation between y_j and y_k also increases, but nonlinearly, in time:

$$Corr\left(y_j, y_k\right) = \sqrt{\frac{j}{k}}, \quad j > k.$$

This model is applicable to longitudinal data which are observed on a regular grid, however, its continuous time analogue permits this restriction to be relaxed. An important special case is the continuous time analogue to the random walk, the Weiner process, which has covariance function $Cov\left(y\left(t_i\right), y\left(t_j\right)\right) = \sigma^2 \min\left(t_i, t_j\right)$.

Random Coefficient Models

Random coefficient models are a broad class of models often used for clustered or longitudinal data. They offer reasonable flexibility for characterizing dependency structure but remain parsimonious because the number of model parameters is unrelated to the number of repeated measurements and can be applied to non-rectangular data. The formulation of the covariance structure for these models is most usually a consideration of regressions that vary across subjects rather than a consideration of within-subject similarity, which is why they are most often considered distinct from parametric covariance models. Still, they yield parametric covariance structures that generally have non-constant variances and non-stationary correlations. A general form of the random coefficient model for $p_i \times 1$ vector of measurements on subject i is given by

$$Y_i = X_i\beta + Z_i\gamma_i + \epsilon_i, \quad i = 1, \ldots, N, \tag{2.9}$$

where $Y_i = \left(y_{i1}, \ldots, y_{ip_i}\right)'$ are measurements taken at equally-spaced times $t_{i1}, \ldots, t_{ip_i}$, the Z_i are specified matrices, the γ_i are vectors of random coefficients distributed independently as $N\left(0, G_i\right)$, the G_i are positive definite but otherwise unstructured matrices, and the ϵ_i are distributed independently (of the γ_i and of each other) as $N\left(0, \sigma^2 \mathrm{I}_{p_i}\right)$. The G_i are usually assumed to be equal across subjects, so the covariance matrix of Y_i is taken to be $\Sigma_i = Z_i G Z_i' + \sigma^2 \mathrm{I}_{p_i}$. Special cases include the linear random coefficients (RCL) and quadratic random coefficients (RCQ) models. In the linear case, $Z_i = \left[1_{p_i}, \left(t_{i1}, \ldots, t_{i,p_i}\right)'\right]$ and

$$G = \begin{bmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{10} & \sigma_{11} \end{bmatrix}$$

In the quadratic case, $Z_i = \left[1_{p_i}, \left(t_{i1}, \ldots, t_{i,p_i}\right)', \left(t_{i1}^2, \ldots, t_{i,p_i}^2\right)'\right]$. It is worth noting that when $Z_i = 1_{p_i}$, the random coefficient model corresponds to the compound symmetric model (2.3). The covariance structure for a subject having measurements taken at measurement times $t_1 = 1, \ldots, t_{p_i} = p_i$ is given by

$$\sigma_{ij} = \begin{cases} \sigma^2 + \sigma_{00} + 2\sigma_{01} j + \sigma_{11} j^2 & i = j, \\ \sigma_{00} + \sigma_{01}\left(i + j\right) + \sigma_{11} ij & i \neq j \end{cases} \tag{2.10}$$

These models permit variance and covariances exhibiting several kinds of time dependency, including increasing or decreasing variances and correlations of which some are negative while others are positive. However, this model does not permit variances which are concave-down in time, and it precludes the variances from being constant if the same-lag correlations are different.

2.2 Shrinking the Sample Covariance Matrix

The simple structure of parametric models is typically accompanied by straightforward interpretation of model coefficients and minimal computational issues. While the choices for parametric model structure are seemingly unlimited, specifying the appropriate parametric covariance structure is challenging even for the experts, and model misspecification can lead to considerably

biased estimates. From this standpoint, it is prudent to allow the data to drive the formulation of the dependency structure. Estimates of parametric models are one extreme, typically exhibiting low variance but potentially high bias. The sample covariance matrix could characterize the other extreme. An unbiased estimator for the $p\left(p+1\right)/2$ parameters of an unstructured covariance matrix, it trades stability for flexibility. Between these poles lies a broad class of estimators which seeks to balance these two objectives.

Approaches rooted in decision theory yield stable estimators which are scalar multiples of the sample covariance matrix; these estimators distort the eigenstructure of Σ unless the sample size is much greater than the dimension, $N \gg p$ (Dempster, 1972). There is a vast body of work which addresses the efficient estimation of the covariance matrix of a normal distribution by correcting the eigenstructure distortion or reducing the number of parameters to be estimated. See Stein (1975), Lin (1985), Yang and Berger (1994), Daniels and Kass (1999), and Champion (2003).

Stein's Estimator

Stein (1975) observed that the sample covariance matrix systematically distorts the eigenstructure of Σ, especially when p is large. His work spurred efforts in the improvement of S, which he did by simply shrinking its eigenvalues. Given the spectral decomposition of the sample covariance matrix

$$S = \hat{P}\hat{\Lambda}\hat{P}' = \sum_{i=1}^{p} \hat{\lambda}_i \hat{e}_i \hat{e}_i',$$

he considered estimators of the form

$$\hat{\Sigma} = \hat{P}\Phi\left(\hat{\lambda}\right)\hat{P}', \tag{2.11}$$

where $\hat{\lambda} = \left(\hat{\lambda}_1, \ldots, \hat{\lambda}_p\right)'$, $\hat{\lambda}_1 > \cdots > \hat{\lambda}_p$ are the ordered eigenvalues of S, $\hat{P}$ is the orthogonal matrix whose i^{th} column is the normalized eigenvector of S corresponding to $\hat{\lambda}_i$, and $\Phi\left(\hat{\lambda}\right) =$

$diag\left(\phi_1, \ldots, \phi_p\right)$ is the diagonal matrix where $\phi_j\left(\hat{\lambda}\right)$ is an estimate of the j^{th} largest eigenvalue of Σ. Letting $\phi_j\left(\hat{\lambda}\right) = \hat{\lambda}_j$ corresponds to the usual unbiased estimator S. It is known that $\hat{\lambda}_1$ and $\hat{\lambda}_p$ are biased low and high, respectively, so Stein specified $\Phi\left(\hat{\lambda}\right)$ to shrink the eigenvalues toward central values to counteract the biases of the sample eigenvalues. The modified estimators of the eigenvalues of Σ are given by $\phi_j = \frac{N\hat{\lambda_j}}{\alpha_j}$, where

$$\alpha_j\left(\hat{\lambda}\right) = N - p + 2\hat{\lambda}_j \sum_{i \neq j} \frac{1}{\hat{\lambda}_j - \hat{\lambda}_i}. \tag{2.12}$$

The Stein estimators ϕ_j differ from the sample eigenvalues when they are nearly equal and N/p is not small. The work of Lin (1985) includes an algorithm to modify any ϕ_j's which are negative and or do not satisfy $\phi_1 > \cdots > \phi_p$.

Ledoit and Wolf's Estimator

The estimator proposed by Ledoit and Wolf (2004) is motivated by the fact that the sample covariance matrix is unbiased but has high variance - the risk associated with S is considerable when $p \gg N$, and even in cases when the dimension is close to the sample size. In contrast, very little estimation error is associated with a highly structured estimator of a covariance matrix, like those presented in Section 2.1, but when the model is misspecified, these can exhibit severe bias. A natural inclination is to define an estimator as a linear combination of the two extremes, letting

$$\hat{\Sigma} = \alpha_1 I + \alpha_2 S, \tag{2.13}$$

where α_1 and α_2 are chosen to minimize

$$\frac{1}{p}||\hat{\Sigma} - \Sigma||_F^2 = \frac{1}{p}\text{tr}\left[\left(\hat{\Sigma} - \Sigma\right)^2\right].$$

They show that the optimal α_i depend on only four characteristics of the true covariance matrix:

$$\begin{aligned} \mu &= \text{tr}\left(\Sigma\right)/p, \\ \alpha^2 &= ||\Sigma - \mu I||^2, \\ \beta^2 &= ||S - \Sigma||^2, \\ \delta^2 &= ||S - \mu I||^2. \end{aligned} \tag{2.14}$$

Ledoit and Wolf (2004) give consistent estimators of these quantities, so that substitution of these in $\hat{\Sigma}$ produces a positive definite estimator of Σ. They demonstrate the superiority of their estimator to several others including the sample covariance matrix and the empirical Bayes estimator (Haff, 1980).

A broad class of estimators aim to stabilize the sample covariance matrix by applying shrinkage elementwise to each of its entries. Many have explored the use of thresholding, banding, and tapering to stabilize the covariance matrix, resulting in estimators that are computationally inexpensive due to their convenient construction. This convenience, however, comes with a tradeoff: because the estimators are constructed by elementwise transformations of the sample covariance, they are not guaranteed to be positive definite. Nonetheless, certain types of elementwise shrinkage estimators enjoy attractive asymptotic properties (Bickel and Levina, 2008) which, in addition to their straightforwardness, perhaps offset their finite sample shortcomings.

Banding the Sample Covariance Matrix

Setting certain entries of the sample covariance matrix to zero is one approach to stabilize the estimator by reducing the dimension of the parameter space. Time series analysis is an example of the classic situation in which $p \gg N$. One typically observes a sample size of $N = 1$, with the data being a single, long realization of the random vector, which severely necessitates a reduction in the dimension of the parameter space. One way to do this is to assume stationarity of the process,

which reduces the number of distinct parameters of the $p \times p$ covariance matrix Σ from $p\left(p+1\right)/2$ to p, which could still be large. Moving average and autoregressive models reduce the number of parameters in the same way as banding a covariance or inverse covariance matrix (Bickel and Levina, 2008; Wu and Pourahmadi, 2009). For a given sample covariance matrix $S = [s_{ij}]$ and integer k, $0 < k < p$, the k-banded sample covariance matrix is given by

$$B_k\left(S\right) = \left[s_{ij} 1\left(|i-j| \leq k\right)\right]. \tag{2.15}$$

This kind of regularization is ideal when the indices have been arranged so that

$$|i-j| > k \Rightarrow \sigma_{ij} = 0.$$

Such structure often implies that variables far apart with respect to time ordering are only weakly correlated, such as when, for example, y_t, $t = 1, \ldots, p$ follow a finite heterogeneous moving average process

$$y_t = \sum_{j=1}^{k} \theta_{t,t-j} \epsilon_j,$$

where the ϵ_j's are iid mean zero errors having finite variance. Banding estimators are a special case of tapering estimators, which have the form

$$\hat{\Sigma} = R * S, \tag{2.16}$$

where R is a positive definite tapering matrix, and the $*$ operator denotes the Schur matrix multiplication (the element-wise matrix product). The Schur product of two positive definite matrices is also guaranteed to be positive definite, so the tapering estimator's positive definiteness is dependent on the choice of tapering matrix R. Banding the sample covariance matrix is equivalent to premultiplying S by

$$R = [r_{ij}] = \left[1\left(|i-j| \leq k\right)\right],$$

which is not positive definite.

Asymptotic analysis of banding estimators is available when N, p, and k are large. Bickel and Levina (2008) establish consistency of the banded estimator in the operator norm, and uniform consistency over the class of "approximately bandable" matrices under a normal likelihood. Convergence requires that $\log p/N \to 0$, and they derive an explicit rate of convergence which depends on the rate at which k grows. Cai et al. (2010) proposed the following tapering estimator of the sample covariance matrix:

$$S^{\omega} = \left[\omega_{ij}^{k} s_{ij}\right], \tag{2.17}$$

where the ω_{ij}^{k} are given by

$$\omega_{ij}^{k} = k_h^{-1}\left[(k - |i-j|)_{+} - (k_h - |i-j|)_{+}\right].$$

The weights ω_{ij}^{k} are indexed with superscript to indicate that they are controlled by a tuning parameter, k, which can take integer values between 0 and p, the dimension of the covariance matrix. If $k_h = k/2$ is even, then the weights may be rewritten as

$$\omega_{ij} = \begin{cases} 1, & |i-j| \leq k_h \\ 2 - \frac{i-j}{k_h}, & k_h < |i-j| \leq k, \\ 0, & \text{otherwise.} \end{cases}$$

This expression indicates how the selection of k controls the amount of shrinkage applied to a particular element of the sample covariance matrix. Elements of S belonging to the subdiagonals closest to the main diagonal are left unregularized. The shrinkage applied to elements increases as we move away from the diagonal: a multiplicative shrinkage factor of $2 - \frac{i-j}{k_h}$ is applied to elements belonging to subdiagonals $k_h, \ldots, k-1, k$, and elements further than k subdiagonals from the main diagonal are shrunk to zero. Cai et al. (2010) derived optimal rates of convergence under the operator norm for their estimator and presented simulations demonstrating that it nearly uniformly outperforms the banded estimator of Bickel and Levina (2008).

Thresholding the Sample Covariance Matrix

When both N and p are large, it is reasonable to assume that Σ is sparse, so that many elements of the covariance matrix are equal to 0. In this case, setting certain elements of the sample estimate to zero can improve the quality of the estimator. Thresholding was originally a method developed in nonparametric function estimation, but recently Bickel et al. (2008) and Rothman et al. (2009) have utilized thresholding for estimating large covariance matrices. Shrinkage and thresholding estimators can be viewed as the solution to the problem of minimizing a penalized quadratic loss function, and since the thresholding operator is applied elementwise to the sample covariance S, these optimization problems are univariate. Rothman et al. (2009) presented a class of generalized thresholding estimators constructed by applying a thresholding operator to each element of the sample covariance matrix. This class includes the soft-thresholding estimator given by

$$S^{\lambda} = \left[\text{sign}\left(s_{ij}\right)\left(s_{ij} - \lambda\right)_{+}\right],$$

where s_{ij} denotes the i-j^{th} entry of the sample covariance matrix, and λ is a penalty parameter controlling the amount of shrinkage applied to S.

These estimators are simple to compute compared to competitor estimates, like the L_1-penalized likelihood estimator, but they suffer from the lack of guaranteed positive definiteness. However, similar to the result for banded estimators, Bickel et al. (2008) have established the consistency of the threshold estimator in the operator norm, uniformly over the class of matrices that satisfy a certain sparsity requirement. Soft thresholding can result in zeros irregularly placed in the resulting estimator, which may not be an optimal choice for sparsity pattern when there is a natural ordering of the variables as with longitudinal data.

Alternately, for estimating the covariance of a random vector which is assumed to have a natural (time) ordering, several have proposed applying kernel smoothing methods directly to elements of

the sample covariance matrix or a function of the sample covariance matrix. Zeger and Diggle (1994) introduced a nonparametric estimator obtained by kernel smoothing the sample variogram and squared residuals. Yao et al. (2005) applied a local linear smoother to the sample covariance matrix in the direction of the diagonal and a local quadratic smoother in the direction orthogonal to the diagonal to account for the presence of additional variation due to measurement error. The latter work is one of the few nonparametric methods utilizing smoothing in both dimensions of the covariance matrix, which was an inspiration of sorts for the work we present in Chapter 3. Like other elementwise shrinkage estimators, however, their proposed estimator is not guaranteed to be positive definite.

2.3 Matrix Decompositions

The positive definite constraint poses a challenge in most covariance estimation settings. In this section, we demonstrate the role of matrix decompositions in removing it from the estimation procedure altogether. These approaches decompose the covariance matrix into its variance and dependence components, and are closely connected to the use of generalized linear models for covariance estimation. In this light, this overview serves as a prerequisite to Section 2.4 which will discuss covariance estimation from the generalized linear modeling perspective.

2.3.1 The Variance-Correlation Decomposition

The variance-correlation decomposition of Σ parameterizes the covariance matrix according to

$$\Sigma = DRD, \tag{2.18}$$

where $D = \text{diag}\left(\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{pp}}\right)$ denotes the diagonal matrix with diagonal entries equal to the square-roots of those of Σ, and R is the corresponding correlation matrix. This parameterization

enjoys attractive practicality because the standard deviations are on the same scale as the responses, and because the estimation of D and R can be separated by iteratively fixing one sequence of parameters to estimate the other. In some applications, one set of parameters may be more important than the others; the dynamic correlation model presented in Engle (2002) is actually motivated by the fact that variances (volatilities) of individual assets are more important than their time-varying correlations.

While the natural log of the diagonal entries of D are unconstrained, the correlation matrix R is constrained to have unit diagonal entires and off-diagonal entries to be less than or equal to 1 in absolute value. Consequently, the variance-correlation decomposition does not lend to modeling its components with the use of covariates. In the literature of longitudinal data analysis and other areas of application which frequently handle correlated data, preferred models for the variance-correlation decomposition typically involve structured correlation matrices with a few parameters, in the interest of parsimony and ensuring positive definiteness (Zimmerman and Núñez-Antón, 1997).

2.3.2 The Spectral Decomposition

The spectral decomposition is the basis of several methods in multivariate statistics, including principal component analysis and factor analysis (Anderson, 1984; Hotelling, 1933). The spectral decomposition of a covariance matrix Σ is given by

$$\Sigma = P\Lambda P' = \sum_{i=1}^{p} \lambda_i e_i e_i', \tag{2.19}$$

where Λ is a diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_p$, and P is the orthogonal matrix of normalized eigenvectors, having e_i as its i^{th} column. The entries of Λ and P can be interpreted as the variances and coefficients of the p principal components. The matrix P is constrained by its orthogonality, so modeling it within the framework to reduce parameter dimension is inconvenient. In

spite of this, Chiu et al. (1996) proposed an new unconstrained reparameterization of a covariance matrix using the spectral decomposition, modeling the matrix logarithm:

$$\log \Sigma = P \left(\log \Lambda \right) P' = \sum_{i=1}^{p} \log \left(\lambda_i \right) e_i e_i', \tag{2.20}$$

The components $\log \lambda_i$ are free but lack any relevant statistical interpretability. Interestingly, this highlights the tradeoff between the requirements for unconstrained parameterization of covariance matrices and the statistical interpretability of the new parameters. We further discuss the log-linear GLM for covariance matrices in Section 2.4.2.

2.3.3 The Cholesky Decomposition

The Cholesky decomposition has received a lot of attention in recent developments in covariance estimation. Unlike the spectral decomposition, it offers an unconstrained parameterization without sacrificing the interpretability of the components of the decomposition. The Cholesky decomposition of a positive-definite matrix is given by

$$\Sigma = CC', \tag{2.21}$$

where $C = [c_{ij}]$ is a unique lower-triangular matrix with positive diagonal entries. This factorization is frequently encountered in optimization techniques and matrix computation (Golub and Van Loan, 2012). It is difficult to attach any statistical interpretation to the entries of C in this form (Pinheiro and Bates, 1996). However, statistical interpretation of the diagonal entries of C and the resulting unit lower-triangular matrix is available by transforming C to a unit lower-triangular matrix, dividing the i^{th} column of C by its i^{th} diagonal element c_{ii}. Letting $D^{1/2} = diag\left(c_{11}, \ldots, c_{pp}\right)$, the standard Cholesky decomposition (2.21) can be written

$$\Sigma = CD^{-1/2}DD^{-1/2}C' = LDL', \tag{2.22}$$

where $L = D^{-1/2}C$. This is commonly referred to as the modified Cholesky decomposition (MCD) of Σ. It is common to write (2.22) in terms of the lower triangular matrix that diagonalizes Σ:

$$D = T\Sigma T', \tag{2.23}$$

where $T = L^{-1}$. Like the orthogonal matrix P in the spectral decomposition, the lower triangular matrix T diagonalizes Σ, however the entries of T can be written as the coefficients of a particular regression model, and are therefore unconstrained. The elements of the diagonal matrix D can also be interpreted as parameters 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 $\epsilon_t = y_t - \hat{y}_t$ denote the corresponding prediction error with variance $Var(\epsilon_t) = \sigma_t^2$. Standard regression machinery gives us that there exist unique scalars ϕ_{tj} so that

$$y_t = \begin{cases} \epsilon_t, & t = 1 \\ \sum_{j=1}^{t-1} \phi_{tj} y_j + \epsilon_t, & t = 2, \ldots, p, \end{cases} \tag{2.24}$$

and the mean zero prediction errors are independently distributed. Denote the variance of the prediction errors by $Var(\epsilon_t) = \sigma_t^2$. The connection between the Cholesky decomposition and the autoregressive model (2.24) is established by noting that the Cholesky factor contains the negatives of the regression coefficients and the prediction error variances are the diagonal elements of D. Let $\epsilon = (\epsilon_1, \ldots, \epsilon_p)'$ denote the vector of uncorrelated prediction residuals with

$$Cov(\epsilon) = D = diag\left(\sigma_1^2, \ldots, \sigma_p^2\right).$$

Then model (2.24) can be written

$$\epsilon = TY, \tag{2.25}$$

where the (t, j) entry of T is $-\phi_{tj}$, and the (t, t) entry of D is the variance of the t^{th} prediction residual: $\sigma_t^2 = Var(\epsilon_t)$.

$$\begin{bmatrix} 1 & & & & \\ -\phi_{21} & 1 & & & \\ -\phi_{31} & -\phi_{32} & 1 & & \\ \vdots & & & \ddots & \\ -\phi_{p1} & -\phi_{p2} & \dots & -\phi_{p,p-1} & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{bmatrix} \tag{2.26}$$

Table 2.2 illustrates how the components of a covariance matrix are obtained through successive regressions. Specifically, this representation demonstrates how 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, for any unstructured covariance matrix, these and the log variances are unconstrained. The regression coefficients of the model in (2.24) are referred to as the *generalized autoregressive parameters* (GARP) and *innovation variances* (IV) (Pourahmadi, 1999, 2000). The powerful implication of the parallel regression framework of decomposition (2.23) is the accessibility of the entire portfolio of regression methods for the service of modeling covariance matrices. Moreover, the estimator $\hat{\Sigma}^{-1} = \hat{T}'\hat{D}^{-1}T$ constructed from the unconstrained parameters ϕ_{tj}, σ_j^2 is guaranteed to be positive definite.

Table 2.2: *Autoregressive coefficients and prediction error variances of successive regressions.*

y_1	y_2	y_3	$\dots$	y_{p-1}	y_p
1					
ϕ_{21}	1				
ϕ_{31}	ϕ_{32}	1			
$\vdots$	$\vdots$		$\ddots$		
$\vdots$	$\vdots$			$\ddots$	
ϕ_{p1}	ϕ_{p2}	$\dots$	$\dots$	$\phi_{p,p-1}$	1
σ_1^2	σ_2^2	$\dots$	$\dots$	σ_{p-1}^2	σ_p^2

2.4 Generalized Linear Models for Covariances

The positive-definiteness constraint and parameter space dimensionality are the major hurdles plaguing covariance estimation. However, within the context of regression analysis for modeling the mean vector μ of a random vector $Y = (y_1, \ldots, y_p)'$, similar challenges have been handled successfully through the use of generalized linear models (GLM). The GLM framework McCullagh and Nelder (1989) merges numerous seemingly disconnected approaches for modeling the mean of a distribution. Much of the success of the GLM is due to the use of a link function $g(\cdot)$ and a linear predictor $g(\cdot) = X\beta$, where X is a design matrix containing covariates which characterize the behaviour of the response. The link function and linear predictor together induce an unconstrained parameterization and reduce the parameter space dimension simultaneously. The covariance matrix, which is defined $\Sigma = E(Y - \mu)(Y - \mu)'$, can be viewed a mean-like parameter, so it is a natural inclination to exploit the idea of the GLM for covariance estimation. In the GLM setting, simply applying a link function componentwise to the constrained mean vector μ permits its unconstrained estimation. Unfortunately, employing the same approach to covariance matrices isn't viable since positive-definiteness is a simultaneous constraint on all entries of a matrix.

In addition to providing an avenue for sidestepping the positive definite constraint, the use of the GLM allows for the explicit use of covariates for estimating a covariance matrix, which is particularly attractive for longitudinal data or spatial data, where the variables exhibit a natural ordering. Extensions of the GLM to large classes of models include nonparametric and generalized additive models, Bayesian GLM, and generalized linear mixed models; see Hastie and Tibshirani (1990), Dey et al. (2000), and McCulloch and Neuhaus (2001). An analogous framework for modeling covariance matrices facilitates further developments in covariance estimation from the Bayesian, nonparametric and other paradigms. Successfully employing a link function for

unconstrained estimation of a general covariance matrix necessitates decomposing a covariance matrix into its "variance" and "dependence" components. In the previous section, we discussed the variance-correlation decomposition, the spectral decomposition, and the Cholesky decomposition, which factor Σ in such a way, and described the advantages that the Cholesky decomposition enjoys over the other two.

2.4.1 Linear Models for Covariance

Gabriel (1962) was among the first to implicitly parameterize a multivariate normal distribution in terms of entries of the precision matrix Σ^{-1}. Dempster (1972) recognized the entries of $\Sigma^{-1} = [\sigma^{ij}]$ as the canonical parameters of the exponential family of normal distributions with mean zero and unknown covariance matrix Σ:

$$\log f\left(Y|\Sigma^{-1}\right) = -\frac{1}{2}\text{tr}\Sigma^{-1}\left(Y'Y\right) + \log|\Sigma|^{-1/2} - p\log\sqrt{\pi}.$$

Soon thereafter, the simple structures of time series and variance components models motivated Anderson (1973) to define the class of linear covariance models:

$$\Sigma = \sum_{i=1}^{q} \alpha_q U_q, \tag{2.27}$$

where the U_is are known symmetric matrices and the α_is are unknown parameters, restricted to ensure that Σ is positive definite. This class of models is general enough to include all linear mixed effects models as well as certain time series and graphical models. For q large enough, any covariance matrix admits representation of the form (2.27), since one can decompose every covariance matrix as

$$\Sigma = \sum_{i=1}^{p}\sum_{j=1}^{p} \sigma_{ij} U_{ij}, \tag{2.28}$$

where U_{ij} is a $p \times p$ matrix with a 1 in the (i, j) position, and zeros everywhere else. The linear model (2.27) can be viewed as modeling the link-transformed covariance $g\left(\Sigma\right) = \sum_{i=1}^{q} \alpha_q U_q$,

where $g(\cdot)$ is the identity link. Despite the convenient parameterization, the positive definite constraint (2.1) makes estimation an arduous task.

Inducing sparsity by setting certain elements of the covariance matrix or its inverse to zero is a common approach to reducing the dimensionality of a covariance structure. Inspection of model (2.27) and the covariance parameterization given in (2.28) makes it easy to see that this can be achieved by eliminating certain U_{ij} from the covariates in the linear covariance model. On the extreme end of the sparsity spectrum is the case of independent variables and Σ is diagonal, eliminating all U_{ij} from the linear model covariates for $i \neq j$. Connection between the linear covariance model and other models for covariance discussed in previous sections can be established if we consider intermediary cases, such as classes of stationary moving average (MA) and autoregressive (AR) models introduced in the early times series literature. The $MA(q)$ model corresponds to a banded covariance matrix, setting

$$\sigma_{ij} = 0 \quad \text{for } |i-j| > q, \tag{2.29}$$

while the $AR(k)$ model corresponds to a banded inverse:

$$\sigma^{ij} = 0 \quad \text{for } |i-j| > k. \tag{2.30}$$

Of course, there are the nonstationary analogues to these classes of models, some of which were discussed in Section 2.1. We will review others which are related to antedependence models and Gaussian graphical models. Random variables $y_1, \ldots, y_p$, which correspond to observation times $t_1, \ldots, t_p$, with multivariate normal joint distribution are said to be k^{th}-order antedependent or $AD(k)$ (Gabriel, 1962) if y_t and y_{t+s+1} are independent given the intervening values $y_{t+1}, \ldots, y_{t+s}$ for $t = 1, \ldots, k-s-1$ and all $s \geq k$. A random vector $Y = (y_1, \ldots, y_p)$ is $AD(k)$ if and only if its covariance matrix satisfies (2.30). Closely connected are the classes of variable order AD

models and varying order, varying coefficient autoregressive models (Kitagawa and Gersch, 1985) in which the coefficients and order of antedependence depend on time.

2.4.2 Log-Linear Covariance Models

The constraint on the α_is in (2.27) was eliminated with the introduction of log-linear covariance models (Chiu et al., 1996; Pinheiro and Bates, 1996). For a general covariance matrix having spectral decomposition $\Sigma = P\Lambda P'$ its matrix logarithm, $\log \Sigma$, defined

$$\log \Sigma = P \left(\log \Lambda\right) P'$$

is a symmetric matrix with unconstrained entries taking values in $\Re$. Application of the log-link function leads to the log-linear model for Σ:

$$g\left(\Sigma\right) = \log \Sigma = \sum_{i=1}^{q} \alpha_i U_i, \tag{2.31}$$

where the U_is are as before in (2.27) and the α_is are now unconstrained. The α_is, however, now lack statistical interpretation since $g\left(A\right) = \log A$ is a highly nonlinear operation. But for diagonal Σ, $\log \Sigma = \text{diag}\left(\sigma_{11}, \ldots, \sigma_{pp}\right)$, and model (2.31) reduces to modeling of heterogeneous variances, which has been extensively studied. Detailed presentation is given in Carroll and Ruppert (1988), Verbyla (1993) and in references therein.

Rice and Silverman (1991) were the first to pursue nonparametric estimation of the spectral decomposition for functional data, which arise from experiments which produce observed responses in the form of curves. See Ramsay (2006), Ramsay and Silverman (2007). The covariance structure is estimated via functional principal component analysis (fPCA); principal components of functional data are estimated using penalized least squares of the normalized eigenvectors, subject to the orthogonality constraint. Additionally, Boente and Fraiman (2000) proposeds kernel-based

PCA, but maintaining orthogonality of the smooth principal components remains a major computational challenge in both approaches.

2.4.3 The Cholesky Decomposition as a Generalized Linear Model

The log link resolves the issues presented by the constrained parameter space associated with the identity link, leading to unconstrained parameterization of a covariance matrix. However, the parameters of the matrix logarithm lack any meaningful statistical interpretation. The Cholesky decomposition leads to unconstrained and statistically meaningful reparameterization of the covariance matrix so that the ensuing GLM overcomes most of the shortcomings of the linear and log-linear models.

The nonredundant entries of $(T, \log D)$ are unconstrained, allowing them to be modeled using any desired technique, including parametric, semi- and nonparametric, and Bayesian approaches. For a random sample of mean zero p-dimensional vectors $Y_1, \ldots, Y_N$ from a normal density with covariance matrix Σ, the form of the likelihood allows for relatively simple computation of the MLE of the parameters. Up to a constant, the log likelihood satisfies

$$\begin{aligned} -2\ell\left(\Sigma | Y_1, \ldots, Y_N\right) &= \sum_{i=1}^{N} \left(\log|\Sigma| + Y_i'\Sigma^{-1}Y_i\right) \\ &= N\log|D| + N\text{tr}\left(\Sigma^{-1}S\right) \\ &= N\log|D| + N\text{tr}\left(D^{-1}TST'\right), \end{aligned} \tag{2.32}$$

where $S = N^{-1}\sum_{i=1}^{N} Y_i Y_i'$. The negative log likelihood (2.32) is quadratic in T for fixed D, so the MLE for the ϕ_{tj} has closed form. Similarly, the MLE for D for fixed T has closed form. See Pourahmadi (2000).

While the MLE is flexible under a saturated model, this advantage can be offset with high variance. Many have attempted to balance the tradeoff between bias and variance by reducing the dimension of the parameter space under model (2.24) in a number of ways. Because the Cholesky

decomposition can be viewed as a link function corresponding to a GLM for the covariance matrix, this can be done in a straightforward way with the use of covariates to elicit parametric models for ϕ_{jk} and $\log \sigma_j^2$. For example, the entries of T and $\log D$ can be modeled as follows:

$$\begin{aligned} \phi_{jk} &= x'_{jk}\beta, \\ \log \sigma_j^2 &= z'_j\gamma, \end{aligned} \tag{2.33}$$

where x_{tj} and z_t denote $q \times 1$ and $d \times 1$ vectors of known covariates, and $\beta = (\beta_1, \ldots, \beta_q)'$ and $\gamma = (\gamma_1, \ldots, \gamma_d)'$ are the parameters relating these covariates to the dependence among the elements of Y and the innovation variances. Covariates most frequently used in the analysis of real longitudinal data sets are low order polynomials of lag and time. Pourahmadi (1999), Pourahmadi (2000), and Pan and Mackenzie (2003) parameterize ϕ_{tj} and $\log \sigma_t^2$ using covariates

$$\begin{aligned} x'_{jk} &= \left(1, t_j - t_k, (t_j - t_k)^2, \ldots, (t_j - t_k)^{q-1}\right)' \\ z'_j &= \left(1, t_j, \ldots, t_j^{d-1}\right)'. \end{aligned} \tag{2.34}$$

They prescribe methods for identifying models of the form (2.33) using model selection criteria such as AIC and regressograms, which are a nonstationary analogue of the correlelogram one typically encounters in the time series literature. Pan and Mackenzie (2003) jointly estimate the mean and covariance of longitudinal data using maximum likelihood, iterating between estimation of the mean vector μ, the log innovation variances $\log \sigma_t^2$, and the generalized autoregressive parameters ϕ_{tj}. Score functions can be computed by direct differentiation of the normal log likelihood. Optimization is carried out by solving the score functions via iterative quasi-Newton method.

Modeling the covariance in such a way reduces a potentially high dimensional problem to something much more computationally feasible; if one models the innovation variances σ_t^2 similarly using a d-dimensional vector of covariates, the problem reduces to estimating $(q + d)$ unconstrained parameters, where much of the dimensionality reduction is a result of characterizing the GARPs in terms of only the difference between pairs of observed time points, and not the time

points themselves. This model specification of ϕ is equivalent to specifying a Toeplitz structure for Σ.

With the entries of T unconstrained, the Cholesky decomposition is ideal for nonparametric estimation and regularization methods. Many have alternatively proposed nonparametric and semiparametric techniques to reduce dimensionality without the risk of model misspecification often accompanying parametric models. Wu and Pourahmadi (2003) proposed local polynomial smoothers to individually estimate the subdiagonals of T. The idea of smoothing along the subdiagonals rather than down the rows or columns, or viewing T as a bivariate function is analogous to the successive regressions in (2.24). A similar procedure by Dahlhaus et al. (1997) uses varying coefficient regression models for each subdiagonal of T:

$$y_t = \sum_{j=1}^{t-1} f_j(t)\, y_{t-j} + \epsilon_t. \tag{2.35}$$

Wu and Pourahmadi (2003) give details of smoothing and selection of the order k of the autoregression under the assumption that the N subjects share common observation times. In the first step, they derive a raw estimate of the covariance matrix and the estimated covariance matrix is subject to the modified Cholesky decomposition. In the second step, they apply local polynomial smoothing to the diagonal elements of D and the subdiagonals of T.

The connection between the entries of T and the family of regression models (2.24) makes it conceivable that T exhibits sparsity, having some of its entries could be zero or close to zero. Smith and Kohn (2002) propose a prior distribution that allows for zero entries in T and have obtained a parsimonious model for Σ without assuming a parametric structure. Similar results are reported in Huang et al. (2006) using penalized likelihood with L_1-penalty to estimate T for Gaussian data. Similar in spirit to the tapering estimators based on the sample covariance matrix

(Section 2.2), several have proposed imposing sparsity by banding the Cholesky factor, including Wu and Pourahmadi (2003) and Huang et al. (2006). Levina et al. (2008) adaptively band the Cholesky factor using penalized maximum likelihood estimation. Their novel 'nested Lasso' penalty produces an estimator with an adaptive bandwidth for each row of the Cholesky factor. This structure has more flexibility than regular banding, but, unlike regular Lasso applied to the entries of the Cholesky factor, results in a sparse estimator for the inverse of the covariance matrix.

Incoherence of Generalized Autoregressive Parameters with Unbalanced Data

The aforementioned methods require balanced longitudinal data; it is unclear how they can be applied directly to irregular or incomplete data without 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.

Alternatively, the framework within which the data are generated may assume that a fixed number of measurements are to be collected at a common set of times for all subjects. In this case, unbalanced longitudinal data arises as a result of missing observations. To our knowledge, Huang et al. (2012) was the first to explicitly discuss the problems presented by unbalanced data within this framework, in the context of model (2.24). These issues are closely related to the ambiguity surrounding the definition of a discrete lag in the absense of a regular measurement grid, which Huang et al. (2012) refers to as incoherence in the autoregressive parameters (as well

as the prediction variances). They demonstrate incoherence with a simple example: let y_{it} denote the t^{th} repeated measurement on subject i. Consider modeling

$$y_{it} = \phi y_{i,t-1} + \epsilon_{it}, \tag{2.36}$$

for $t = 2, 3, 4$ with $y_{i1} = \epsilon_{i1}$, where $\epsilon_i = \left(\epsilon_{i1}, \ldots, \epsilon_{ip_i}\right)'$, $\epsilon_i \sim N(0, I)$. For a subject with a complete set of observations, the diagonal matrix of innovation variances is given by $D = I_4$, and the corresponding T and Σ are given by

$$T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \phi & 1 & 0 & 0 \\ 0 & \phi & 1 & 0 \\ 0 & 0 & \phi & 1 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & \phi & \phi^2 & \phi^3 \\ \phi & 1+\phi^2 & \phi^2+\phi^3 & \phi^3+\phi^4 \\ \phi^2 & \phi^2+\phi^3 & 1+\phi^2+\phi^4 & \phi+\phi^3+\phi^5 \\ \phi^3 & \phi^3+\phi^4 & \phi+\phi^3+\phi^5 & 1+\phi^2+\phi^4+\phi^6 \end{bmatrix}$$

Consider a pair of subjects, with Subject 1 having $p_1 = 3$ measurements at $t = 1, 2, 4$, and Subject 2 having $p_2 = 3$ measurements at $t = 1, 3, 4$. The covariance matrix for Subject 1, Σ_1, can be obtained by deletion of the third row and column of Σ, and similarly Σ_2 can be obtained by deletion of the second row and column of Σ. The Cholesky decompositions of the subject-specific covariance matrices are given by

$$T_1 = \begin{bmatrix} 1 & 0 & 0 \\ -\phi & 1 & 0 \\ 0 & -\phi^2 & 1 \end{bmatrix}, \quad D_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1+\phi^2 \end{bmatrix},$$

$$T_2 = \begin{bmatrix} 1 & 0 & 0 \\ -\phi^2 & 1 & 0 \\ 0 & -\phi & 1 \end{bmatrix}, \quad D_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1+\phi^2 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

The parameter ϕ_{ijk} denotes the coefficient associated with regressing the j^{th} measurement on the k^{th} measurement taken on subject i. For example, ϕ_{i21} is interpreted as the coefficient when regressing the second measurement on the first, they take different values for each subject. For Subject 1, the measurement at time 2 is regressed on the measurement at time 1, and for Subject 2, the measurement at time 3 is regressed on the measurement at time 1. This results in a discrepancy between the autoregressive coefficients, which are given by $\phi_{121} = \phi$ and $\phi_{221} = \phi^2$. There is similar discordance between the innovation variances.

This incoherence indicates that a naive approach to estimating the regression model (2.24) is inappropriate when the data are unbalanced. Huang et al. (2012) assume that there is a common set of observation times define a "grand" covariance matrix Σ, which is common to all subjects, the measurements on subject i can be modeled with covariance matrix Σ_i which is a principal minor of Σ. They propose handling data from longitudinal studies with dropouts and intermittent missing values by imputation, using the EM algorithm when the data are missing at random. Huang et al. (2007) employ a similar approach, assuming the same framework surrounding the data generation as Huang et al. (2012). They jointly model the mean and covariance matrix of longitudinal data using basis function expansions. They treat the subdiagonals of T as smooth functions which they approximate using B-splines, and carry out estimation via maximum (normal) likelihood. They regularize the estimated covariance matrix through the choice of k, the number of nonzero subdiagonals, and the total number of basis functions used to approximate the k smoothed diagonals, which are selected using Bayesian information criterion (BIC).

Chapter 3: A Reproducing Kernel Hilbert Space Framework for Covariance Estimation

We propose an alternate route for estimating the Cholesky decomposition of a covariance matrix when the data are unbalanced. In this chapter, we present a functional varying coefficient model to extend model (2.24). The functional varying coefficient model serves as a flexible alternative to parametric models for the GARPs and accommodates unbalanced data without the need for imputation. We propose a blueprint for the construction of an estimator of a covariance matrix for longitudinal data by modeling T as smooth two-dimensional surface, and present a reproducing kernel Hilbert space framework for estimating the functional components of the Cholesky decomposition. Chapter 4 demonstrates multidimensional smoothing with penalized B-splines as a flexible and computationally convenient alternative to the Hilbert space methods.

There has been substantial interest recently in the use of varying coefficient models for extending parametric models for longitudinal data (Noh and Park, 2010; Şentürk et al., 2013; Şentürk and Müller, 2008; Chiang et al., 2001; Hoover et al., 1998; Fan and Zhang, 1999). Given a sample of repeated measurements on N independent subjects, it is convenient to model the observed data collected on an individual as sampled from a realization of a continuous-time stochastic process $Y(t)$. Allowing individual-specific observation times, let $t_i = \left\{t_{i1} < \cdots < t_{i,p_i}\right\}$ denote the time

points at which the sequence of measurements on the i^{th} subject were taken, and let

$$Y_i = \left(y_{i1}, \ldots, y_{i,p_i}\right)'$$

denote the corresponding measurements, $i = 1, \ldots, N$. We assume that measurement times are drawn from some distribution having compact domain $\mathcal{T}$; without loss of generality, we take $\mathcal{T} = [0, 1]$. We use varying coefficient models to extend the linear model corresponding to the Cholesky decomposition (2.24). Consider the following model as a generalization of (2.24):

$$y\left(t_{ij}\right) = \sum_{k<j} \tilde{\phi}\left(t_{ij}, t_{ik}\right) y\left(t_{ik}\right) + \epsilon\left(t_{ij}\right), \qquad \begin{array}{l} i = 1, \ldots, N \\ j = 2, \ldots, p_i, \end{array} \tag{3.1}$$

where the prediction errors $\epsilon\left(t\right)$ follow a mean zero Gaussian process, with variance function $\sigma^2\left(t\right)$. The coefficient associated with regressing the measurement taken at time t on the measurement taken at time s is given by the value of the autoregressive coefficient function evaluated at (t, s):

$$\tilde{\phi}\left(t, s\right), \quad 0 \leq s < t \leq 1.$$

Under Model (3.1), the negative log likelihood satisfies

$$-2\ell\left(\phi, \sigma^2 | Y_1, \ldots, Y_N\right) = \sum_{i=1}^{N} \sum_{j=2}^{p_i} \log \sigma_{ij}^2 + \sum_{i=1}^{N} \sum_{j=2}^{p_i} \frac{1}{\sigma_{ij}^2} \left(y_{ij} - \sum_{k<j} \tilde{\phi}\left(t_{ij}, t_{ik}\right) y_{ik} \right)^2, \tag{3.2}$$

where $\sigma_{ij}^2 = \sigma^2\left(t_{ij}\right)$.

A number of methods have been developed for estimating the varying coefficients for the mean trajectory of repeated measurements. Wu and Pourahmadi (2003) and Dahlhaus et al. (1997) have explored one dimensional varying coefficient models for the Cholesky decomposition (2.35) for balanced longitudinal data. Writing the varying coefficient as a bivariate function, we can model unbalanced longitudinal data, and even accommodate longitudinal data for which there is no associated fixed set of observation times. Our goal is to use bivariate smoothing to estimate $\tilde{\phi}\left(t, s\right)$ for

$0 \leq s < t \leq 1$. In similar fashion, we estimate the innovation variance function $\sigma^2(t)$, $0 \leq t \leq 1$ by smoothing the squared prediction residuals as a function of t.

This model formulation grants access to the abundance of regularization techniques that are accessible in the usual function estimation setting. Nonparametric models are often used for "checking" or eliciting parametric models; see Cox et al. (1988) and Liu and Wang (2004). In this light, it is convenient to parameterize $\tilde{\phi}$ so that the fitted function can easily be used as a diagnostic tool or for suggesting parsimonious or structured models for the Cholesky decomposition. Given the prevalence of stationary covariance models in the applied literature, including those specifying the elements of T as a function of the lag between observations, we take a convenient parameterization of the varying coefficient function with inputs

$$l = t - s \text{ and } m = \frac{t+s}{2}, \tag{3.3}$$

and model

$$\phi(l, m) = \phi\left(t - s, \frac{1}{2}(s+t)\right) = \tilde{\phi}(t, s), \tag{3.4}$$

where l is the continuous analogue of the usual discrete lag between time points t and s, and m is its orthogonal direction. Stationary covariance models specify that the covariance between a pair of measurements taken at times t and s can be written as a function of $|t - s|$ only, so that

$$Cov(y(t), y(s)) = G(|t - s|)$$

for some positive definite function G. Model (3.1) corresponds to a stationary process when ϕ can be written as a function of l only and the innovation variances are constant in t. Taking stationarity as a form of simplicity or parsimony in covariance models, our approach is to regularize nonstationarity in the autoregressive varying coefficient and the innovation variance function so that simultaneous application of heavy penalization to both functions results in models that are close to stationary covariance matrices.

For estimation of ϕ, we employ the smoothing spline framework which can naturally incorporate structural differences in the functional components into modeling (see Kimeldorf and Wahba (1971) and Wahba (1990) for comprehensive presentation). To enhance the statistical interpretability of model parameters, we decompose ϕ into functional components similar to the notion of the main effect and the interaction terms in classical analysis of variance. We adopt the smoothing spline analogue of the classical ANOVA model proposed by Gu (2013), and estimation is achieved through similar computational strategies.

3.1 The Function Space for Smoothing Spline ANOVA Models

Smoothing spline ANOVA models (Gu, 2002) are a versatile family of smoothing methods that are applicable for both univariate and multivariate problems. These models are rooted in the theory of reproducing kernel Hilbert spaces and have been studied extensively for nonparametric function estimation (see Aronszajn (1950), Wahba (1990), and Berlinet and Thomas-Agnan (2011) for detailed examinations). However, to our knowledge, they have received little attention in the context of covariance modeling. Before we demonstrate the estimation of ϕ using a smoothing spline ANOVA model, we first must establish some notation and review the relevant mathematical details of reproducing kernel Hilbert spaces.

3.1.1 Properties of 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 $f \in \mathcal{H}$,

I. $K(\cdot, x) \in \mathcal{H}$

II. $f(x) = \langle f, K(\cdot, x) \rangle_{\mathcal{H}}$

The second property is called the reproducing property of K. Every reproducing kernel uniquely determines the RKHS, and in turn, every RKHS has 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 $\{a_1, \ldots, a_{n_1}\}$, $\{b_1, \ldots, b_{n_2}\} \in \Re$,

$$\langle \sum_{i=1}^{n_1} a_i K(\cdot, x_i), \sum_{j=1}^{n_2} b_j K(\cdot, u_j) \rangle_{\mathcal{H}} = \sum_i \sum_j a_i b_j K(x_i, u_j). \tag{3.5}$$

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

3.1.2 The Smoothing Spline Model Space

Suppose that $J(f)$ is a penalty functional defined on $\mathcal{H}$ measuring the roughness of f. When $J(f)$ is in the form of a squared semi-norm, it induces an orthogonal decomposition of $\mathcal{H}$. Let $\mathcal{H}_0 = \{f : J(f) = 0\}$ denote the null space of J, and consider the decomposition

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

where $\mathcal{H}_1$ is the subspace of $\mathcal{H}$ with $J(f)$ as its squared norm. For the cubic smoothing spline defined on $\chi = [0, 1]$, the roughness penalty corresponds to

$$J(f) = \int_0^1 (f''(x))^2 \, dx. \tag{3.6}$$

The penalty on the squared second derivative induces a decomposition of the function space

$$C^{(2)}[0, 1] = \left\{ f : \int_0^1 (f''(x))^2 \, dx < \infty \right\},$$

which is a Hilbert space if equipped with inner product

$$\langle f, g\rangle_{\mathcal{H}} = (M_0 f)(M_0 g) + (M_1 f)(M_1 g) + \int_0^1 f''(x)\, g''(x)\, dx, \tag{3.7}$$

where the i^{th} order differential operator M_i is given by $M_i f = \int_0^1 f^{(i)}(x)\, dx$.

Given inner product (3.7), the reproducing kernel K can be expressed in terms of the scaled Bernoulli polynomials $\left\{k_j(x) = \frac{1}{j!}B_j(x)\right\}$ for $x \in [0,1]$, where B_j is defined according to:

$$B_0(x) = 1$$
$$\frac{d}{dx}B_j(x) = jB_{j-1}(x),\ j = 1, 2, \ldots$$

One can verify that $\int_0^1 k_i^{(j)}(x)\, dx = \delta_{ij}$ for $i, j = 0, 1$, where δ_{ij} is the Kronecker delta. This implies that $\{k_0, k_1\}$ form an orthonormal basis for $\mathcal{H}_0 = \left\{f \in C^{(2)}[0,1] : f'' = 0\right\}$ under the inner product $\langle f, g\rangle_0 = (M_0 f)(M_0 g) + (M_1 f)(M_1 g)$ and that

$$K_0(x, y) = k_0(x)\, k_0(y) + k_1(x)\, k_1(y)$$

is the reproducing kernel for $\mathcal{H}_0$. One can further decompose $\mathcal{H}_0$ into the tensor sum of the subspaces spanned by k_0 and k_1:

$$\mathcal{H}_0 = \mathcal{H}_{00} \oplus \mathcal{H}_{01} = \{f :\ f \propto 1\} \oplus \{f :\ f \propto k_1\}, \tag{3.8}$$

where the corresponding reproducing kernels for each subspace are given by 1 and $k_1(x)\, k_1(y)$, respectively. The subspaces of $\mathcal{H}$ which are orthogonal to $\mathcal{H}_0$ are comprised of functions f satisfying

$$\mathcal{H}_1 = \{f : M_0 f = M_1 f = 0, \quad \int_0^1 (f''(x))^2\ dx < \infty\}.$$

One can show that the representer for the evaluation functional $[x](\cdot)$ in $\mathcal{H}_1$ with inner product $\langle f, g\rangle_{\mathcal{H}_1} = \int_0^1 f''(x)\, g''(x)\ dx$ is given by the function

$$K_1(x, y) = k_2(x)\, k_2(y) - k_4(x - y). \tag{3.9}$$

See Example 2.3.3 in Gu (2002) for proof. It is obvious that $\mathcal{H}_0 \bigcap \mathcal{H}_1 = \{0\}$, so the converse of Theorem 2.5 in Gu (2013) gives us that the reproducing kernel for the full space

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1, \tag{3.10}$$

is given by $K = K_0 + K_1$. Using the decomposition of $\mathcal{H}_0$ into the constant and linear subspaces in (3.8), we can further decompose $\mathcal{H}$ into

$$\mathcal{H} = \mathcal{H}_{00} \oplus \mathcal{H}_{01} \oplus \mathcal{H}_1, \tag{3.11}$$

where $\mathcal{H}_{01} \oplus \mathcal{H}_1$ forms the contrast in a one-way ANOVA decomposition with averaging operator $\mathcal{A}f = \int_0^1 f(x)\ dx$. The reproducing kernel $K = K_{00} + K_{01} + K_1$ can be defined in terms of the corresponding reproducing kernels

$$\begin{aligned} K_{00}(x, y) &= 1, \\ K_{01}(x, y) &= k_1(x) k_1(y), \text{ and} \\ K_1(x, y) &= k_2(x) k_2(y) - k_4(x - y). \end{aligned} \tag{3.12}$$

The kernel K_{00} generates the "mean" space. Together, the kernels K_{01} and K_1 generate the "contrast" space, with K_{01} contributing to the "parametric contrast" and K_1 to the "nonparametric contrast."

3.1.3 The Tensor Product Smoothing Spline Model Space

To estimate a bivariate function using the ANOVA decomposition given in (3.11), one may construct a tensor product reproducing kernel Hilbert space. The space can be constructed through the reproducing kernel, which is constructed using the reproducing kernels on each of the marginal domains. One-way ANOVA decompositions on the marginal domains naturally induce an ANOVA decomposition on the product domain. It can be shown that the products of reproducing kernels on the marginal domains form reproducing kernels on the product domain; see Theorem 2.6 in Gu (2013).

Let $\mathcal{H}_{[1]}$ and $\mathcal{H}_{[2]}$ denote reproducing kernel Hilbert spaces on marginal domains $[0, 1]$ equipped with corresponding reproducing kernels K_1 and K_2, each defined as in (3.12). The RKHS corresponding to the tensor product smoothing spline is given by

$$\mathcal{H} = \mathcal{H}_{[1]} \otimes \mathcal{H}_{[2]}$$

and has reproducing kernel

$$K(\boldsymbol{x}, \boldsymbol{y}) = K_1(x_1, y_1) K_2(x_2, y_2),$$

where $\boldsymbol{x} = (x_1, x_2)$ and $\boldsymbol{y} = (y_1, y_2)$.

The tensor product space can be constructed with nine tensor sum terms, which are defined by the decomposition of the marginal subspaces

$$\mathcal{H}_{[i]} = \mathcal{H}_{00[i]} \oplus \mathcal{H}_{01[i]} \oplus \mathcal{H}_{1[i]}, \quad i = 1, 2.$$

Table 3.1 gives the tensor sum terms defining the decomposition of $\mathcal{H}$ and the functional components corresponding to each subspace. The reproducing kernels for each of the subspaces are given in Table 3.2.

Table 3.1: *Construction of the tensor product cubic spline function space from marginal subspaces $\mathcal{H}_{[1]}$, $\mathcal{H}_{[2]}$ and the corresponding functional components, where "n" and "p" mean "parametric" and "nonparametric," respectively.*

	$\mathcal{H}_{00[2]}$	$\mathcal{H}_{01[2]}$	$\mathcal{H}_{1[2]}$
$\mathcal{H}_{00[1]}$	$\mathcal{H}_{00[1]} \otimes \mathcal{H}_{00[2]}$	$\mathcal{H}_{00[1]} \otimes \mathcal{H}_{01[2]}$	$\mathcal{H}_{00[1]} \otimes \mathcal{H}_{1[2]}$
$\mathcal{H}_{01[1]}$	$\mathcal{H}_{01[1]} \otimes \mathcal{H}_{00[2]}$	$\mathcal{H}_{01[1]} \otimes \mathcal{H}_{01[2]}$	$\mathcal{H}_{01[1]} \otimes \mathcal{H}_{1[2]}$
$\mathcal{H}_{1[1]}$	$\mathcal{H}_{1[1]} \otimes \mathcal{H}_{00[2]}$	$\mathcal{H}_{1[1]} \otimes \mathcal{H}_{01[2]}$	$\mathcal{H}_{1[1]} \otimes \mathcal{H}_{1[2]}$

	$\{1\}$	$\{k_1\}$	$\mathcal{H}_{1[2]}$
$\{1\}$	mean	p-main effect	np-main effect
$\{k_1\}$	p-main effect	$p \times p$-interaction	$p \times np$-interaction
$\mathcal{H}_{1[1]}$	np-main effect	$np \times p$-interaction	$np \times np$-interaction

Table 3.2: *Reproducing kernels corresponding to the subspaces for the cubic tensor product smoothing spline given in Table 3.1.*

Subspace	Reproducing kernel
$\mathcal{H}_{00[1]} \otimes \mathcal{H}_{00[2]}$	1
$\mathcal{H}_{01[1]} \otimes \mathcal{H}_{00[2]}$	$k_1(x_1)\, k_1(y_1)$
$\mathcal{H}_{00[1]} \otimes \mathcal{H}_{01[2]}$	$k_1(x_2)\, k_1(y_2)$
$\mathcal{H}_{01[1]} \otimes \mathcal{H}_{01[2]}$	$k_1(x_1)\, k_1(y_1)\, k_1(x_2)\, k_1(y_2)$
$\mathcal{H}_{1[1]} \otimes \mathcal{H}_{00[2]}$	$k_2(x_1)\, k_2(y_1) - k_4(x_1 - y_1)$
$\mathcal{H}_{00[1]} \otimes \mathcal{H}_{1[2]}$	$k_2(x_2)\, k_2(y_2) - k_4(x_2 - y_2)$
$\mathcal{H}_{1[1]} \otimes \mathcal{H}_{01[2]}$	$[k_2(x_1)\, k_2(y_1) - k_4(x_1 - y_1)]\, k_1(x_2)\, k_1(y_2)$
$\mathcal{H}_{01[1]} \otimes \mathcal{H}_{1[2]}$	$k_1(x_1)\, k_1(y_1)\, [k_2(x_2)\, k_2(y_2) - k_4(x_2 - y_2)]$
$\mathcal{H}_{1[1]} \otimes \mathcal{H}_{1[2]}$	$[k_2(x_1)\, k_2(y_1) - k_4(x_1 - y_1)]\, [k_2(x_2)\, k_2(y_2) - k_4(x_2 - y_2)]$

The penalty functional driving the ANOVA decomposition of the marginal subspaces can be generalized to penalize the m^{th} order derivative by letting

$$J(f) = \int_0^1 \left(f^{(m)}(x)\right)^2 \, dx.$$

For example, letting $m = 1$ corresponds to the space for a linear smoothing spline, where the null space of the penalty functional is spanned by constant functions. For detailed derivations of the smoothing spline ANOVA decomposition with arbitrary penalty order m, we refer the reader to Chapter 2 in Gu (2013).

3.1.4 A General Form for Multiple-Term Reproducing Kernel Hilbert Spaces

The previous construction of the RKHS for the tensor product cubic spline space contains multiple tensor sum terms. We can write

$$\mathcal{H} = \bigoplus_{\beta} \mathcal{H}_\beta, \tag{3.13}$$

where β is a generic index. The subspaces $\mathcal{H}_\beta$ have reproducing kernels K_β and corresponding inner products $\langle f_\beta, g_\beta \rangle_{\mathcal{H}_\beta}$, where $f_\beta = P_\beta f$ denotes the projection of f into the subspace $\mathcal{H}_\beta$. For example, one can write the RKHS for the tensor product smoothing spline according to (3.13) using the subspaces given in Table 3.2.

The subspaces $\mathcal{H}_\beta$ are independent modules, and the inner products $\langle f_\beta, g_\beta \rangle_{\mathcal{H}_\beta}$ are not necessarily comparable between subspaces. To standardize across the subspaces, an inner product in $\mathcal{H}$ can be specified via

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\beta} \theta_\beta^{-1} \langle f, g \rangle_{\mathcal{H}_\beta}. \tag{3.14}$$

where $\theta_\beta \in (0, \infty)$ are additional smoothing parameters. The corresponding reproducing kernel for $\mathcal{H}$ is given by

$$K = \sum_{\beta} \theta_\beta K_\beta, \tag{3.15}$$

which can be used to specify the penalty $J(f)$. Subspaces which don't contribute to $J(f)$ form $\mathcal{H}_0 = \{f : J(f) = 0\}$, the null space of $J(f)$. The subspaces contributing to $J(f)$ form the space $\mathcal{H}_1 = \mathcal{H} \ominus \mathcal{H}_0$, in which $J(f)$ is a squared norm. For this specification, denote the penalty constructed as such by

$$J(f) = \sum_{\beta} \theta_{\beta}^{-1} \langle f_{\beta}, f_{\beta} \rangle_{\mathcal{H}_{\beta}} = \sum_{\beta} \theta_{\beta}^{-1} J_{\beta}(f_{\beta}). \tag{3.16}$$

The $\{\theta_{\beta}\}$ are implicit in notation henceforth to permit ease of exposition.

3.2 A Reproducing Kernel Hilbert Space Framework for the Generalized Autoregressive Varying Coefficient

We can construct the model space for the generalized autoregressive varying coefficient ϕ using the previous recipe for constructing a tensor product RKHS. Let $\mathcal{H}_{[l]}$ denote the RKHS for the domain of $l \in [0,1]$ with reproducing kernel $K_{[l]}$, and similarly, let $\mathcal{H}_{[m]}$ denote the RKHS for the domain of $m \in [0,1]$ with reproducing kernel $K_{[m]}$. The function space for $\phi(l,m) \in \mathcal{H}$

$$\mathcal{H} = \mathcal{H}_{[l]} \otimes \mathcal{H}_{[m]} = \mathcal{H}_0 \oplus \mathcal{H}_1$$

is obtained as in Section 3.1, with reproducing kernel $K = K_{[l]} K_{[m]}$.

Let $\boldsymbol{v}_{ijk} = \left(t_{ij} - t_{ik}, \frac{1}{2}(t_{ij} + t_{ik})\right) = (l_{ijk}, m_{ijk})$, $\boldsymbol{v}_{ijk} \in \mathcal{V} = [0,1]^2$ denote the tuple corresponding to the transformed pair of observation times. Fixing the innovation variances $\sigma_{ij}^2 = \sigma^2(t_{ij})$ in (3.2), the negative log likelihood satsifies

$$-2\ell\left(\phi | Y_1, \ldots, Y_N, \sigma^2\right) = \sum_{i=1}^{N} \sum_{j=2}^{p_i} \frac{1}{\sigma_{ij}^2} \left(y_{ij} - \sum_{k<j} \phi(\boldsymbol{v}_{ijk}) y_{ik} \right)^2. \tag{3.17}$$

The roughness penalty associated with reproducing kernel K can be written as $J(\phi) = ||P_1 \phi||^2$, the squared norm of the projection of ϕ onto $\mathcal{H}_1$. Appending this to (3.17), the penalized negative

log likelihood may be written

$$-2\ell\left(\phi|Y_1,\ldots,Y_N,\sigma^2\right)+\lambda J\left(\phi\right)=\sum_{i=1}^{N}\sum_{j=2}^{p_i}\frac{1}{\sigma_{ij}^2}\left(y_{ij}-\sum_{k<j}\phi\left(\boldsymbol{v}_{ijk}\right)y_{ik}\right)^2+\lambda||P_1\phi||^2. \quad (3.18)$$

3.2.1 A Representer Theorem

Wahba (1990) established an explicit form for the minimizer of the penalized sums of squares in the usual function estimation setting. The following theorem establishes the form for the minimizer of the penalized negative log likelihood (3.18) for the varying coefficient model (3.1). Define

$$V=\bigcup_{i,j,k}\{\boldsymbol{v}_{ijk}\}\equiv\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_{|V|}\}$$

as the set of unique within-subject pairs of observation times.

Theorem 3.2.1. *Let* $\{\nu_1,\ldots,\nu_{\mathcal{N}_0}\}$ *span* $\mathcal{H}_0$*, the null space of* $J\left(\phi\right)=||P_1\phi||^2$*. Then the minimizer* ϕ_λ *of* (3.18) *is given by*

$$\phi_\lambda\left(\boldsymbol{v}\right)=\sum_{i=1}^{\mathcal{N}_0}d_i\nu_i\left(\boldsymbol{v}\right)+\sum_{j=1}^{|V|}c_jK_1\left(\boldsymbol{v}_j,\boldsymbol{v}\right), \quad (3.19)$$

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, which is similar in spirit to the proof of Theorem 1.3.1 in Wahba (1990), can be found in Appendix A.

3.2.2 Model Fitting

Let Y denote the vector of length $n_Y=\sum_i p_i-N$ constructed by stacking the N observed response vectors $Y_1,\ldots,Y_N$ less their first element y_{i1} one on top of each other:

$$Y=\left(Y_1',Y_2',\ldots,Y_N'\right)' \quad (3.20)$$

$$=\left(y_{12},y_{13},\ldots,y_{1p_1},\ldots,y_{N2},\ldots,y_{Np_N}\right)'. \quad (3.21)$$

Define X_i to be the $(p_i - 1) \times |V|$ matrix containing the covariates necessary for regressing each measurement $y_{i2}, \ldots, y_{i,p_i}$ on its predecessors as in Model (3.1), and stack these on top of one another to obtain

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}, \tag{3.22}$$

which has dimension $n_Y \times |V|$. Using the Representer Theorem in (3.19), the penalized negative log likelihood in (3.18) can be expressed as

$$-2\ell\left(\phi|Y_1, \ldots, Y_N, \sigma^2\right) + \lambda J\left(\phi\right) = ||D^{-1/2}\left(Y - X\left(Bd + K_{_V}c\right)\right)||^2 + \lambda c' K_{_V} c, \tag{3.23}$$

where the (i, j) entry of the $|V| \times |V|$ matrix $K_{_V}$ is given by $K_1\left(\boldsymbol{v}_i, \boldsymbol{v}_j\right)$, and the $|V| \times \mathcal{N}_0$ matrix B has (i, j) element equal to $\nu_j\left(\boldsymbol{v}_i\right)$. The diagonal matrix D holds the n_Y innovation variances σ^2_{ij}. The following examples demonstrate how to construct the subject-specific design matrices $X_1, \ldots, X_N$ when observation times are common across all subjects and when observation times are subject-specific.

Example 1. Construction of X_i with complete data

Construction of the autoregressive design matrix X_i is straightforward in the case that there are an equal number of measurements on each subject at a common set of measurement times $t_1, \ldots, t_p$. When complete data are available for measurement times $t_1, \ldots, t_p$,

$$X_i = \begin{bmatrix} y_{i1} & 0 & 0 & \ldots & & \ldots & 0 \\ 0 & y_{i1} & y_{i2} & & & \ldots & 0 \\ \vdots & & & \ddots & & & \\ 0 & 0 & \ldots & & y_{i1} & \ldots & y_{i,p-1} \end{bmatrix} \tag{3.24}$$

for all $i = 1, \ldots, N$. Note that this design matrix specification does not require that measurement times be regularly spaced.

Example 2. Construction of X_i with incomplete data

We demonstrate the construction of the autoregressive design matrices when subjects do not share a universal set of observation times for $N = 2$; the construction extends naturally for an arbitrary number of trajectories. Let subjects have corresponding sample sizes $p_1 = 4$, $p_2 = 4$, with measurements on subject 1 taken at $t_{11} = 0, t_{12} = 0.2, t_{13} = 0.5, t_{14} = 0.9$ and on subject 2 taken at $t_{21} = 0, t_{22} = 0.1, t_{23} = 0.5, t_{24} = 0.7$. Then the unique within-subject pairs of observation times (t, s) such that $0 \leq s < t \leq 1$ are given by

i	2	1	1, 2	2	1	2	2	2	1	1	1
t	0.1	0.2	0.5	0.5	0.5	0.7	0.7	0.7	0.9	0.9	0.9
s	0.0	0.0	0.0	0.1	0.2	0.0	0.1	0.5	0.0	0.2	0.5

Here, the top row indicates which subject was observed at each pair (t, s). This gives that $V = \{\boldsymbol{v}_{121}, \ldots, \boldsymbol{v}_{143}\} \bigcup \{\boldsymbol{v}_{221}, \ldots, \boldsymbol{v}_{243}\} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{11}\}$, where the distinct observed $\boldsymbol{v} = (l, m)$ are

l	0.10	0.20	0.50	0.40	0.30	0.70	0.60	0.20	0.90	0.70	0.40
m	0.05	0.10	0.25	0.30	0.35	0.35	0.40	0.60	0.45	0.55	0.70

Then a potential construction of the autoregressive design matrix for subject is given by:

$$X_1 = \begin{bmatrix} 0 & y_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & y_{11} & 0 & y_{12} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & y_{11} & y_{12} & y_{13} \end{bmatrix}$$

and similarly, for subject 2:

$$X_2 = \begin{bmatrix} y_{21} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & y_{21} & y_{22} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & y_{21} & y_{22} & y_{23} & 0 & 0 & 0 \end{bmatrix}$$

Finding the Solution

Defining $\tilde{Y} = D^{-1/2}Y$, $\tilde{B} = D^{-1/2}XB$, and $\tilde{K}_{_V} = D^{-1/2}XK_{_V}$, the penalized negative log likelihood (3.23) may be written

$$-2\ell\left(c, d|\tilde{Y}, \tilde{B}, \tilde{K}_{_V}\right) + \lambda J\left(\phi\right) = \left[\tilde{Y} - \tilde{B}d - \tilde{K}_{_V}c\right]'\left[\tilde{Y} - \tilde{B}d - \tilde{K}_{_V}c\right] + \lambda c' K_{_V} c. \tag{3.25}$$

Taking partial derivatives with respect to d and c and setting them equal to zero yields normal equations:

$$\begin{aligned} \tilde{B}'\tilde{B}d + \tilde{B}'\tilde{K}_{_V}c &= \tilde{B}'\tilde{Y} \\ \tilde{K}'_{_V}\tilde{B}d + \tilde{K}'_{_V}\tilde{K}_{_V}c + \lambda K_{_V}c &= \tilde{K}'_{_V}\tilde{Y}. \end{aligned} \tag{3.26}$$

Thus, for fixed smoothing parameters, the solution ϕ is obtained by finding c and d which satisfy

$$\begin{bmatrix} \tilde{B}'\tilde{B} & \tilde{B}'\tilde{K}_{_V} \\ \tilde{K}'_{_V}\tilde{B} & \tilde{K}'_{_V}\tilde{K}_{_V} + \lambda K_{_V} \end{bmatrix} \begin{bmatrix} d \\ c \end{bmatrix} = \begin{bmatrix} \tilde{B}'\tilde{Y} \\ \tilde{K}'_{_V}\tilde{Y} \end{bmatrix}. \tag{3.27}$$

Fixing smoothing parameters λ and θ_β (hidden in $K_{_V}$ and $\tilde{K}_{_V}$ if present), assuming that $\tilde{K}_{_V}$ is full column rank, (3.27) can be solved by the Cholesky decomposition of the $(\mathcal{N}_0 + |V|) \times (\mathcal{N}_0 + |V|)$ matrix, followed by forward and backward substitution. See Golub and Van Loan (2012). Singularity of $\tilde{K}_{_V}$ demands special consideration. Write the Cholesky decomposition

$$\begin{bmatrix} \tilde{B}'\tilde{B} & \tilde{B}'\tilde{K}_{_V} \\ \tilde{K}'_{_V}\tilde{B} & \tilde{K}'_{_V}\tilde{K}_{_V} + \lambda K_{_V} \end{bmatrix} = \begin{bmatrix} C'_1 & 0 \\ C'_2 & C'_3 \end{bmatrix} \begin{bmatrix} C_1 & C_2 \\ 0 & C_3 \end{bmatrix} \tag{3.28}$$

where $\tilde{B}'\tilde{B} = C'_1C_1$, $C_2 = \left(C'_1\right)^{-1}\tilde{B}'\tilde{K}_{_V}$, and $C'_3C_3 = \lambda K_{_V} + \tilde{K}'_{_V}\left(I - \tilde{B}\left(\tilde{B}'\tilde{B}\right)^{-1}\tilde{B}'\right)\tilde{K}_{_V}$. Using an exchange of indices known as pivoting, one may write

$$C_3 = \begin{bmatrix} H_1 & H_2 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} H \\ 0 \end{bmatrix},$$

where H_1 is nonsingular. Define

$$\tilde{C}_3 = \begin{bmatrix} H_1 & H_2 \\ 0 & \delta I \end{bmatrix}, \quad \tilde{C} = \begin{bmatrix} C_1 & C_2 \\ 0 & \tilde{C}_3 \end{bmatrix}; \tag{3.29}$$

then

$$\tilde{C}^{-1} = \begin{bmatrix} C_1^{-1} & -C_1^{-1}C_2\tilde{C}_3^{-1} \\ 0 & \tilde{C}_3^{-1} \end{bmatrix}. \tag{3.30}$$

Premultiplying (3.28) by $(\tilde{C}')^{-1}$, straightforward algebra gives

$$\begin{bmatrix} I & 0 \\ 0 & (\tilde{C}_3')^{-1}C_3'C_3\tilde{C}_3^{-1} \end{bmatrix} \begin{bmatrix} \tilde{d} \\ \tilde{c} \end{bmatrix} = \begin{bmatrix} (C_1')^{-1}\tilde{B}'\tilde{Y} \\ (\tilde{C}_3')^{-1}\tilde{K}_V' \left(I - \tilde{B}\left(\tilde{B}'\tilde{B}\right)^{-1}\tilde{B}' \right) \tilde{Y} \end{bmatrix} \tag{3.31}$$

where $\left(\tilde{d}' \;\; \tilde{c}'\right)' = \tilde{C}'\,(d \;\; c)'$. Partition $\tilde{C}_3 = \begin{bmatrix} F & L \end{bmatrix}$; then $HF = I$ and $HL = 0$. So

$$\begin{aligned} (\tilde{C}_3')^{-1}C_3'C_3\tilde{C}_3^{-1} &= \begin{bmatrix} F' \\ L' \end{bmatrix} C_3'C_3 \begin{bmatrix} F & L \end{bmatrix} \\ &= \begin{bmatrix} F' \\ L' \end{bmatrix} H'H \begin{bmatrix} F & L \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

If $L'C_3'C_3L = 0$, then $L'\tilde{K}_V' \left(I - \tilde{B}\left(\tilde{B}'\tilde{B}\right)^{-1}\tilde{B}' \right) \tilde{K}_V L = 0$, so $L'\tilde{K}_V' \left(I - \tilde{B}\left(\tilde{B}'\tilde{B}\right)^{-1}\tilde{B}' \right) \tilde{Y} = 0$. Thus, the linear system has form

$$\begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{d} \\ \tilde{c}_1 \\ \tilde{c}_2 \end{bmatrix} = \begin{bmatrix} * \\ * \\ 0 \end{bmatrix}, \tag{3.32}$$

which can be solved, but with $\tilde{c}_2$ arbitrary. One may perform the Cholesky decomposition of (3.27) with pivoting, replace the trailing 0 with δI for appropriate value of δ, and proceed as if $\tilde{K}_V$ were of full rank.

Solving for the coefficients gives

$$\begin{bmatrix} \hat{d} \\ \hat{c} \end{bmatrix} = \tilde{C}^{-1}(\tilde{C}')^{-1} \begin{bmatrix} \tilde{B}' \\ \tilde{K}_V' \end{bmatrix} \tilde{Y}. \tag{3.33}$$

It follows that

$$\widehat{\tilde{Y}} = \tilde{B}\hat{d} + \tilde{K}_V\hat{c} = \begin{bmatrix} \tilde{B} & \tilde{K}_V \end{bmatrix} \tilde{C}^{-1}(\tilde{C}')^{-1} \begin{bmatrix} \tilde{B}' \\ \tilde{K}_V' \end{bmatrix} \tilde{Y} = \tilde{A}_{\lambda,\boldsymbol{\theta}}\tilde{Y}, \tag{3.34}$$

where

$$\begin{aligned}\tilde{A}_{\lambda,\boldsymbol{\theta}} &= \begin{bmatrix}\tilde{B} & \tilde{K}_V\end{bmatrix} \tilde{C}^{-1}(\tilde{C}')^{-1} \begin{bmatrix}\tilde{B}' \\ \tilde{K}_V'\end{bmatrix} \\ &= G + (I-G)\,\tilde{K}_V \left[\tilde{K}_V'\,(I-G)\,\tilde{K}_V + \lambda K_V\right]^{-1} \tilde{K}_V'\,(I-G)\,,\end{aligned} \tag{3.35}$$

for $G = \tilde{B}\left(\tilde{B}'\tilde{B}\right)^{-1}\tilde{B}'$.

3.2.3 Smoothing Parameter Selection

By varying smoothing parameters λ and θ_β, the minimizer ϕ_λ of (3.27) defines a family of potential estimates. In practice, we need to choose a specific estimate from the family, which requires effective methods for smoothing parameter selection. We consider two criteria that are commonly used for smoothing parameter selection in the context of smoothing spline models for longitudinal data. The first score is an unbiased estimate of a relative loss and assumes known variances σ_t^2. The unbiased risk estimate has attractive asymptotic properties; see Gu (2013) for a comprehensive examination. The second score, the leave-one-subject-out cross validation (LosoCV) score, provides an estimate of the same loss without assuming a known variance function. We review a computationally convenient approximation of the LosoCV score proposed by Xu et al. (2012), who demonstrate the shortcut score's asymptotic optimality. To simplify notation for the initial presentation, we only make explicit the dependence of estimates and their components on λ and conceal any dependence on θ_β.

Unbiased Risk Estimate

Define $\tilde{Y}$, $\tilde{B}$, and $\tilde{K}$ as before. Let $\tilde{\epsilon} = D^{-1/2}\epsilon$ denote the vector of length $n_Y = \sum_{i=1}^N p_i - N$ containing the standardized prediction errors $\tilde{\epsilon}_{ij} \sim N(0,1)$. Let $\mu = E[Y|X] = X\Phi$ denote the mean vector of Y conditional on its predecessors, where Φ is the $|V| \times 1$ vector resulting from the

evaluation of ϕ at $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{|V|} \in V$. The elements of $\mu = \left(\mu_{11}, \ldots, \mu_{N p_N}\right)'$ are given by

$$\begin{aligned}\mu_{ij} &= E\left[y_{ij} | y_{i1}, \ldots, y_{i,j-1}\right] \\ &= \sum_{k<j} \phi\left(\boldsymbol{v}_{ijk}\right) y_{ik}, \quad i = 1, \ldots, N,\ j = 2, \ldots, p_i.\end{aligned}$$

Let $\tilde{X} = D^{-1/2} X$, and let $\tilde{\mu} = E\left[\tilde{Y} | \tilde{X}\right] = \tilde{X}\Phi$ denote the conditional mean vector of the standardized observations. We can assess $\hat{\tilde{Y}}_\lambda$, an estimate of the mean of $\tilde{Y}$ based on observed data $\{y_{ij}\}$ using the loss function

$$\begin{aligned}L(\lambda) &= \sum_{i=1}^{N} \sum_{j=1}^{p_i} \left(\hat{\tilde{y}}_{ij} - \tilde{\mu}_{ij}\right)^2 \\ &= ||\widehat{\tilde{Y}} - \tilde{\mu}||^2.\end{aligned} \tag{3.36}$$

Writing $\widehat{\tilde{Y}} = \tilde{A}_{\lambda,\boldsymbol{\theta}} \tilde{Y}$, straightforward algebra yields that

$$L(\lambda) = \tilde{\mu}' \left(I - \tilde{A}_{\lambda,\boldsymbol{\theta}}\right)^2 \tilde{\mu} - 2\tilde{\mu}' \left(I - \tilde{A}_{\lambda,\boldsymbol{\theta}}\right)^2 \tilde{A}_{\lambda,\boldsymbol{\theta}} \tilde{\epsilon} + \tilde{\epsilon}' \tilde{A}^2_{\lambda,\boldsymbol{\theta}} \tilde{\epsilon} \tag{3.37}$$

Define the unbiased risk estimate

$$U(\lambda) = \tilde{Y}' \left(I - \tilde{A}_{\lambda,\boldsymbol{\theta}}\right)^2 \tilde{Y} + 2 \text{ tr} \left(\tilde{A}_{\lambda,\boldsymbol{\theta}}\right).$$

Adding $\tilde{\mu}$ to and subtracting $\tilde{\mu}$ from the quadratic terms, one can verify with straightforward algebra that

$$\begin{aligned}U(\lambda) &= \left(\tilde{Y} - \tilde{\mu} + \tilde{\mu} - \tilde{A}_{\lambda,\boldsymbol{\theta}} \tilde{Y}\right)' \left(\tilde{Y} - \tilde{\mu} + \tilde{\mu} - \tilde{A}_{\lambda,\boldsymbol{\theta}} \tilde{Y}\right) + 2 \text{ tr} \left(\tilde{A}_{\lambda,\boldsymbol{\theta}}\right) \\ &= \left(\tilde{A}_{\lambda,\boldsymbol{\theta}} \tilde{Y} - \tilde{\mu}\right)' \left(\tilde{A}_{\lambda,\boldsymbol{\theta}} \tilde{Y} - \tilde{\mu}\right) + \tilde{\epsilon}'\tilde{\epsilon} + 2\tilde{\epsilon}' \left(I - \tilde{A}_{\lambda,\boldsymbol{\theta}}\right) \tilde{\mu} - 2\left(\tilde{\epsilon}' \tilde{A}_{\lambda,\boldsymbol{\theta}} \tilde{\epsilon} - \text{tr}\left(\tilde{A}_{\lambda,\boldsymbol{\theta}}\right)\right).\end{aligned}$$

This gives

$$U(\lambda) - L(\lambda) - \tilde{\epsilon}'\tilde{\epsilon} = 2\tilde{\epsilon}' \left(I - \tilde{A}_{\lambda,\boldsymbol{\theta}}\right) \tilde{\mu} - 2\left(\tilde{\epsilon}' \tilde{A}_{\lambda,\boldsymbol{\theta}} \tilde{\epsilon} - \text{tr}\left(\tilde{A}_{\lambda,\boldsymbol{\theta}}\right)\right),$$

where $\tilde{\epsilon} = (\tilde{\epsilon}_{11}, \ldots, \tilde{\epsilon}_{Npn})'$ denotes the vector of standardized errors $\tilde{\epsilon}_{ij} = \sigma_{ij}^{-2}\left[y_{ij} - \sum_{k<j}\phi\left(\boldsymbol{v}_{ijk}\right)y_{ik}\right]$. This shows that $U\left(\lambda\right)$ is unbiased for the relative loss $L\left(\lambda\right) + \tilde{\epsilon}'\tilde{\epsilon}$. Under mild conditions on the risk function

$$R\left(\lambda\right) = E\left[L\left(\lambda\right)\right],$$

one can establish that U is also a consistent estimator. See Chapter 3 of Gu (2013) for a formal theorem and proof.

Leave-one-subject-out Cross Validation

The conditions under which the the cross validation and generalized cross validation scores traditionally used for smoothing parameter selection yield desirable properties generally do not hold when the data are clustered or longitudinal in nature. Instead, the leave-one-subject-out (LosoCV) cross validation score has been widely used for smoothing parameter selection for semiparametric and nonparametric models for longitudinal or functional data. The LosoCV criterion is defined as

$$V_{loso}\left(\lambda\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{Y}_i - \widehat{\tilde{\mu}}_i^{[-i]}\right)'\left(\tilde{Y}_i - \widehat{\tilde{\mu}}_i^{[-i]}\right) \tag{3.38}$$

where $\widehat{\tilde{\mu}}_i^{[-i]}$ is the estimate of $E\left[\tilde{Y}_i|X_i\right]$ based on the data when $\tilde{Y}_i$ is omitted. Intuitively, the LosoCV score is appealing because it preserves any within-subject dependence by leaving out all observations from the same subject together in the cross-validation. However, despite its prevalent use, theoretical justifications for its use have not been established. In their seminal work, Rice and Silverman (1991) were the first to present a heuristic justification of LosoCV by demonstrating that it mimics the mean squared prediction error. Consider new observations $\tilde{Y}_i^* = \left(\tilde{y}_{i1}^*, \tilde{y}_{i1}^*, \ldots, \tilde{y}_{i,p_i}^*\right)$. We may write the mean squared prediction error for the new observations as follows:

$$
\begin{aligned}
MSPE &= \frac{1}{N}\sum_{i=1}^{N} E\left[||\tilde{Y}_i^* - \widehat{\tilde{\mu}}_i||^2\right] \\
&= \frac{1}{N}\sum_{i=1}^{N} E\left[||\tilde{Y}_i^* - \tilde{\mu}_i + \tilde{\mu}_i - \widehat{\tilde{\mu}}_i||^2\right] \\
&= \frac{1}{N}\sum_{i=1}^{N} \left\{p_i + E\left[||\tilde{\mu}_i - \widehat{\tilde{\mu}}_i||^2\right]\right\}
\end{aligned} \tag{3.39}
$$

When $\{\sigma^2(t)\}$ is known, $\tilde{\epsilon}_i$ is a mean zero multivariate normal vector with $Cov(\tilde{\epsilon}_i) = I_{p_i}$, which gives the last equality. Since $\tilde{Y}_i$ and $\widehat{\tilde{\mu}}_i^{[-i]}$ are independent, the expected LosoCV score can be written

$$
E[V_{loso}(\lambda)] = \frac{1}{N}\sum_{i=1}^{N} \left\{p_i + E\left[||\widehat{\tilde{\mu}}_i^{[-i]} - \tilde{\mu}_i||^2\right]\right\}. \tag{3.40}
$$

When N is large, we expect that $\widehat{\tilde{\mu}}_i$ should be close to $\widehat{\tilde{\mu}}_i^{[-i]}$, so $E[V_{loso}(\lambda)]$ should be a good approximation to the mean-squared prediction error. For a formal proof of consistency, see Xu et al. (2012).

The definition of V_{loso} would lead one to initially believe that calculation of the score requires solving N separate minimization problems. However, Xu et al. (2012) established a computational shortcut that requires solving only one minimization problem that involves all data.

Lemma 3.2.2 (Shortcut formula for LosoCV)**.** *The LosoCV score satisfies the following identity:*

$$
V_{loso}(\lambda) = \frac{1}{N}\sum_{i=1}^{N} \left(\tilde{Y}_i - \widehat{\tilde{Y}}_i\right)' \left(I_{p_i} - \tilde{A}_{ii}\right)^{-2} \left(\tilde{Y}_i - \widehat{\tilde{Y}}_i\right),
$$

where $\tilde{A}_{ii}$ is the diagonal block of smoothing matrix $\tilde{A}_{\lambda,\boldsymbol{\theta}}$ corresponding to the observations on subject i, and I_{p_i} is a $p_i \times p_i$ identity matrix.

A detailed presentation and proof can be found in Xu et al. (2012) and supplementary materials Xu and Huang (2012). The authors additionally proposed an approximation to the LosoCV score to further reduce the computational cost of evaluating V_{loso}, which can be expensive due to the

inversion of the $I_{p_i} - \tilde{A}_{ii}$. Using the Taylor expansion of $\left(I_{p_i} - \tilde{A}_{ii}\right)^{-1} \approx I_{p_i} + \tilde{A}_{ii}$, we can use the following to approximate V_{loso}:

$$V^*_{loso}(\lambda) = \frac{1}{N}|| \left(I - \tilde{A}_{\lambda,\boldsymbol{\theta}}\right) \tilde{Y}||^2 + \frac{2}{N} \sum_{i=1}^{N} \hat{\tilde{e}}_i' \tilde{A}_{ii} \hat{\tilde{e}}_i, \tag{3.41}$$

where $\hat{\tilde{e}}_i$ is the portion of the vector of prediction errors $\left(I - \tilde{A}_{\lambda,\boldsymbol{\theta}}\right) \tilde{Y}$ corresponding to subject i. They show that under mild conditions, and for fixed, nonrandom λ, the approximate LosoCV score V^*_{loso} and the true LosoCV score V_{loso} are asymptotically equivalent. See Theorem 3.1 of Xu et al. (2012).

Selection of Multiple Smoothing Parameters

With the definition of the unbiased risk estimate and the leave-one-subject-out criteria, the expression of the smoothing matrix in (3.35) permits straightforward evaluation of both scores $U(\lambda, \boldsymbol{\theta})$ and $V^*_{loso}(\lambda, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$ denotes the vector of smoothing parameters associated with each RK. In this section, we discuss an algorithm to minimize the unbiased risk estimate $U(\lambda, \boldsymbol{\theta})$ with respect to λ and $\boldsymbol{\theta}$ hidden in $K = \sum_\beta \theta_\beta K_\beta$, where the (i, j) entry of K_β is given by $K_\beta(\boldsymbol{v}_i, \boldsymbol{v}_j)$, for $\boldsymbol{v}_i, \boldsymbol{v}_j \in V$. We present minimization of the unbiased risk estimate explicitly, but the mechanics of the optimization are very similar to those necessary for optimizing the leave-one-subject-out cross validation criterion. The details of a procedure for explicitly minimizing the alternative criterion are presented in Xu et al. (2012), which is based on the algorithms of Gu and Wahba (1991), Kim and Gu (2004) and Wood (2004). The algorithm in Kim and Gu (2004) is the basis for the following algorithm. The key difference between the minimization of U and the minimization of V^*_{loso} lies in the calculation of the gradient and the Hessian matrix in the Newton update. To minimize the unbiased risk estimate,

I. Fix $\boldsymbol{\theta}$; minimize $U(\lambda|\boldsymbol{\theta})$ with respect to λ.

II. Update $\boldsymbol{\theta}$ using the current estimate of λ.

Executing I follows immediately from the expression for the smoothing matrix. Performing the update in II requires approximating the gradient and the Hessian of $U(\boldsymbol{\theta}|\lambda)$ with respect to $\boldsymbol{\kappa} = \log(\boldsymbol{\theta})$. Optimizing with respect to $\boldsymbol{\kappa}$ rather than on the original scale is motivated by two driving factors. First, $\boldsymbol{\kappa}$ is invariant to scale transformations because the derivatives of $U(\cdot)$ with respect to $\boldsymbol{\kappa}$ are invariant to such transformations, while the derivatives with respect to $\boldsymbol{\theta}$ are not. The second motivation for optimizing with respect to $\boldsymbol{\kappa}$ is that it converts a constrained optimization ($\theta_\beta \geq 0$) problem to an unconstrained one.

Algorithms

The following presents the main algorithm for minimizing $U(\lambda, \boldsymbol{\theta})$ and its key components are presented in the section to follow. The minimization of U is done via two nested loops. Fixing tuning parameter λ, the outer loop minimizes U with respect to smoothing parameters θ_β via quasi-Newton iteration of Dennis Jr and Schnabel (1996), as implemented in the `nlm` function in `R`. The inner loop then minimizes $-2\ell + \lambda J(\phi)$ with fixed tuning parameters via Newton iteration. Fixing the θ_βs in $J(\phi) = \sum_\beta \theta_\beta^{-1} J_\beta(\phi_\beta)$, the outer loop with a single λ is straightforward.

Algorithm 1 Selection of multiple smoothing parameters for the SSANOVA model.

1: **Initialization:**
2: Set $\Delta\boldsymbol{\kappa} := 0;\ \boldsymbol{\kappa}_{-} := \boldsymbol{\kappa}_0;\ U_{-} = \infty;$
3: **Iteration:**
4: **while** not converged **do**
5: For current value $\boldsymbol{\kappa}^* = \boldsymbol{\kappa}_{-} + \Delta\boldsymbol{\kappa}$, compute $K_\theta^* = \sum_{\beta=1}^{g} \theta_\beta^* K_\beta$, scale so that tr (K_β) is fixed.
6: Compute $\tilde{A}(\lambda|\boldsymbol{\theta}^*) = \tilde{A}(\lambda, \exp(\boldsymbol{\kappa}^*))$.
7: Minimize $U(\lambda|\boldsymbol{\kappa}^*) = \tilde{Y}' \left(I - \tilde{A}_{\lambda,\boldsymbol{\theta}}\right)^2 \tilde{Y} + 2\ \mathrm{tr}\tilde{A}_{\lambda,\boldsymbol{\theta}}$
8: Set $U_* := \min_{\lambda} U(\lambda|\boldsymbol{\kappa}^*)$
9: **if** $U^* > U_{-}$ **then**
10: Set $\Delta\boldsymbol{\kappa} := \Delta\boldsymbol{\kappa}/2$
11: Go to 5.
12: **else**
13: Continue
14: **end if**
15: Evaluate the approximation of gradient $\mathbf{g} = (\partial/\partial\boldsymbol{\kappa})\, U(\boldsymbol{\kappa}|\lambda)$
16: Evaluate the approximation of Hessian $H = (\partial^2/\partial\boldsymbol{\kappa}\partial\boldsymbol{\kappa}')\, U(\boldsymbol{\kappa}|\lambda)$.
17: Calculate step $\Delta\boldsymbol{\kappa}$:
18: **if** H positive definite **then**
19: $\Delta\boldsymbol{\kappa} := -H^{-1}\mathbf{g}$
20: **else**
21: $\Delta\boldsymbol{\kappa} := -\tilde{H}^{-1}\mathbf{g}$, where $\tilde{H} = \mathrm{diag}\,(e)$ is positive definite.
22: **end if**
23: **end while**
24: **Calculate optimal model:**
25: **if** $\Delta\kappa_\beta < -\gamma$, for γ large **then**
26: Set $\kappa_\beta^* := -\infty$
27: **end if**
28: Compute $K_\theta^* = \sum_{\beta=1}^{g} \theta_\beta^* K_\beta$;
29: Calculate $\begin{bmatrix} d \\ c \end{bmatrix} = \tilde{C}^{-1} \left(\tilde{C}'\right)^{-1} \begin{bmatrix} \tilde{B}' \\ \tilde{K}_\theta^{*\prime} \end{bmatrix} \tilde{Y}$ as in (3.33)

Algorithm step (21) returns a descent direction even when H is not positive definite by adding positive mass e to the diagonal elements of H if necessary to produce $\tilde{H} = G'G$ where G is upper triangular. See Chapter 4 in Gill et al. (1981) for details. Gu and Wahba (1991) present details

on convergence criteria based on those suggested in Gill et al. (1981). Gill et al. (1981) provide detailed discussion of the Newton method based on the Cholesky decomposition necessary for calculating the update direction for $\boldsymbol{\kappa}$.

The unbiased risk estimate $U(\lambda, \boldsymbol{\theta})$ is fully parameterized by $\{\lambda_\beta\} = \left\{\lambda\theta_\beta^{-1}\right\}$, so the smoothing parameters $\left(\lambda, \left\{\theta_\beta^{-1}\right\}\right)$ over-parameterize the score, which is the reason for scaling the trace of K_β. The starting values for the θ quasi-Newton iteration are obtained with two passes of the fixed-θ outer loop as follows:

I. Set $\breve{\theta}_\beta^{-1} \propto \text{tr}\,(K_\beta)$, and minimize $U(\lambda)$ with respect to λ to obtain $\breve{\phi}$.

II. Set $\bar{\theta}_\beta^{-1} \propto J_\beta\left(\breve{\phi}_\beta\right)$, and minimize $U(\lambda)$ with respect to λ to obtain $\bar{\phi}$.

The first pass allows equal opportunity for each penalty to contribute to U, allowing for arbitrary scaling of $J_\beta(\phi_\beta)$. The second pass grants greater allowance to terms exhibiting strength in the first pass. The following θ iteration fixes λ and starts from $\breve{\theta}_\beta$. These are the starting values adopted by Gu and Wahba (1991); the starting values for the first pass loop are arbitrary, but are invariant to scalings of the θ_β. The starting values in II for the second pass of the outer are based on more involved assumptions derived from the background formulation of the smoothing problem. After the first pass, the initial fit $\breve{\phi}$ reveals where the structure in the true ϕ lies in terms of the components of the subspaces $\mathcal{H}_\beta$. Less penalty should be applied to terms exhibiting strong signal.

3.3 A Reproducing Kernel Hilbert Space Framework for the Innovation Variance Function

Fixing ϕ in (3.2), the negative log likelihood of the data $Y_1, \ldots, Y_N$ satisfies

$$-2\ell\left(\sigma^2 | Y_1, \ldots, Y_N, \phi\right) = \sum_{i=1}^{N}\sum_{j=1}^{p_i} \log\sigma_{ij}^2 + \sum_{i=1}^{N}\sum_{j=1}^{p_i} \frac{\epsilon_{ij}^2}{\sigma_{ij}^2}; \tag{3.42}$$

where $\epsilon_{ij} = y_{ij} - \sum_{k<j} \phi\left(\boldsymbol{v}_{ijk}\right) y_{ik}$. Let

$$\text{RSS}\left(t\right) = \sum_{i,j:t_{ij}=t} \left(y_{ij} - \sum_{k<j} \phi\left(\boldsymbol{v}_{ijk}\right) y_{ik} \right)^2 \tag{3.43}$$

denote the squared innovations for the observations y_{ij} having corresponding measurement time $t = t_{ij}$. Then $\text{RSS}\left(t\right)/\sigma^2\left(t\right) \sim \chi^2_{df_t}$, where the degrees of freedom df_t corresponds to the number of observations y_{ij} having corresponding measurement time t. In this light, for fixed ϕ, the penalized likelihood (3.42) is that of a variance model with the ϵ^2_{ij} serving as the response. This corresponds to a generalized linear model with Gamma errors and known scale parameter equal to 2. Let $z_{ij} = \epsilon^2_{ij}$, and let $Z_i = \left(z_{i1}, \ldots, z_{i,p_i}\right)'$ denote the vector of squared innovations for the i^{th} observed trajectory.

The Gamma distribution is parameterized by shape parameter α and scale parameter β. Let σ^2 denote the mean of the distribution given by $\alpha\beta$. For a single observation Z, reparameterizing the Gamma likelihood in terms of (α, σ^2) and dropping terms that don't involve $\sigma^2\left(\cdot\right)$ gives

$$\begin{aligned} -\ell\left(\sigma^2, \alpha | z\right) &\propto \alpha \left[\log \sigma^2 + \frac{z}{\sigma^2} \right] \\ &= \alpha \left[\eta + z e^{-\eta} \right], \end{aligned} \tag{3.44}$$

where α^{-1} is the dispersion parameter and $\eta = \log \sigma^2$. The log likelihood of the squared working residuals $Z_1, \ldots, Z_N$ becomes

$$-2\ell\left(\sigma^2 | Z_1, \ldots, Z_N\right) = \sum_{i=1}^{N} \sum_{j=1}^{p_i} \eta_{ij} + \sum_{i=1}^{N} \sum_{j=1}^{p_i} z_{ij} e^{-\eta_{ij}}, \tag{3.45}$$

where $\eta_{ij} = \eta\left(t_{ij}\right)$. This coincides with a Gamma distribution with scale parameter $\alpha = 2$. Smoothing spline ANOVA models for exponential families have been studied extensively; see Wahba et al. (1995), Wang (1997), and Gu (2013). Fixing ϕ, we take the estimator of $\eta\left(t\right) = \log \sigma^2\left(t\right)$ to be the minimizer of the penalized negative log likelihood:

$$-2\ell\left(\eta | Z_1, \ldots, Z_N\right) + \lambda J\left(\eta\right) = \sum_{i=1}^{N} \sum_{j=1}^{p_i} \eta\left(t_{ij}\right) + \sum_{i=1}^{N} \sum_{j=1}^{p_i} z_{ij} e^{-\eta(t_{ij})} + \lambda J\left(\eta\right), \tag{3.46}$$

for $\eta \in \mathcal{H}$, where the penalty J can be written as a square norm and decomposed as in (3.16), with

$$J(\eta) = \sum_{\beta} \theta_{\beta}^{-1} \langle \eta, \eta \rangle_{\mathcal{H}_\beta}.$$

The first two terms in (3.45) serve as a measure of the goodness of fit of η to the data, and only depend on η through the evaluation functional $[t_{ij}]\,\eta$. Thus, the argument justifying the form of the minimizer in (3.19) applies to η. Let $\mathcal{T} = \bigcup_{i,j} \{t_{ij}\}$ denote the unique values of the observations times pooled across subjects. The minimizer of the penalized likelihood (3.46) has the form

$$\eta_\lambda(t) = \sum_{i=1}^{\mathcal{N}_0} d_i \nu_i(t) + \sum_{j=1}^{|\mathcal{T}|} c_j K_1(t_j, t), \tag{3.47}$$

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 $\mathcal{T}$, viewed as a function of t.

3.3.1 Model Fitting

The Gamma penalized negative log likelihood (3.47) is non-quadratic, so η_λ must be computed using iteration even for fixed smoothing parameters. Standard theory for exponential families gives us that the functional

$$L(\eta) = \sum_{i=1}^{N} \sum_{j=1}^{p_i} \eta(t_{ij}) + \sum_{i=1}^{N} \sum_{j=1}^{p_i} z_{ij} e^{-\eta(t_{ij})} \tag{3.48}$$

is continuous and convex in $\eta \in \mathcal{H}$. We assume that the $|V| \times \mathcal{N}_0$ matrix B which has (i,j) element $\nu_j(t_i)$ is full column rank, so that $L(\eta)$ is strictly convex in $\mathcal{H}$ and the minimizer of (3.46) uniquely exists. See Wahba et al. (1995).

The minimizer can be computed via Newton iteration using a quadratic approximation of (3.48) at a point $\tilde{\eta}$. Letting $\tilde{u}_{ij} = -z_{ij} e^{-\tilde{\eta}_{ij}}$, the Newton iteration uses the minimizer of the penalized weighted sums of squares

$$\sum_{i=1}^{N} \sum_{j=1}^{p_i} \left(\tilde{z}_{ij} - \eta(t_{ij})\right)^2 + \lambda J(\eta) \tag{3.49}$$

to update $\tilde{\eta}$, where $\tilde{z}_{ij} = \tilde{\eta}(t_{ij}) - \tilde{u}_{ij}$.

3.3.2 Smoothing Parameter Selection for Exponential Families

Performance-oriented is a typical choice for method of smoothing parameter selection when data are generated from a distribution belonging to exponential families. This section provides a brief overview of the the performance-oriented iteration, specifically for selecting the optimal degree of smoothing for $\sigma^2(t)$. This approach is just one of many in the inventory of model selection techniques for penalized regression with exponential families. We refer the reader desiring detailed examination to Zhang and Lin (2006), Xiang and Wahba (1996), Wahba et al. (1995), Wood (2004), and Wood (2017).

A measure of the discrepancy between distributions belonging to an exponential family having densities of the form $p(z) = \exp\{(z\eta - b(\eta))/a(\phi) + c(z,\phi)\}$ is the Kullback-Leibler distance

$$\begin{aligned} \mathrm{KL}(\eta, \eta_\lambda) &= E_\lambda\left[Z(\eta - \eta_\lambda) - (b(\eta) - b(\eta_\lambda))\right]/a(\phi) \\ &= \left[b'(\eta)(\eta - \eta_\lambda) - (b(\eta) - b(\eta_\lambda))\right]/a(\phi). \end{aligned} \tag{3.50}$$

For the Gamma distribution, the KL distance simplifies to

$$\mathrm{KL}(\eta, \eta_\lambda) = -\sigma^2\left(e^{-\eta} - e^{-\eta_\lambda}\right) - (\eta - \eta_\lambda),$$

which is not symmetric. Thus, a natural choice of loss function for measuring the performance of an estimator $\eta_\lambda(t)$ of $\eta(t)$ is the symmeterized Kullback-Leibler distance averaged over the observed time points $t_{11}, \ldots, t_{N,p_N}$. For the Gamma distribution, this is given by

$$L(\eta, \eta_\lambda) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{p_i}\sum_{j=1}^{p_i}\left(\frac{\sigma^2(t_{ij})}{\sigma^2_\lambda(t_{ij})} - \frac{\sigma^2_\lambda(t_{ij})}{\sigma^2(t_{ij})} - 2\right)\right]. \tag{3.51}$$

The ideal smoothing parameters are those which minimize (3.51). One can derive an unbiased risk estimate U for the Gamma distribution as in Section 3.2.3 for the Gaussian case. Theorem 5.2 in Gu (2013) gives that the minimizer of U, which relies only on the data, approximately minimizes and quadratic approximation to (3.51). To find the optimal value of the smoothing parameter, the

performance-oriented iteration tracks loss $L(\eta, \eta_\lambda)$ (through U) indirectly, simultaneously updating λ, θ_β. Since it does not explicitly keep track of $L(\eta, \eta_\lambda)$ itself, it may not be the most effective way to search for the optimal smoothing parameters, but it is numerically efficient. Instead of fixing smoothing parameters and moving according to a particular Newton update, one chooses an update from among a family of Newton updates that is perceived to be better performing according to $L(\eta, \eta_\lambda)$. If the smoothing parameters stabilize at, say, $\left(\lambda^*, \theta_\beta^*\right)$ and the corresponding Newton iteration converges at η^*, then it is clear that $\eta^* = \eta_{\lambda^*}$ is the minimizer. In a neighborhood of η^* where the corresponding values of the quadratic approximation of L closely approximate the penalized likelihood functional (3.48) for smoothing parameters close to $\left(\lambda^*, \theta_\beta^*\right)$, then the η_{λ,η^*}s are, in turn, hopefully close approximations to the η_λs. Thus, through indirect comparison η^* is perceived to be better performing among the other η_λs in the neighborhood. See Chapter 5, Section 2 of Gu (2013) for thorough discussion.

An alternative to the performance-oriented iteration is to choose the optimal smoothing parameters by comparing candidate η_λs directly. The generalized approximate cross validation (GACV) score in Xiang and Wahba (1996) keeps track of $L(\eta, \eta_\lambda)$, approximating the score which is analogous to the generalized cross validation score (GCV) in the usual penalized regression setting (Wahba, 1990). We refer the reader to the aforementioned sources for extensive discussion. For the same reason that we utilized the LosoCV criterion rather than leave-one-out or generalized cross validation for smoothing parameter selection when estimating ϕ, we did not explore using GACV for model selection for the innovation variance function.

To jointly estimate the autoregressive coefficient function and the innovation variance function, we adopt an iterative approach in the spirit of Huang et al. (2006), Huang et al. (2007), and

Pourahmadi (2000). A procedure for minimizing

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

starts with initializing $e^{\eta_{ij}}=\sigma_{ij}^2=1$ for $i=1,\ldots,N$, $j=1,\ldots,p_i$. For fixed η, we take ϕ^* to minimize the penalized negative log likelihood

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

Given ϕ^* and setting $\phi=\phi^*$, we update our estimate of η by taking η^* to minimize the penalized negative log likelihood of the working residuals

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

This process of iteratively updating ϕ^* and η^* is repeated until convergence. Alternatively, one could initialize the innovation variances using parsimonious approximation of the autoregressive varying coefficient ϕ to calculate an initial estimate of the σ_{ij}^2. For example, one might initialize the innovation variances using the squared prediction errors after regressing y_{ij} on just its most recent preceding measurement $y_{i,j-1}$. Better starting values may lead to faster convergence of the overall procedure, though we have not yet explored this concern and leave it to future investigation.

Chapter 4: A P-spline Model for the Cholesky Decomposition

In this chapter, we demonstrate multidimensional smoothing with penalized B-splines, or *P-splines*, as a flexible and computationally convenient alternative to the Hilbert space methods presented in Chapter 3. P-spline models are an extension of (generalized) linear regression models. They exploit the attractive properties of the B-spline basis along with the use of computationally convenient difference penalties. The formulation of the penalty is independent of the basis, which provides added modeling flexibility due to the ease with which one can employ various types of regularization. The B-spline functions have compact support, making them more attractive than the smoothing spline basis when the function to be estimated exhibits compact support as well, such as covariance matrices having banded Cholesky factor. Despite their flexibility, fitting P-spline models are only as computationally intensive as fitting regression models.

4.1 Tensor Product B-splines for Multidimensional Smoothing

Splines are piecewise polynomial functions, where the piecewise polynomials are joined at certain values of the domain called knots. B-splines are a basis for splines. Given a set of knots, B-splines can be easily computed recursively for any polynomial degree; see De Boor et al. (1978) and Dierckx (1995). The smoothness of a fitted curve can be controlled by the number of B-splines used in the basis expansion used to approximate the curve. Fewer knots (thus, fewer basis functions) lead to smoother fits, and there is an extensive body of research focused on the choice

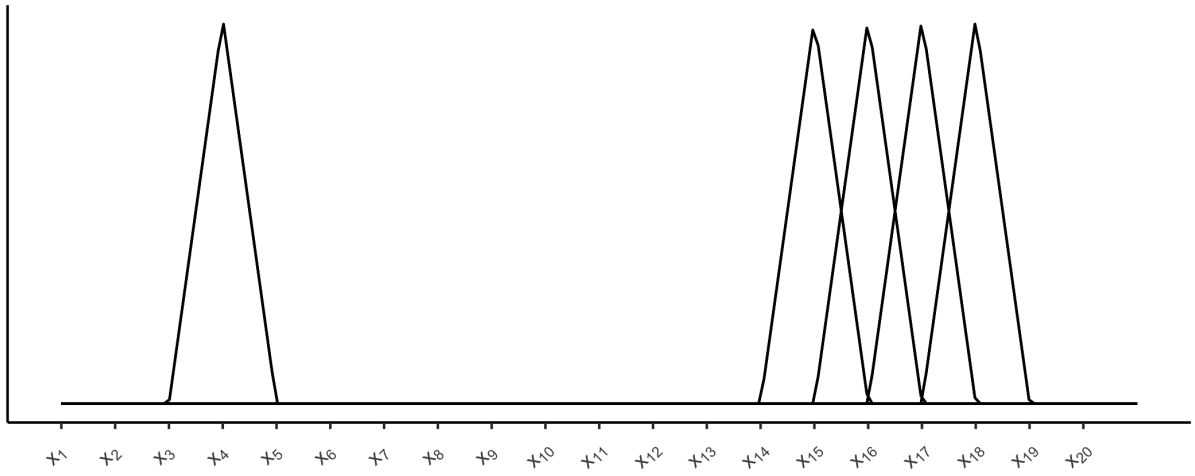
of knot placement. Some authors have proposed adaptive smoothing techniques which attempt to automatically optimize the number and the positions of the knots; see Friedman and Silverman (1989), Kooperberg and Stone (1991). However, this problem is nontrivial and requires nonlinear optimization, and is still an open problem today. However, limiting the number of B-splines is not the only approach to controlling the complexity of the fitted function.

Instead, Eilers and Marx (1996) propose alternative an approach to nonparametric smoothing based on finite difference penalties. The difference penalties are trivial to compute and can be done so independently of the basis, unlike the smoothing spline penalty functional (3.6). Their approach circumvents the choice of knot specification. They achieve smoothness in fitted functions by purposefully overfitting the smooth coefficient vectors using a B-spline basis with a large number of equally spaced knots. Augmenting the log likelihood with the difference penalty prevents overfitting and accommodates a potentially ill-conditioned fitting procedure.

Analogous to the smoothing spline representation (3.19), we can represent ϕ using a B-spline basis. But first, in order to illustrate the ideas in the sections to follow, it is pragmatic to first review some basic properties of B-splines. For an exhaustive and more formal mathematical review, see De Boor et al. (1978) and Dierckx (1995). A B-spline is a function constructed from piecewise polynomial functions which are connected in a very particular way. Their values can be computed recursively; for a non-decreasing sequence of knots $\{t_i\}$, the value of the i^{th} B-spline of order k can be defined using

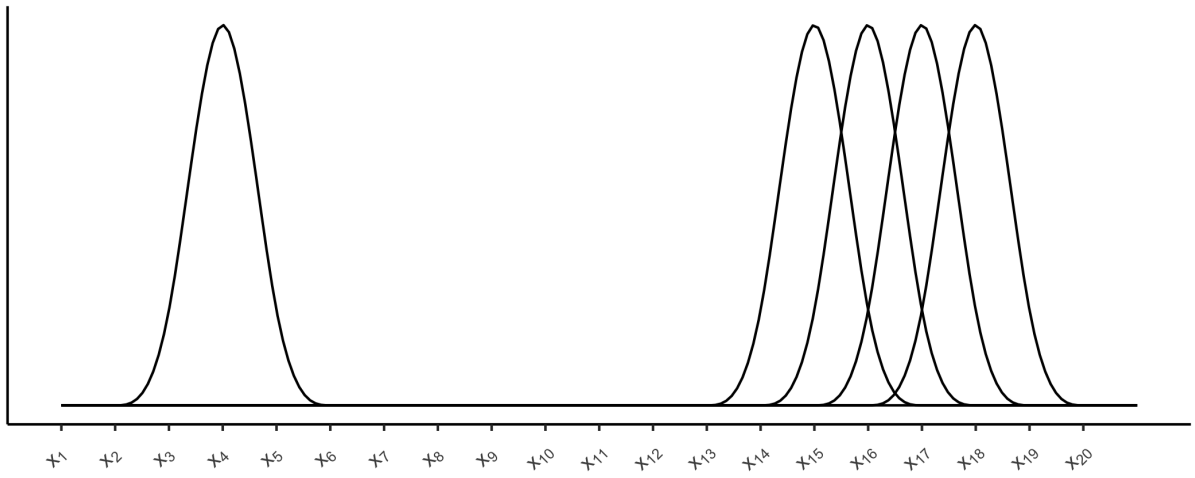
$$\begin{aligned} B_{i1}(x) &= \begin{cases} 1, & t_i \leq x < t_{i+1} \\ 0, & otherwise \end{cases} \\ B_{ik}(x) &= \frac{x - t_i}{t_{i+k-1} - t_i} B_{i,k-1}(x) + \frac{t_{i+k} - x}{t_{i+k} - t_{i+1}} B_{i+1,k-1}(x). \end{aligned} \tag{4.1}$$

Figure 4.1 displays a set of linear B-splines, and Figure 4.1 displays a set of B-splines of degree 2. A single isolated B-spline is shown on the left side of the axis in each panel. In Figure 4.1, the single B-spline of degree 1 consists of two linear pieces: one piece from x_3 to x_4, and the other from x_4 to x_5, which are the knots that define its support. In the right part of Figure 4.1, three more B-splines of degree 1 are shown. Each one based on three knots. Comparing these with the overlapping quadratic B-splines in Figure 4.2, we can see that the extent to which neighboring B-splines overlap depends on the polynomial degree of the basis.



(a)

Figure 4.1: *B-splines of degree 1*



(a)

Figure 4.2: *B-splines of degree 2*

B-splines make attractive basis functions for nonparametric regression; a linear combination of B-spline basis functions gives a smooth curve. Once a B-spline basis is computed, their application is no more difficult than polynomial regression, and extension to two-dimensional smoothing is

available with the use of tensor products. To construct a B-spline representation for ϕ, we need to equip the l and m axes each with a B-spline basis: let

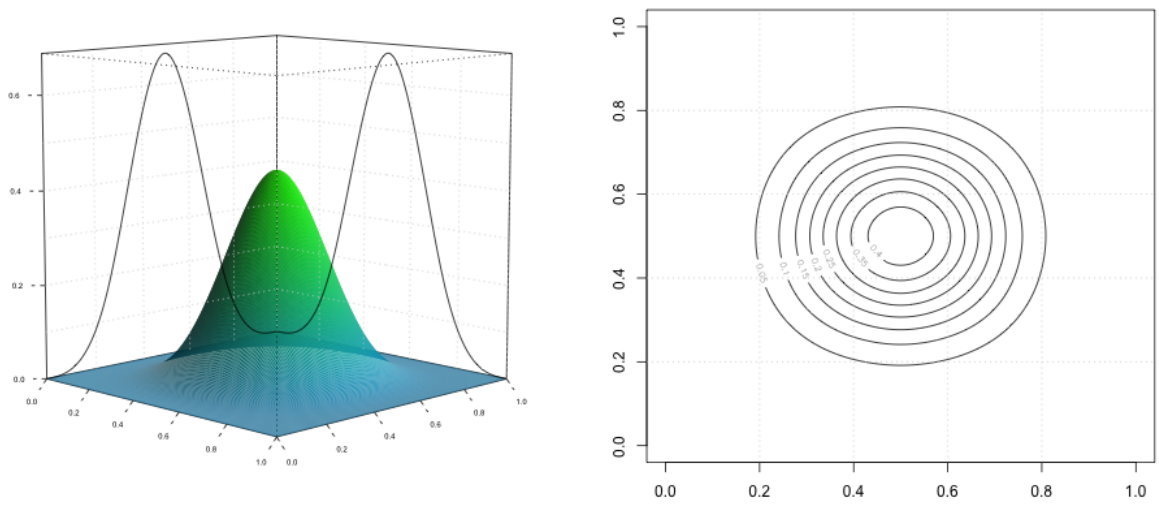
$$B_{l_1}(l), \ldots, B_{l_{k_l}}(l) \text{ and } B_{m_1}(m), \ldots, B_{m_{k_m}}(m)$$

denote the B-spline bases for l and m, each having a set of equally spaced knots along their respective domain. It is worth noting that one is free to specify a different basis for each dimension either by using different order B-spline or using different numbers of knots. Order of the basis will be indicated only when necessary and otherwise suppressed to maintain simplicity of notation. The tensor product basis functions

$$T_{kk'}(l, m) = B_{l_k}(l) B_{m_{k'}}(m)$$

carve the l-m domain into rectangles. Figure 4.3 shows a single $T_{kk'}$, where the marginal B-spline bases are of degree 2. For a given knot grid, we can approximate ϕ by

$$\phi(l, m) = \sum_{r=1}^{k_l} \sum_{c=1}^{k_m} \theta_{rc} B_{l_r}(l) B_{m_c}(m). \tag{4.2}$$



B-splines of degree 1 *B-splines of degree 2*

Figure 4.3: *Tensor product of two quadratic B-splines*

4.2 Difference Penalties

The specification of a P-spline model provides a simple way to avoid the issue of optimal knot selection for the l and m bases. This is done by constructing the marginal B-spline bases with a large number of knots - more than necessary. Application of a smoothness penalty prevents overfitting. A penalty based on discrete differences of the B-spline coefficients controls the fit in a way much like the classical second derivative penalty (3.6) does. However, its construction is trivial even when the number of basis functions is very large. Using the properties of B-splines, it is straightforward to show that the difference penalty of order d approximates the integrated square of the d^{th} derivative well, so little is lost by using it in place of the derivative-based penalty. O'Sullivan (1986) established that for $f(x) = \sum_{j=1}^{k} \theta_j B_j(x)$, one can derive a banded matrix P

using the properties of B-splines such that $J(f) = \int_0^1 (f''(x))^2$ can be written

$$J(f) = \theta' P \theta$$

where $\theta = (\theta_1, \ldots, \theta_k)$ denotes the vector of B-spline basis coefficients. The (i, j) element of the penalty matrix P is given by

$$p_{ij} = \int_0^1 B_i''(x) B_j''(x).$$

Wand and Ormerod (2008) extend the work in O'Sullivan (1986) to higher order derivatives for general degree B-splines and derive an exact matrix algebraic expression for the penalty matrices. The computation of P is nontrivial and becomes very tedious when the third and fourth derivative are used as the roughness measure. In the cubic case, the expression is a result of the application of Simpson's Rule applied to the inter-knot differences since each $B_i'' B_j''$ is a piecewise quadratic function. The penalty may be written

$$P = (B'')' \operatorname{diag}(\omega) B'',$$

where B'' is the $3(k+7) \times (k+4)$ matrix with i-j^{th} entry given by $B_j''(x_i^*)$, x_i^* is the i^{th} element of

$$\left(\theta_1, \frac{\theta_1 + \theta_2}{2}, \theta_2, \theta_2, \frac{\theta_2 + \theta_3}{2}, \theta_3, \ldots, \theta_{k+7}, \frac{\theta_{k+7} + \theta_{k+8}}{2}, \theta_{k+8}\right),$$

and ω is the $3(k+7) \times 1$ vector given by

$$\omega = \left(\frac{1}{6}(\Delta\theta)_1, \frac{4}{6}(\Delta\theta)_1, \frac{1}{6}(\Delta\theta)_1, \frac{1}{6}(\Delta\theta)_2, \frac{4}{6}(\Delta\theta)_2, \right.$$
$$\left. \frac{1}{6}(\Delta\theta)_2, \ldots, \frac{1}{6}(\Delta\theta)_{n+7}, \frac{4}{6}(\Delta\theta)_{k+7}, \frac{1}{6}(\Delta\theta)_{k+7}\right)$$

where $(\Delta\theta)_j = \theta_{j+1} - \theta_j$. They generalize this to the case of any order penalty and present a table of formulas for constructing any arbitrary penalty matrix, P.

Alternatively, Eilers and Marx (1996) replace the curvature penalty (3.6) with a finite difference penalty on the B-spline coefficients. They suggest enforcing smoothness of fitted functions $f(x) = \sum_{j=1}^{k} \theta_i B_j(x)$ using the d^{th} order difference penalty:

$$J_d(f) = \sum_{j=d}^{k} \left(\Delta^d \theta_j\right)^2. \tag{4.3}$$

where $\Delta\theta_j = \theta_j - \theta_{j-1}$, and $\Delta^2\theta_j = \Delta(\Delta\theta_j) = \theta_j - 2\theta_{j-1} + \theta_{j-2}$. In general, $\Delta^d\theta_j = \Delta\left(\Delta^{d-1}\theta_j\right)$. Let D_d denote the differencing operator:

$$D_d\theta = \Delta^d\theta.$$

Then, (4.3) can be written in terms of the squared norm of the difference operator applied to the vector of B-spline coefficients:

$$\begin{aligned} J_d(f) &= ||D_d\theta||^2 \\ &= \theta' P_d \theta \end{aligned} \tag{4.4}$$

where $P_d = D_d' D_d$. The connection between the second-derivative penalty to the penalty on second-order differences of the B-spline coefficients can be established with straightforward calculus and the recursive property of the B-spline basis functions. The derivative properties of B-splines permits the traditional smoothness penalty applied to f to be written

$$\int_0^1 \left(f''(x)\right)^2 \, dx = \int_0^1 \left[\sum_{i=1}^{k}\sum_{j=1}^{k} \Delta^2\theta_i \Delta^2\theta_j B_{i,1}(x) B_{j,1}(x) \, dx\right].$$

where $B_{j,1}(x)$ is the j^{th} B-spline of order 1. Most of the cross products of $B_{i,1}$ and $B_{j,1}$ vanish since B-splines of degree 1 only overlap when j is $i-1$, i, or $i+1$. Thus, we have that

$$\begin{aligned} \int_0^1 \left(f''(x)\right)^2 \, dx &= \int_0^1 \left[\left(\sum_j \Delta^2\theta_j B_{j,1}(x)\right)^2 + 2\sum_j \Delta^2\theta_j \Delta^2\theta_{j-1} B_{j,1} B_{j-1,1}(x)\right] dx \\ &= \sum_j \left(\Delta^2\theta_j\right)^2 \int_0^1 B_{j,1}^2(x) \, + 2\sum_j \Delta^2\theta_j \Delta^2\theta_{j-1,1} \int_0^1 B_{j,1}(x) B_{j-1,1}(x) \, dx. \end{aligned} \tag{4.5}$$

This can be written as

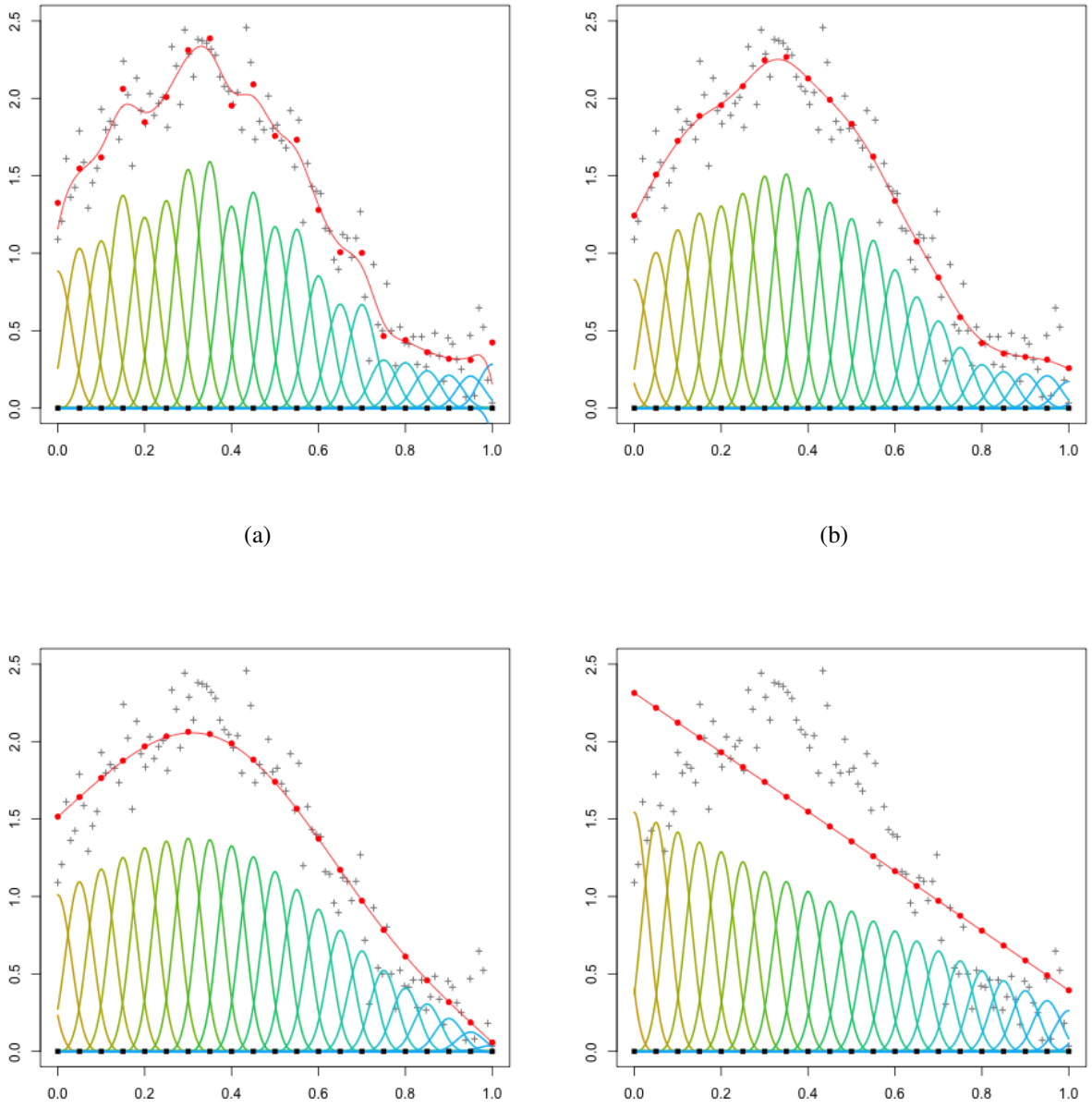
$$\int_0^1 \left(f''(x)\right)^2 = c_1 \sum_j \left(\Delta^2 \theta_j\right)^2 + c_2 \sum_j \Delta^2 \theta_j \Delta^2 \theta_{j-1}. \tag{4.6}$$

Given a set of equidistant knots, the constants c_1 and c_2 are given by

$$\begin{aligned} c_1 &= \int_0^1 \left(B_{j,1}(x)\right)^2 \, dx \text{ and} \\ c_2 &= \int_0^1 B_{j,1}(x) \, B_{j-1,1}(x) \, dx. \end{aligned} \tag{4.7}$$

This establishes that traditional smoothness penalty on the squared second derivative can be written as a linear combination of a penalty on the second-order differences of the B-spline coefficients (4.3) and the sum of the cross products of neighboring second differences. The second term in (4.6) leads to a complex objective function when minimizing the penalized likelihood, where seven adjacent spline coefficients occur, as opposed to five if only the first term in (4.6) is used in the penalty. The added complexity is a consequence of overlapping B-splines, which quickly increases when using higher order differences and higher order B-splines.

A smoother sequence of coefficients leads to a smoother curve, as illustrated in Figure 4.4. The relationship between P-spline curves and their coefficients is easily characterized if we consider the coefficients as the skeleton of the function, and draping the B-splines over them puts the flesh on the bones, so to speak. As long as the coefficient sequence is smooth, the number of basis functions (and coefficients) is unimportant since the penalty ensures the smoothness of the skeleton and that the fitting procedure is well-conditioned.

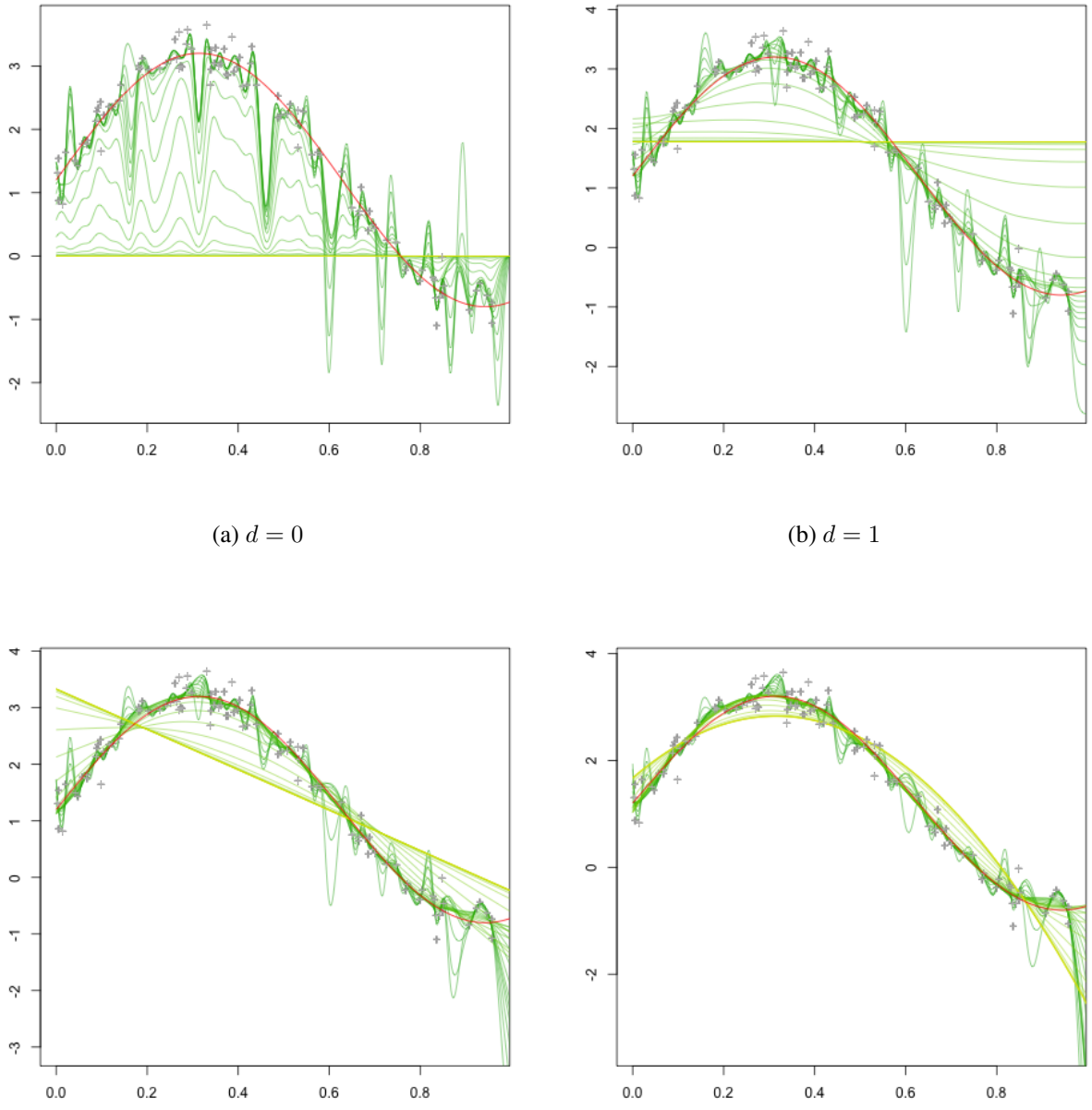


(a) (b)

(c) (d)

Figure 4.4: *Illustration of the impact of the second order difference penalty. The number of B-splines used is the same in each plot, with the value of the penalty parameter increasing from left to right and top to bottom across each plot. The red circles are the values of each of the B-spline coefficients; as the penalty increases, they form as smoother sequence as we move across the four plots, which results in a smoother fitted function. As the penalty parameter approaches infinity, the fit approaches a linear function as shown in the bottom right plot.*

The limiting P-spline fit approaches a polynomial as the smoothing parameter tends to infinity. Under a difference penalty of order d, the fitted function will approach a polynomial of degree $(d-1)$ for large values of the smoothing parameter as long as the degree of the B-splines is greater than or equal to k. Figure 4.5 demonstrates the impact of the order of the penalty on the fitted function as the smoothing parameter increases. To verify this mathematically, we need to use the relationship between the differenced coefficient sequence and the derivative of a B-spline. See Appendix B. Consider using the second-order difference penalty. When λ is large, the penalty dominates the penalized likelihood, so that the minimizer θ must be such that $\sum_j (\Delta^2 \theta_j)^2$ is close to zero. Consequently, each of the individual second differences must also be nearly zero, and thus the second derivative of the fitted function must be close to zero over the entire domain.



(a) $d = 0$

(b) $d = 1$

(c) $d = 2$

(d) $d = 3$

Figure 4.5: *Illustration of the impact of the order of the difference penalty. The number of B-splines used is the same in each plot, with the penalty parameter varying from across the same grid of values. The fitted curves in the upper left plot correspond to the difference penalty of order* 0*, where* $|D_0\theta|^2 = \sum_i \theta_i^2$*, analogous to ridge regression using the B-spline basis as regression covariates. The fitted curves approach polynomials of degree* $d - 1$ *as* $\lambda \to \infty$*.*

4.3 The P-spline Estimator of the Generalized Autoregressive Varying Coefficient

To extend the use of the difference penalty (4.3) to the bivariate setting, the only necessary modification to the one-dimensional differencing procedure is the addition of a second difference penalty so that there is a penalty for each variable, l and m. Let Θ denote the $k_l \times k_m$ matrix of basis coefficients $\{\theta_{rc}\}$. For given Θ, the fitted value $\phi(l, m)$ may be written

$$\sum_{r=1}^{k_l} \sum_{c=1}^{k_m} \theta_{rc} B_{l_r}(l) B_{m_c}(m).$$

Let $\lambda = (\lambda_l, \lambda_m)$ denote the tuple of smoothing parameters for the l and m dimensions, respectively. We take ϕ_λ to be the minimizer of

$$\begin{aligned} -2\ell\left(\phi|Y_1, \ldots, Y_N, \sigma^2\right) + J_\lambda(\phi) = \sum_{i=1}^{N} \sum_{j=2}^{p_i} \frac{1}{\sigma^2(t_{ij})} \left(y_{ij} - \sum_{k=1}^{j-1} \left(\sum_{r=1}^{k_l} \sum_{c=1}^{k_m} \theta_{rc} B_r(l_{ijk}) B_c(m_{ijk}) \right) y_{ik} \right)^2 \\ + \lambda_l \sum_{r=1}^{k_l} ||\theta_{r\cdot} D_{d_l}||^2 + \lambda_m \sum_{c=1}^{k_m} ||D_{d_m} \theta_{\cdot c}||^2, \end{aligned} \tag{4.8}$$

where $\theta_{r\cdot}$ and $\theta_{\cdot c}$ denote the r^{th} row and c^{th} column of Θ, respectively. The second term in (4.8) imposes a difference penalty of order d_l on the rows of the coefficient matrix, while the third term places a difference penalty of order d_m on the columns.

The penalized log likelihood is quadratic in $\theta = \left(\theta_{11}, \ldots, \theta_{k_l, k_m}\right)'$. Demonstration of computation is simple if we express the coefficient matrix Θ in "unfolded" notation so that we can write the mean of the stacked response vector Y as defined in (3.20) as in the usual multiple regression form

$$E[Y] = X \text{vec}\{\phi(\boldsymbol{v})\} = XB\theta,$$

where $\theta = \text{vec}(\Theta)$ denotes the vectorized coefficient matrix constructed by stacking the columns of Θ. The $|V| \times k_l k_m$ tensor product basis B is constructed from the tensor product of the marginal

B-spline bases defined in Eilers et al. (2006) as the *row-wise Kronecker product* of the individual bases:

$$B = B_m \square B_l = \left(B_m \otimes 1'_{k_m}\right) \odot \left(1'_{k_l} \otimes B_l\right). \tag{4.9}$$

The operator $\odot$ denotes the element-wise matrix product; 1_{k_l} (1_{k_m}) denotes the column vector of ones having length k_l (k_m.) The operations in (4.9) construct B such that the i^{th} row of $B_m \square B_l$ is the Kronecker product of the corresponding rows of B_m and B_l. We can compactly express the penalty in (4.8) by writing

$$\lambda_l ||P_l \theta||^2 + \lambda_m ||P_m \theta||^2$$

where $P_l = I_{k_m} \otimes D'_{d_l} D_{d_l}$ and $P_m = D'_{d_m} D_{d_m} \otimes I_{k_l}$. The $n_Y \times k_l k_m$ matrix X is defined as before (3.22). Then the log likelihood (4.8) can be written as

$$-2\ell\left(\phi | Y_1, \ldots, Y_N\right) + J_\lambda\left(\phi\right) = \left(Y - XB\theta\right)' D^{-1} \left(Y - XB\theta\right) + \lambda_l ||P_l \theta||^2 + \lambda_m ||P_m \theta||^2. \tag{4.10}$$

Taking derivatives and setting equal to zero gives normal equations:

$$\left[\left(XB\right)' D^{-1} XB + \lambda_l P_l + \lambda_m P_m\right] \theta = \left(XB\right)' D^{-1} Y. \tag{4.11}$$

The solution ϕ_λ is given by $\phi_\lambda\left(\boldsymbol{v}\right) = \sum_{r=1}^{k_l} \sum_{c=1}^{k_m} \hat{\theta}_{rc} B_{l_r}\left(l\right) B_{m_c}\left(m\right)$, where

$$\hat{\theta} = \left[\left(XB\right)' D^{-1} XB + \lambda_l P_l + \lambda_m P_m\right]^{-1} \left(XB\right)' D^{-1} Y. \tag{4.12}$$

We note that the size of the system of equations (4.11) which determine the basis coefficients remains fixed at $k_l k_m$, even as the number of observations increases. The grid of regression coefficients can be recovered by arranging the elements of $\hat{\theta}$ into a matrix of k_l columns having length k_m. The vector of fitted values is given by

$$\hat{Y} = AY = X\left[\left(XB\right)' D^{-1} XB + \lambda_l P_l + \lambda_m P_m\right]^{-1} \left(XB\right)' D^{-1} Y, \tag{4.13}$$

where $A = X\left[(XB)' D^{-1} XB + \lambda_l P_l + \lambda_m P_m\right](XB)' D^{-1}$ is the "smoothing" matrix, analogous to the smoothing matrix $\tilde{A}$ (3.35) for the smoothing spline estimator in Chapter 3. Its use in smoothing parameter selection and model tuning is similar to the reproducing kernel Hilbert space framework, which we will discuss in the next section.

It is important to note that the construction of the tensor product basis B and penalty matrix P requires special care in this setting, where the domain of $\phi(l, m)$ is restricted to the region satisfying $0 \leq s < t \leq 1$, which is shown in Figure 4.6.

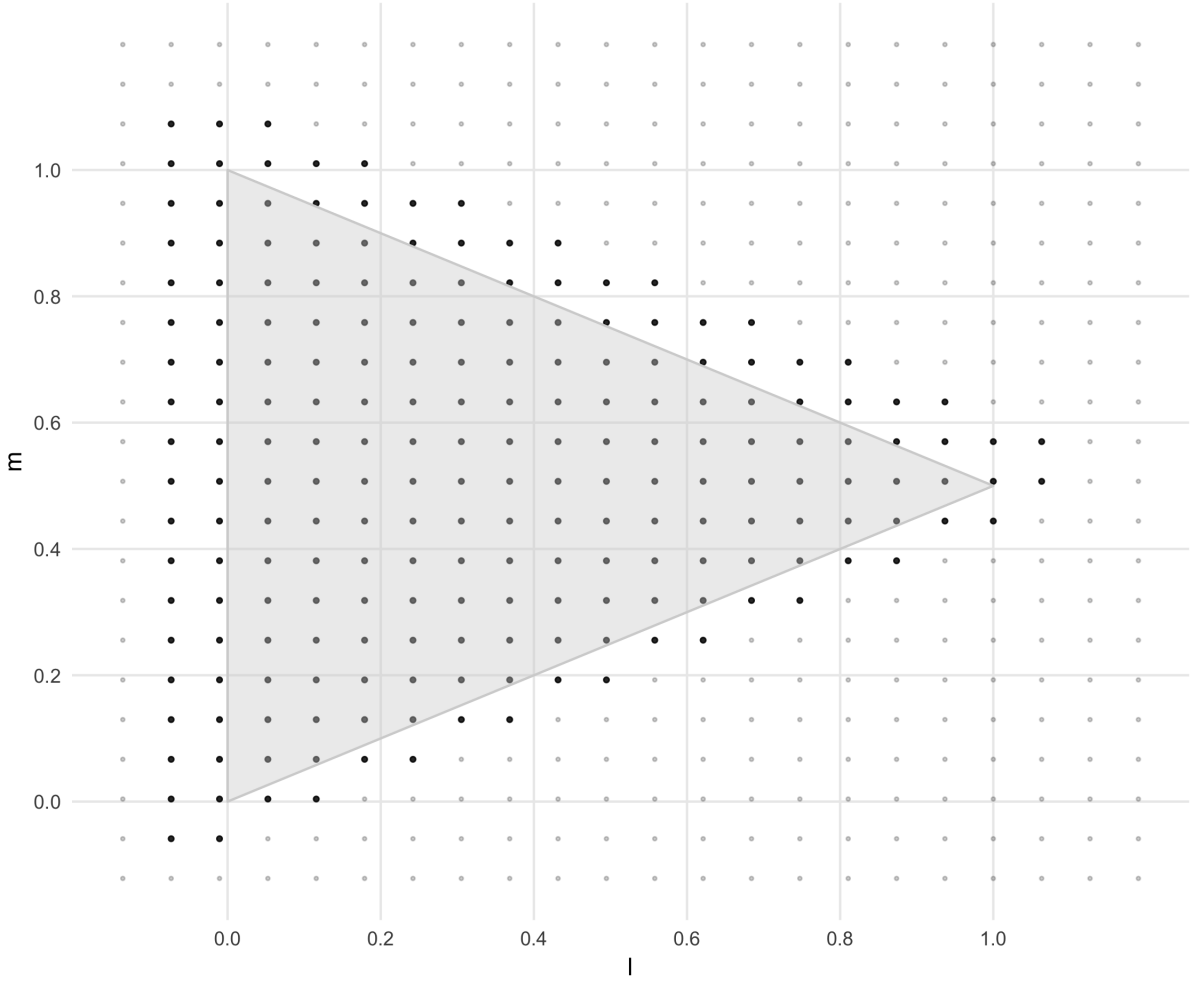


Figure 4.6: $\frac{l}{2} < m < 1 - \frac{l}{2}, \quad 0 < l < 1.$

When the tensor product basis is constructed on the regular grid defined by the cartesian product of the knots of the marginal bases B_l and B_m, a large number of basis functions anchored are at knots near which we have no data, so there is little information about the corresponding basis coefficient. As a result, the resulting tensor product matrix can be ill-conditioned and solving (4.11) results in singularities. In this case, the quality of the estimator can suffer terribly. To correct for this instability, one can simply remove the knots corresponding to tensor products functions which do not overlap with the function domain from the basis, B, and trimming the penalty matrices P_l and P_m as needed. With the trimmed basis and penalties, optimization can be carried out as previously discussed.

Bivariate B-splines are useful for smoothing over arbitrary domains, making them a natural alternative for the construction of a basis for $\boldsymbol{v} = (l, m)$. Multidimensional B-splines are well-developed by mathematicians, but they are rarely used in the statistical community. To smooth over non-rectangular domains, the domain is approximated by a set of triangles, or a triangulation, where each triangle is defined by its three vertices. The B-splines are defined according to a set of triples which correspond to set of knots over the bivariate domain. See Dahmen et al. (1992) and Seidel (1991) for details.

4.4 Smoothing Parameter Selection

As with the RKHS framework and accompanying smoothing spline representation, the smoothing matrix

$$A_\lambda = X\big((XB)' D^{-1} XB + \lambda_l P_l + \lambda_m P_m \big)^{-1} (XB)' D^{-1}$$

and its properties play an integral role in selecting the optimal smoothing parameter in any regularized regression, including the P-spline framework. We discussed the leave-one-subject-out cross validation score (3.38) and its computationally efficient approximation (3.41) in Chapter 3, which

rely directly on the smoothing matrix for calculation. The model selection criteria discussed in Section 3.2.3 can be calculated as in the smoothing spline setting by replacing $\tilde{A}_{\lambda,\boldsymbol{\theta}}$ with A_λ. For detailed discussion of P-spline model selection with respect to multiple smoothing parameters, see Wood (2017).

4.5 The P-spline Estimator for the Innovation Variance Function

The P-spline estimator for the log innovation variance function is constructed via penalized similarly to the smoothing spline estimator in Section 3.3. Fixing $\phi = \phi^*$ at an estimate ϕ^* of ϕ, the the log likelihood of the squared working residuals can be written as in (3.42)

$$-2\ell\left(\sigma^2|Z_1,\ldots,Z_N,\phi,\right) = \sum_{i=1}^{N}\sum_{j=1}^{p_i}\log\sigma_{ij}^2 + \sum_{i=1}^{N}\sum_{j=1}^{p_i}\frac{z_{ij}}{\sigma_{ij}^2},$$

where $\epsilon_{ij} = y_{ij} - \sum_{k<j}\phi_{ijk}^* y_{ik}$, and $z_{ij} = \epsilon_{ij}^2$. We can approximate $\eta(t) = \log\sigma^2(t)$ using a B-spline basis expansion, letting

$$\eta(t) = \sum_{j=1}^{k_t}\theta_j B_j(t).$$

Model estimation and smoothing parameter selection can be carried out using performance-oriented iteration as described in Section 3.3.2, substituting the above expansion for η and trading the smoothing spline penalty for the discrete difference penalty (4.3). For detailed presentation of optimization procedures, see Marx and Eilers (1999).

Chapter 5: Simulation Studies

In this section we compare bivariate spline estimators of the Cholesky factor to other methods of covariance estimation. Our primary comparisons are that with the parametric polynomial estimator proposed by Pourahmadi (1999), Pan and Mackenzie (2003), and Pourahmadi and Daniels (2002), which is also based on the modified Cholesky decomposition, and with the oracle estimator, which effectively gives a lower bound on the risk for given covariance structure. As a benchmark, we also include the sample covariance matrix, and two regularized variants of it: the tapered sample covariance matrix (Cai et al., 2010) and the soft thresholding estimator (Rothman et al., 2009), which do not rely on a natural ordering among the variables. In the simulations, the smoothing spline estimator of the modified Cholesky decomposition was constructed using the framework of a tensor product cubic smoothing spline. For each covariance matrix used for simulation, the P-spline estimator was constructed so that the order of the difference penalties for l and m are treated as additional tuning parameters.

Simulations were carried out for five covariance structures: the diagonal covariance with homogenous variances, a heterogenous autoregressive process with linear varying coefficient function, the same heterogeneous process but truncated to zero to band the inverse covariance matrix, the rational quadratic covariance model, and the compound symmetric model. The two-dimensional surfaces corresponding to each of these are shown left to right in Figure 5.1. The first row of image plots display the surface which coincides with the appropriate discrete covariance

matrix, and in the second row are the surface maps of the corresponding Cholesky factors. Precise models used for simulations are defined in Table 5.1.

Table 5.1: *Covariance models used for data generation in the simulation study.*

I: Mutual independence	
$\Sigma = \mathrm{I}$	$\tilde{\phi}(t,s) = 0, \quad 0 \le s < t \le 1,$ $\sigma^2(t) = 1, \quad 0 \le t \le 1.$
II: Linear varying coefficient function, constant innovation variances	
$\Sigma = T^{-1}DT'^{-1}$	$\tilde{\phi}(t,s) = t - \frac{1}{2}, \quad 0 \le t \le 1,$ $\sigma^2(t) = 0.1^2, \quad 0 \le t \le 1.$
III: Banded linear varying coefficient function, constant innovation variances	
$\Sigma = T^{-1}DT'^{-1}$	$\tilde{\phi}(t,s) = \begin{cases} t - \frac{1}{2}, & t - s \le 0.5 \\ 0, & t - s > 0.5 \end{cases},$ $\sigma^2(t) = 0.1^2, \quad 0 \le t \le 1.$
IV: Rational quadratic covariance	
$\Sigma = [\sigma_{ij}]$	$\sigma_{ij} = \left(1 + \frac{(t_i - t_j)^2}{2\alpha k^2}\right)^{-\alpha}, \quad 0 < t_i, t_j < 1$ $k = 0.6, \quad \alpha = 1$
V: Compound symmetry	
$\Sigma = \sigma^2(\rho \mathrm{J} + (1-\rho)\mathrm{I}),$ $\rho = 0.7, \quad \sigma^2 = 1$	$\tilde{\phi}(t,s) = \frac{\rho}{1+(t-2)\rho}, \quad \begin{matrix} t = 2, \ldots, p, \\ s = 1, \ldots, t-1 \end{matrix}$ $\sigma^2(t) = \begin{cases} 1, & t = 1 \\ 1 - \frac{(t-2)\rho^2}{1+(t-2)\rho}, & t = 2, \ldots, p \end{cases}$

The smallest elements of each matrix correspond to dark green pixels, while the light pink (white) pixels correspond to the large (largest) elements of the matrix. Comparison of the covariance matrices with the generalized autoregressive coefficient function which defines lower triangular surface in the second row demonstrates that covariance structures exhibiting sparsity or parsimony do not necessarily exhibit the same simplicity in the components of the Cholesky decomposition. The Cholesky factor for Model III, the truncated linear varying coefficient AR model, is sparse, with elements on the outer half of the subdiagonals equal to zero. While this corresponds to a banded inverse covariance structure, Σ itself is not sparse. The compound symmetric model has simple structure and is parsimonious; its dependence parameters can be expressed as the evaluation of a function which is constant in time t. However, the elements of the Cholesky factor and diagonal matrix of innovation variances $D = T\Sigma T'$ do not exhibit such elementary structure, the elements of which are nonlinear in t.

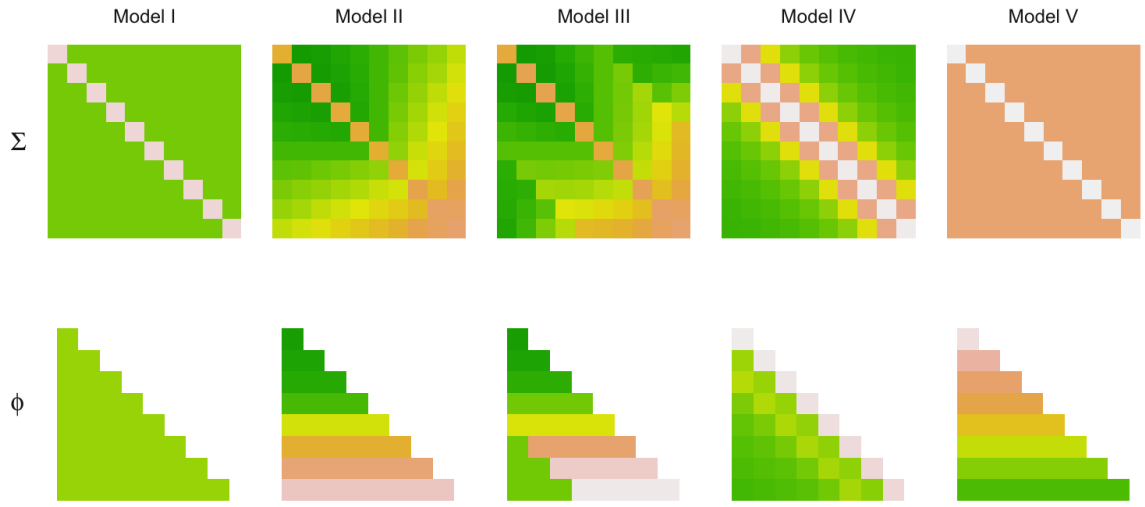


Figure 5.1: *Heatmaps of the true covariance matrices (row 1) under simulation Model I - Model V (see Table 5.1) and the function ϕ defining the corresponding Cholesky factor T (row 2).*

5.1 Loss Functions and Risk Measures

Let $\hat{\Sigma}$ be an estimator of the true $p \times p$ covariance matrix Σ. To assess performance of an estimator $\hat{\Sigma}$, we consider two commonly loss functions:

$$\Delta_1\left(\Sigma, \hat{\Sigma}\right) = tr\left(\left(\Sigma^{-1}\hat{\Sigma} - \mathrm{I}\right)^2\right), \tag{5.1}$$

$$\Delta_2\left(\Sigma, \hat{\Sigma}\right) = tr\left(\Sigma^{-1}\hat{\Sigma}\right) - log|\Sigma^{-1}\hat{\Sigma}| - p. \tag{5.2}$$

Each of these loss functions is 0 when $\hat{\Sigma} = \Sigma$ and is positive when $\hat{\Sigma} \neq \Sigma$. Both measures of loss are scale invariant. If we let random vector Y have covariance matrix Σ, and define the Z as some

linear transformation of Y:

$$Z = CY.$$

for some $p \times p$ matrix C, then Z has covariance matrix $\Sigma_Z = C\Sigma C'$. Given an estimator $\hat{\Sigma}$ of Σ, one immediately obtains an estimator for Σ_Z, $\hat{\Sigma}_Z = C\hat{\Sigma}C'$. If C is invertible, then the loss functions Δ_1 and Δ_2 satisfy

$$\Delta_i\left(\Sigma, \hat{\Sigma}\right) = \Delta_i\left(C\Sigma C', C\hat{\Sigma}C'\right).$$

The first loss Δ_1, or the quadratic loss, measures the discrepancy between $\left(\Sigma^{-1}\hat{\Sigma}\right)$ and the identity matrix with the squared Frobenius norm. The Frobenius norm of a matrix A is given by

$$||A||_F^2 = \text{tr}\left(AA'\right).$$

The second loss Δ_2 is commonly referred to as the entropy loss; it gives the Kullback-Leibler divergence of two multivariate Normal densities with the same mean and the two corresponding covariance matrices. The quadratic loss penalizes overestimates more than underestimates, so "smaller" estimates are favored more under Δ_1 than Δ_2. For example, among the class of estimators comprised of scalar multiples cS of the sample covariance matrix, Haff (1980) established that S is optimal under Δ_2, while the smaller estimator $\frac{N}{N+p+1}S$ is optimal under Δ_1.

Given Σ, the corresponding values of the risk functions are obtained by taking expectations:

$$R_i\left(\Sigma, \hat{\Sigma}\right) = E_\Sigma\left[\Delta_i\left(\Sigma, \hat{\Sigma}\right)\right], \quad i = 1, 2.$$

We prefer an estimator $\hat{\Sigma}$ with smaller risk. Given Σ, we can estimate the risk of an estimator via Monte Carlo approximation.

5.2 Alternative Estimators

The following estimators serve as benchmarks for performance under the five simulation settings outlined above: the MCD polynomial estimator $\hat{\Sigma}_{poly}$, the sample covariance matrix S, the

soft thresholding estimator S^{λ}, and the tapering estimator S^{ω}. We will review the general definitions of these, but for detailed discussion of the construction and properties of these estimators, see Sections 2.2 and 2.3.

In the spirit of the GLM, the MCD polynomial estimator is a particular case of estimators which model the components of the Cholesky decomposition using covariates. The polynomial estimator takes the GARPs and IVs to be polynomials of lag and time, respectively:

$$\phi_{jk} = x'_{jk}\beta$$

$$\log \sigma_j^2 = z'_j\gamma,$$

for $j = 2, \ldots, p$, $k = 1, \ldots, j-1$. The vectors z_j and x_{jk} are of dimension $q \times 1$ and $d \times 1$ which hold covariates

$$\begin{aligned} x'_{jk} &= \left(1, t_j - t_k, (t_j - t_k)^2, \ldots, (t_j - t_k)^{d-1}\right)', \\ z'_j &= \left(1, t_j, \ldots, t_j^{q-1}\right)', \end{aligned} \tag{5.3}$$

where the orders of the polynomials, d and q, are chosen by BIC.

Rothman et al. (2009) presented a class of generalized thresholding estimators, including the soft-thresholding estimator given by

$$S^{\lambda} = \left[\text{sign}\left(s_{ij}\right)\left(s_{ij} - \lambda\right)_+\right],$$

where s_{ij} denotes the i-j^{th} entry of the sample covariance matrix, and λ is a penalty parameter controlling the amount of shrinkage applied to the empirical estimator.

The tapering estimator proposed by Cai et al. (2010) is given by

$$S^{\omega} = \left[\omega_{ij}^k s_{ij}\right],$$

where the ω_{ij}^k are given by

$$\omega_{ij}^k = k_h^{-1}\left[\left(k - |i-j|\right)_+ - \left(k_h - |i-j|\right)_+\right].$$

The weights ω_{ij}^k are controlled by a tuning parameter, k, which can take integer values between 0 and p. Without loss of generality, we assume that $k_h = k/2$ is even. The weights may be rewritten as

$$\omega_{ij} = \begin{cases} 1, & |i-j| \leq k_h \\ 2 - \frac{i-j}{k_h}, & k_h < |i-j| \leq k, \\ 0, & \text{otherwise.} \end{cases}$$

Tuning parameter selection for the regularized versions of the sample covariance matrix was performed using cross validation. Under certain conditions pertaining to the ratio of sample sizes of the training and validation datasets, the K-fold cross validation criterion is a consistent estimator of the Frobenius norm risk. It is defined

$$\mathbf{CV}_F(\lambda) = \arg\min_{\lambda} K^{-1} \sum_{k=1}^{K} ||\hat{\Sigma}^{(-k)} - \tilde{\Sigma}^{(k)}||_F^2. \tag{5.4}$$

There is little established about the optimal method for tuning parameter selection for the class of estimators based on elementwise shrinkage of the sample covariance matrix. However, based on the results of an extensive simulation study presented in Fang et al. (2016), we use 10-fold cross validation to select the tuning parameters for both the tapering estimator S^{ω} and the soft thresholding estimator S^{λ}. The authors implement cross validation for a number of element-wise shrinkage estimators for covariance matrices in the Wang (2014) R package, which was used to calculate the risk estimates for S^{ω} and S^{λ}.

As discussed in Chapter 2, in the limit, soft thresholding produces a positive definite estimator with probability tending to 1 (Rothman et al., 2009), however element-wise shrinkage estimators of the covariance matrix, including the soft thresholding estimator, are not guaranteed to be positive definite. We observed simulations runs which yielded a soft thresholding estimator that was indeed not positive definite. In this case, the estimate has at least one eigenvalue less than or equal to zero, and the evaluation of the entropy loss (5.2) is undefined. To enable the evaluation of the

entropy loss, we coerced these estimates to the "nearest" positive definite estimate via application of the technique presented in Cheng and Higham (1998). For a symmetric matrix A, which is not positive definite, a modified Cholesky algorithm produces a symmetric perturbation matrix E such that $A + E$ is positive definite.

Pan and Mackenzie (2003) present an iterative procedure for estimating coefficient vectors β, γ of the polynomial model (5.2). Their algorithm uses a quasi-Newton step for computing the MLE under the multivariate normal likelihood. Their work is implemented in the JMCM package for R, which we used to compute the polynomial MCD estimates. For implementation details, see Pan and Pan (2017).

In addition to these estimators, we include risk estimates for the oracle estimator for each of the simulation models in Table 5.1, which serves as a practical lower bound for the risk under each generating model. For the case of mutual independence with constant variance, the oracle estimator of the covariance matrix 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_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, p_i$. The oracle estimator for Model II is obtained by fitting the model

$$y\left(t_{ij}\right) = \sum_{k<j} \left(\beta_0 + \beta_1 t_{ik}\right) y\left(t_{ik}\right) + \epsilon\left(t_{ij}\right), \tag{5.5}$$

where $\epsilon\left(t_{ij}\right)$ are independent mean zero Normal random variables with common variance σ^2. The estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, $\hat{\boldsymbol{\beta}}$ is taken to be

$$\arg\min_{\beta} ||Y - XB\beta||^2, \tag{5.6}$$

where X denotes the matrix of autoregressive covariates as defined in (3.22) and Example 1, and the matrix B contains the basis for a linear function of t:

$$\begin{bmatrix} 1 & t_{11} \\ 1 & t_{12} \\ \vdots & \vdots \\ 1 & t_{1,p_1} \\ \vdots & \vdots \\ 1 & t_{N,1} \\ \vdots & \vdots \\ 1 & t_{N,p_N} \end{bmatrix}.$$

The estimator for σ^2 is then the mean of the squared residuals:

$$\hat{\sigma}^2 = \frac{1}{N}\left[\sum_{i=1}^{N} \frac{1}{(p_i - 1)} \sum_{j=1}^{p_i} e_{ij}^2\right],$$

where $e_{i1} = y_{i1}$, $i = 1, \ldots, N$. The oracle estimator for Model III is obtained in the same fashion, but $y(t_{ij})$ is regressed only on its predecessors such that $t_{ij} - t_{ik} < 0.5$:

$$y(t_{ij}) = \sum_{t_{ij} - t_{ik} < 0.5} (\beta_0 + \beta_1 t_{ik})\, y(t_{ik}) + \epsilon(t_{ij}). \tag{5.7}$$

The oracle estimator under Model IV, the rational quadratic covariance model, assumes that $Y_1, \ldots, Y_N$ is a random sample from a mean zero multivariate normal distribution with covariance matrix $\Sigma = \left[\sigma_{ij}\right]$, where the elements of the covariance matrix are defined according to the parametric function given in Table 5.1.

The compound symmetric covariance Model V can be written as a simple random effects model:

$$Y_i = Z_i b_i + \boldsymbol{\epsilon}_i, \tag{5.8}$$

where $\boldsymbol{\epsilon}_i$ is a vector of residuals from a $N(0, \sigma_\epsilon^2)$ distribution, and the b_i are independent $N(0, \sigma_b^2 \mathrm{I})$ random vectors, the elements of which are mutually independent of the elements of ϵ_i. The matrix

of covariates corresponding to the random effects contains only an intercept term:

$$Z_i = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Under this model, the within-subject covariance structure is given by

$$Cov\left(Y_i\right) = \sigma_\epsilon^2 \mathrm{I} + \sigma_b^2 11'.$$

The oracle estimator can be obtained using restricted maximum likelihood estimation under a Normal likelihood with this covariance structure.

5.3 Data Generation Procedures

For each of the covariance models, we generated a set of observations of sample size $N = 50, 100$ from a multivariate normal distribution for each of three different values of within-subject sample size $p = 10, 20, 30$. To generate data according to Models II and III, which are parameterized in terms of the components of the Cholesky decomposition, the Cholesky factor T and diagonal innovation variance matrix D are constructed by evaluating $\tilde{\phi}$ and σ^2 at the fixed observation times. The data are then sampled according to the multivariate normal distribution with covariance matrix $\Sigma = T^{-1}DT'^{-1}$. Given covariance matrix Σ, risk estimates are obtained from$N_{sim} = 100$ samples from an p-dimensional multivariate Normal distribution with mean zero and the same covariance. Since construction of the sample covariance matrix S, S^ω, and S^λ rely on having an equal number of regularly-spaced observations on each subject, simulations comparing performance across estimators were conducted using complete data with common measurement times across all N subjects. The observation times, which are equally spaced, are mapped from the integers $1, 2, \ldots, p$ to the unit interval for estimation.

Our second concern in evaluation of our methods is how performance changes when the data exhibit varying degrees of sparsity. We fix the number of sampled trajectories N and vary p, the size of the set of possible measurement times

$$t_1, \ldots, t_p.$$

We generate irregular data by first generating a complete dataset as we did for the first simulation study:

$$\begin{aligned} Y_1 &= \left(y_1\left(t_1\right), y_1\left(t_2\right), \ldots, y_1\left(t_p\right)\right)' \\ Y_2 &= \left(y_2\left(t_1\right), y_2\left(t_2\right), \ldots, y_2\left(t_p\right)\right)' \\ &\vdots \\ Y_N &= \left(y_N\left(t_1\right), y_N\left(t_2\right), \ldots, y_N\left(t_p\right)\right)', \end{aligned}$$

where $Y_1, \ldots, Y_N$ are independently and identically distributed according to an p-dimensional multivariate Normal distribution with mean zero and having covariance structure identical to one of Models I - V in Table 5.1. To induce sparsity, we subsample from the complete data $\{y_i\left(t_j\right)\}$, $i = 1, \ldots, N$, $j = 1, \ldots, p$, randomly omitting an observation $y_i\left(t_j\right)$ with probability 0.1, 0.2, and 0.3. For both sets of simulations, the smoothing parameters for the smoothing spline and P-spline estimators were selected using both leave-one-subject-out cross validation losoCV (λ) and unbiased risk estimate U (λ). Given the selected values of the tuning parameters, we computed the estimated covariance matrix and compared it to the true covariance matrix via entropy loss and quadratic loss.

5.4 Results

5.4.1 Simulations with Complete Data

Figure 5.2 provides a visual summary of the qualitative differences between the estimates resulting from each of the eight 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 estimation method. The surface plots of the oracle estimate in the second row serve as a point of reference for the 'gold standard' in each scenario, since the oracle estimates were constructed assuming that the functional form of the covariance is known (either the full covariance structure or the components of the Cholesky decomposition). The corresponding estimates of the Cholesky factor T for the estimators based on the modified Cholesky decomposition are shown in Figure 5.3, and the decomposition of the $\hat{T}$ corresponding to the smoothing spline ANOVA estimator $\hat{\Sigma}_{SS}$ into functional components is displayed in Figure 5.4

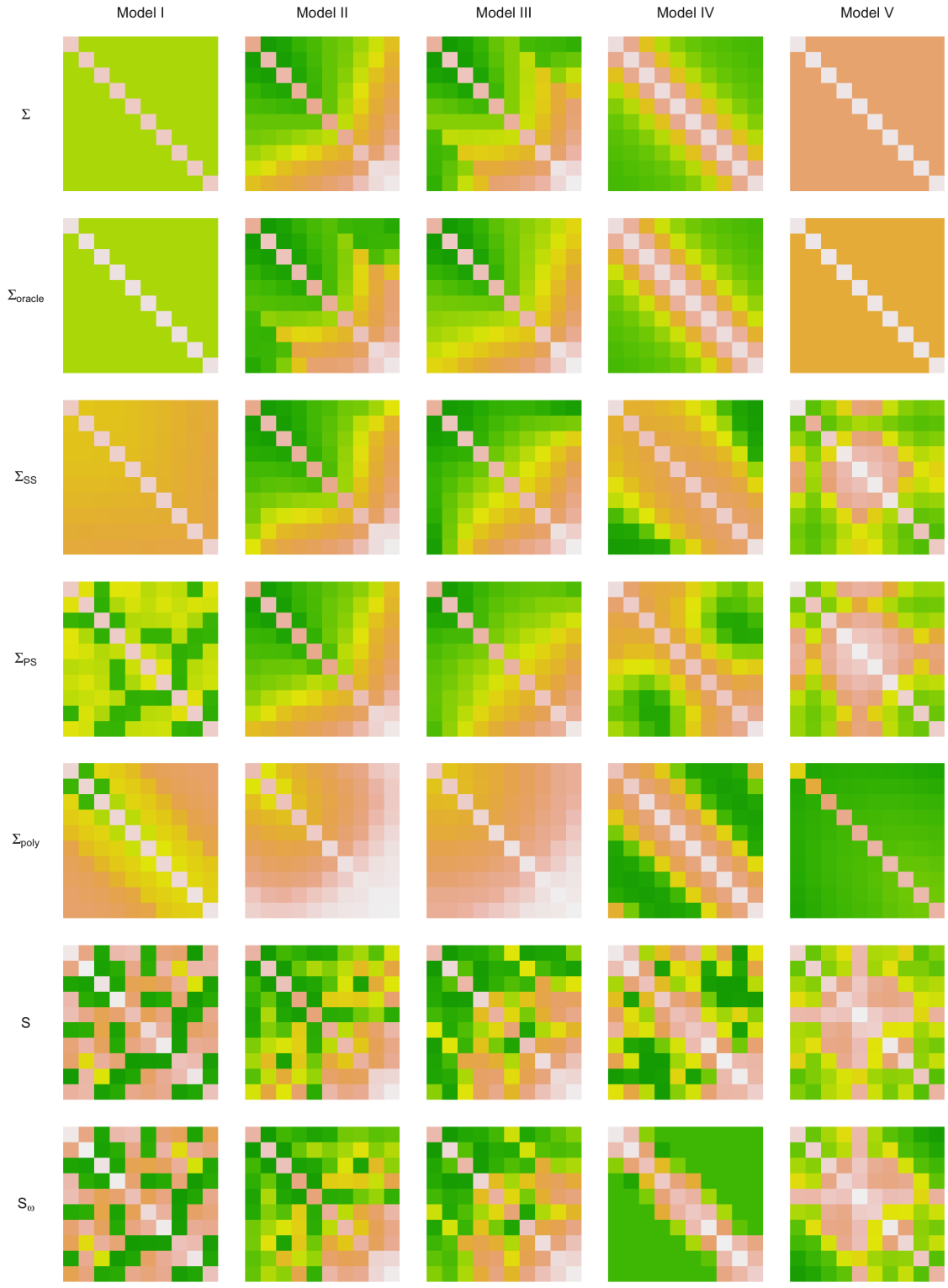


Figure 5.2: *Covariance Model I - Model V (see Table 5.1) used for simulation and corresponding estimates. The columns in the grid correspond to each simulation model. The first row shows the true covariance structure, and each row beneath corresponds to each of the estimators.*

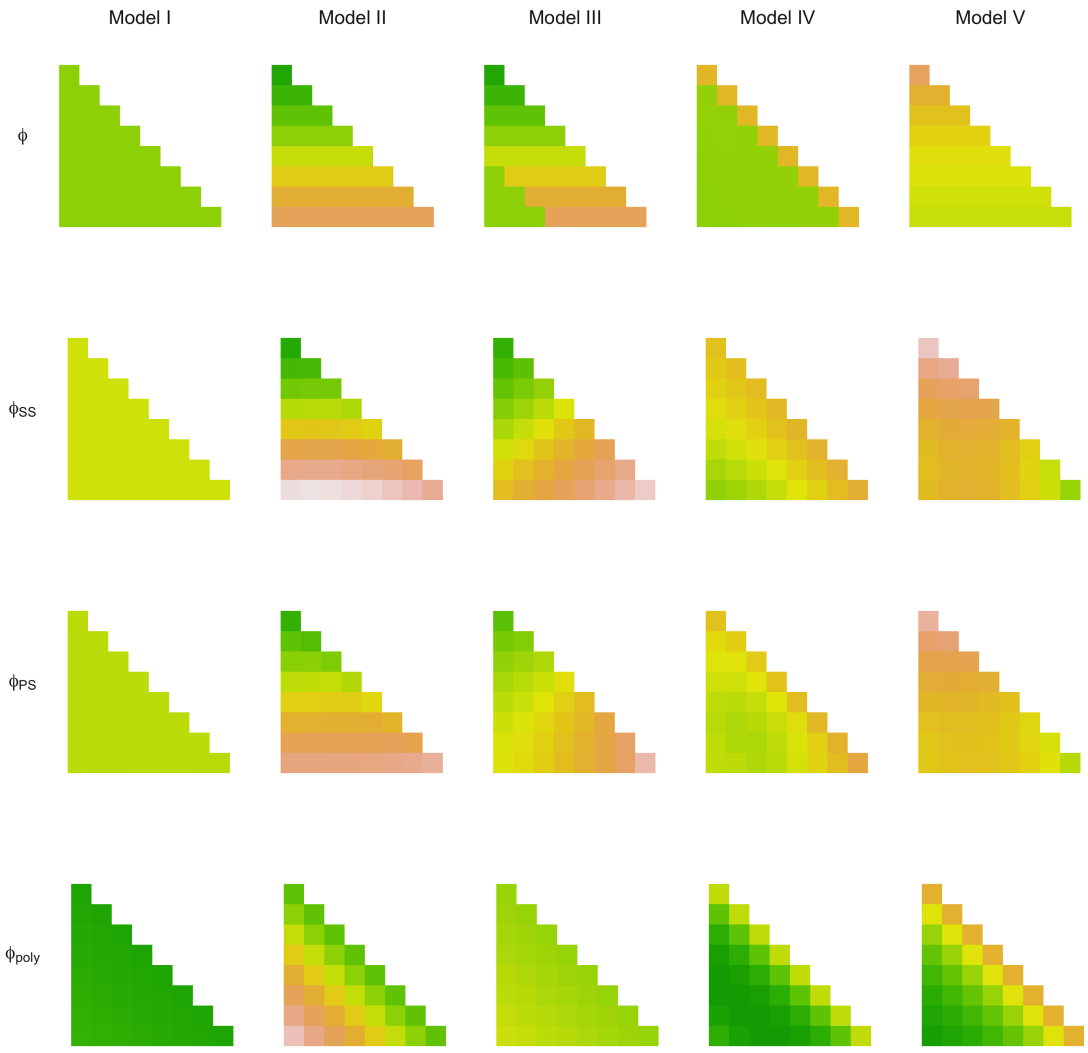


Figure 5.3: *The generalized autoregressive coefficient function ϕ which defines the elements of the true lower triangle of Cholesky factor T corresponding to Model I - Model V and estimates of the same surface for estimators based on the modified Cholesky decomposition. The true covariance structure is displayed across the top row.*

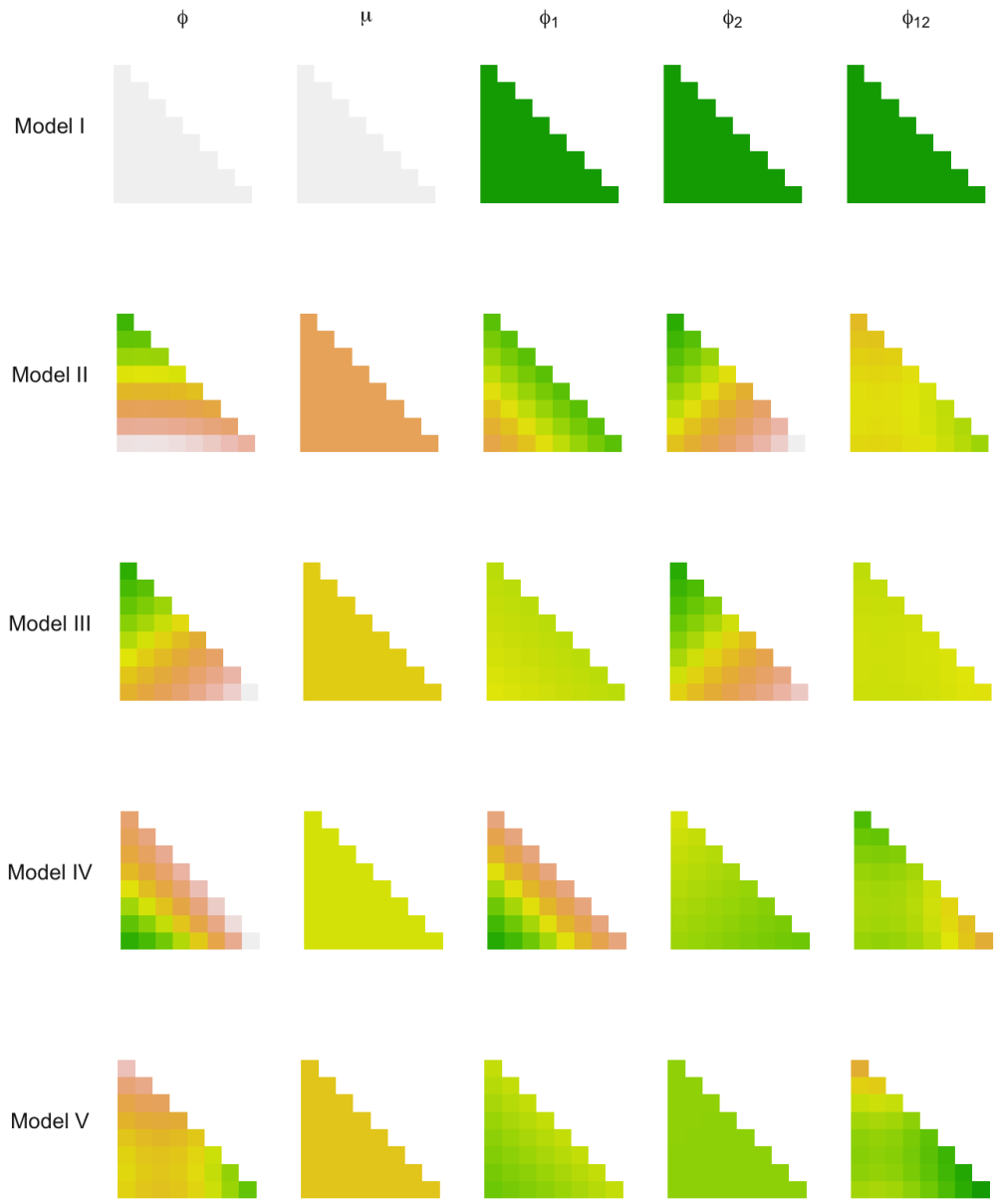


Figure 5.4: *Estimated functional components of the smoothing spline ANOVA decomposition* $\phi = \phi_1 + \phi_2 + \phi_{12}$ *for* $\hat{\Sigma}_{SS}$ *under each simulation Model I - V.*

The results of the simulations for complete data under entropy loss are presented in Tables 5.2 - 5.6, where the smoothing parameters for our smoothing spline estimator $\hat{\Sigma}_{SS}$ and P-spline estimator $\hat{\Sigma}_{PS}$ are chosen using the unbiased risk estimate. Performance of the estimator when the smoothing

parameter is chosen using leave-one-subject-out cross validation is comparable; these results are left to Appendix C. Risk estimates under quadratic loss, while there is not agreement between results every time, qualitatively, they are similar in nature to those with entropy loss and are also presented in Appendix C, Tables C.1-C.5. Since both loss functions are not standardized, they cannot be compared across dimensions p.

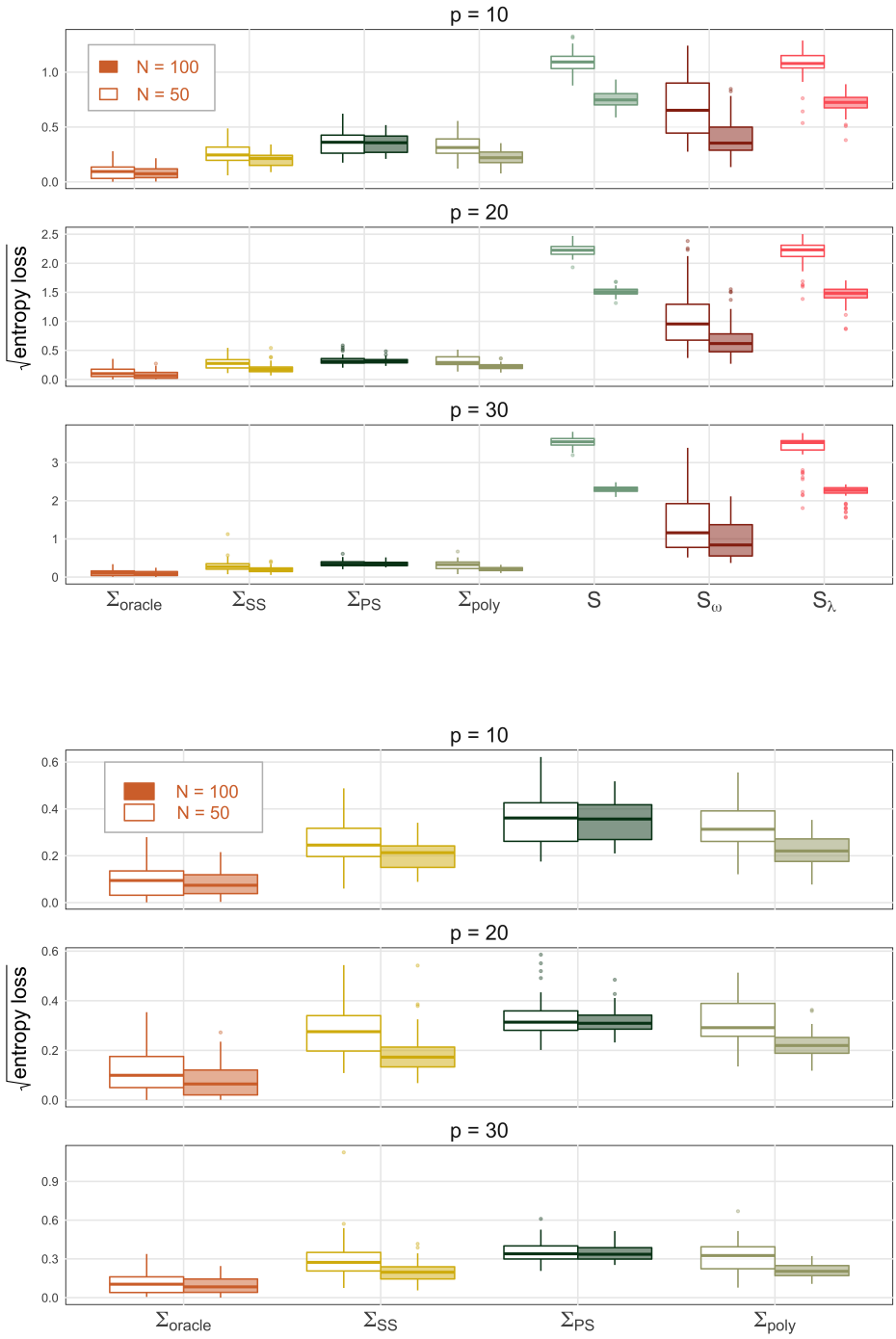


Figure 5.5: *The distribution of the square root of entropy loss Δ_2 under Model I for each of the estimators considered in the simulations with complete data. The top figure displays results under each of the three data dimensions $p = 10, 20, 30$ and sample sizes $N = 50, 100$. The relative performance of each estimator compared to the others is consistent across dimensions. The bottom figure shows a more detailed view for comparison between the oracle estimator and the estimators based on the modified Cholesky decomposition.*

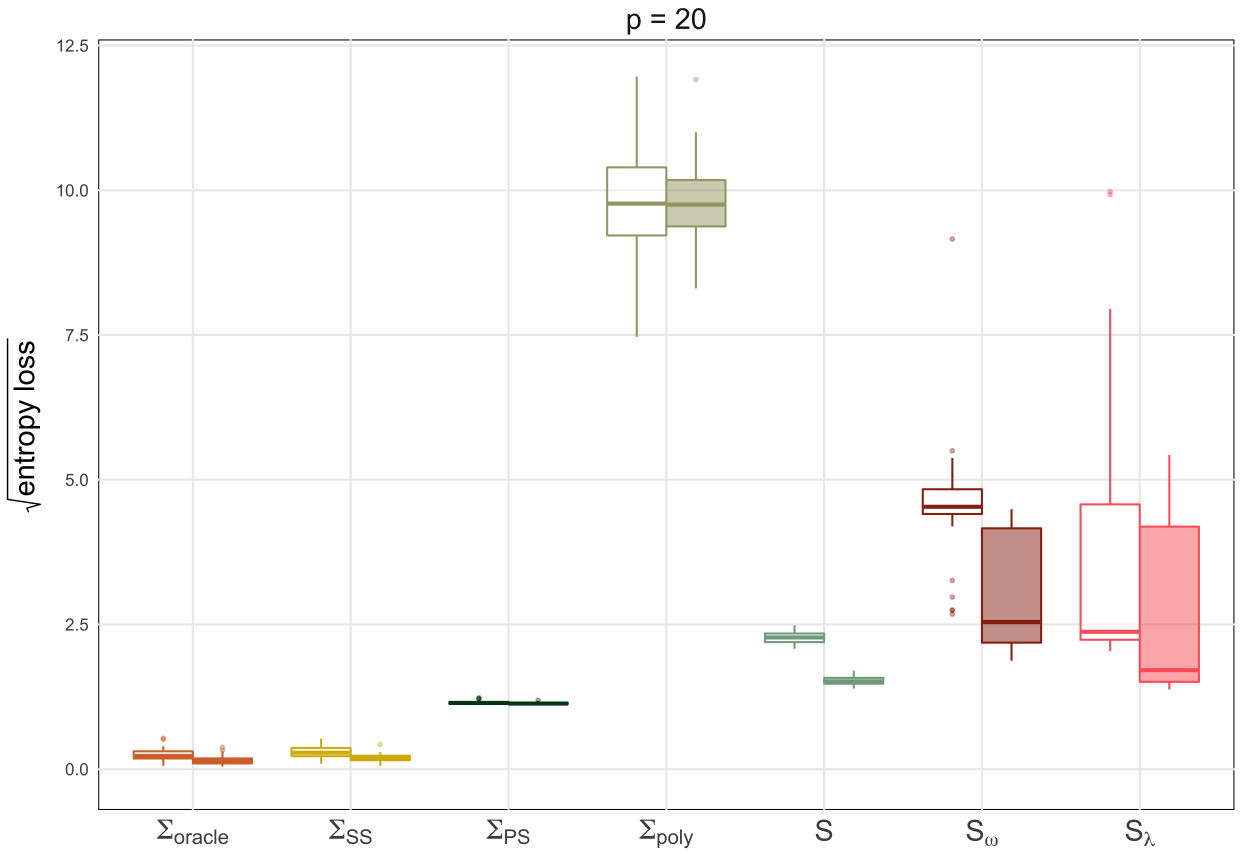


Figure 5.6: *The distribution of the square root of entropy loss Δ_2 under Model III for the smoothing spline estimator $\hat{\Sigma}_{SS}$ for 100 simulated multivariate normal samples of size $N = 50$ (hollow boxplots) and $N = 100$ (filled boxplots) with dimension $p = 20$. The estimator Σ_{poly} based on the parametric model for the modified Cholesky decomposition suffers from model misspecification, which is clearly reflected in its performance as measured by Δ_2.*

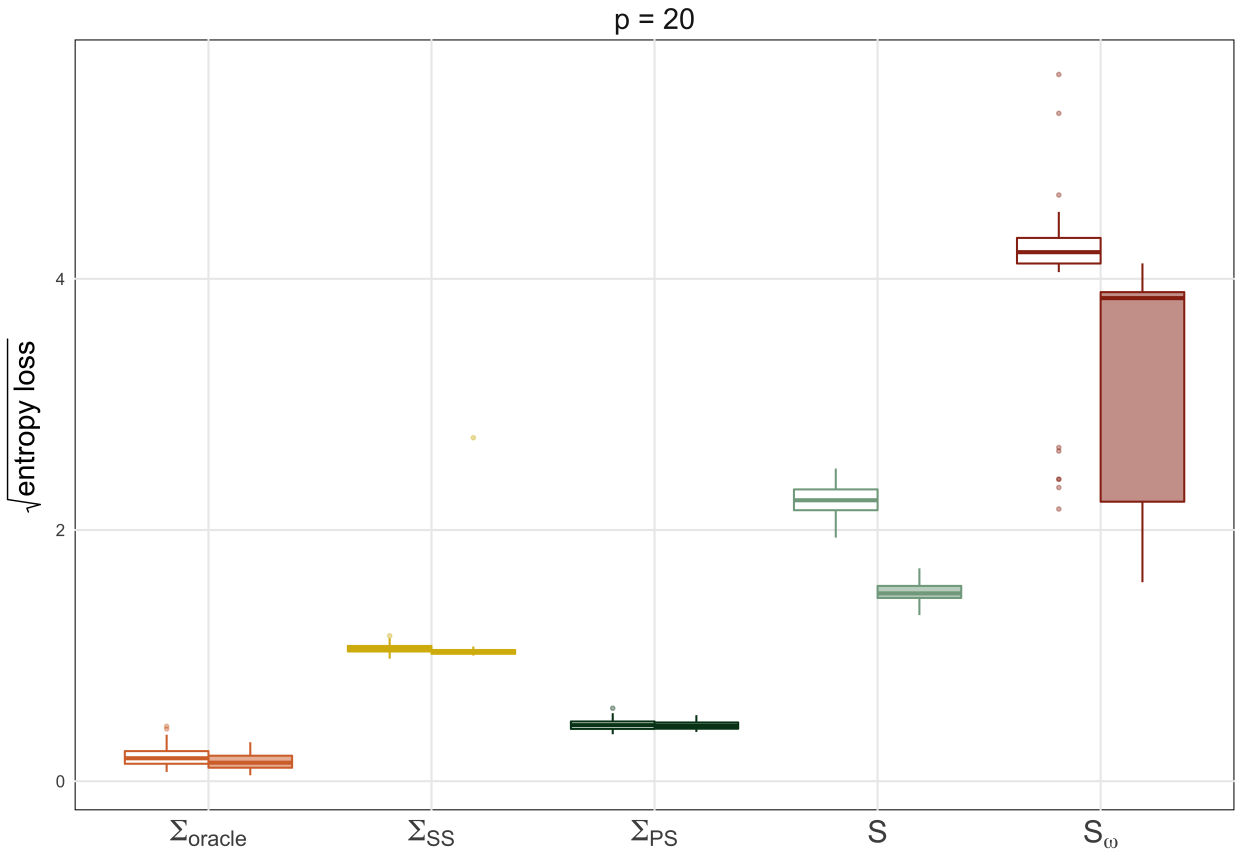


Figure 5.7: *The distribution of the square root of entropy loss Δ_2 under Model III for the smoothing spline estimator $\hat{\Sigma}_{SS}$ for 100 simulated multivariate normal samples of size $N = 50$ (hollow boxplots) and $N = 100$ (filled boxplots) with dimension $p = 20$. The loss distributions are shown for the sample covariance matrix as well as the tapered sample covariance matrix. While the underlying inverse covariance matrix is banded, the covariance matrix itself is not, which explains why S outperforms its regularized variant S_ω.*

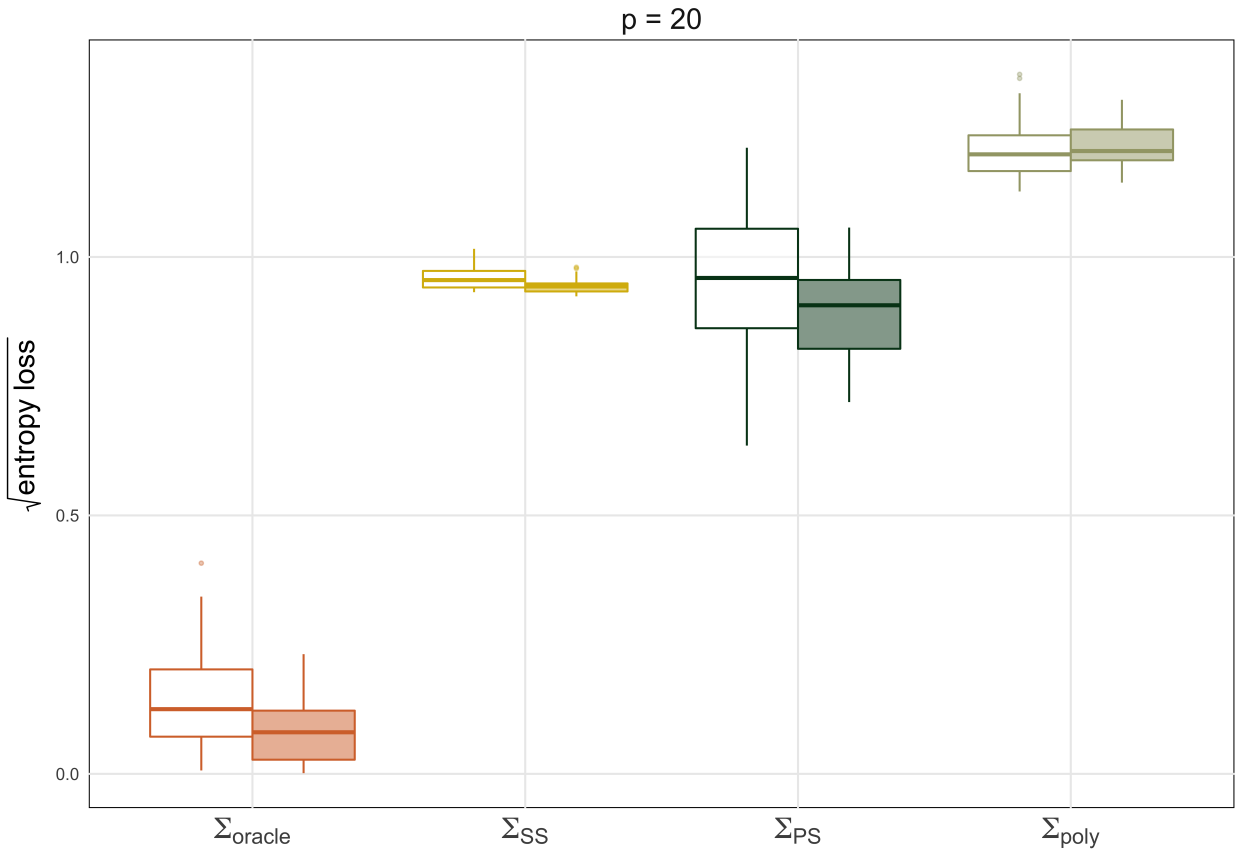


Figure 5.8: *The distribution of the square root of entropy loss Δ_2 under Model IV for the smoothing spline estimator $\hat{\Sigma}_{SS}$ for 100 simulated multivariate normal samples of size $N = 50$ (hollow boxplots) and $N = 100$ (filled boxplots) with dimension $p = 20$. The performance of the estimator Σ_{poly} based on the parametric model for the modified Cholesky decomposition is more comparable to that of our estimators when the underlying covariance matrix is stationary.*

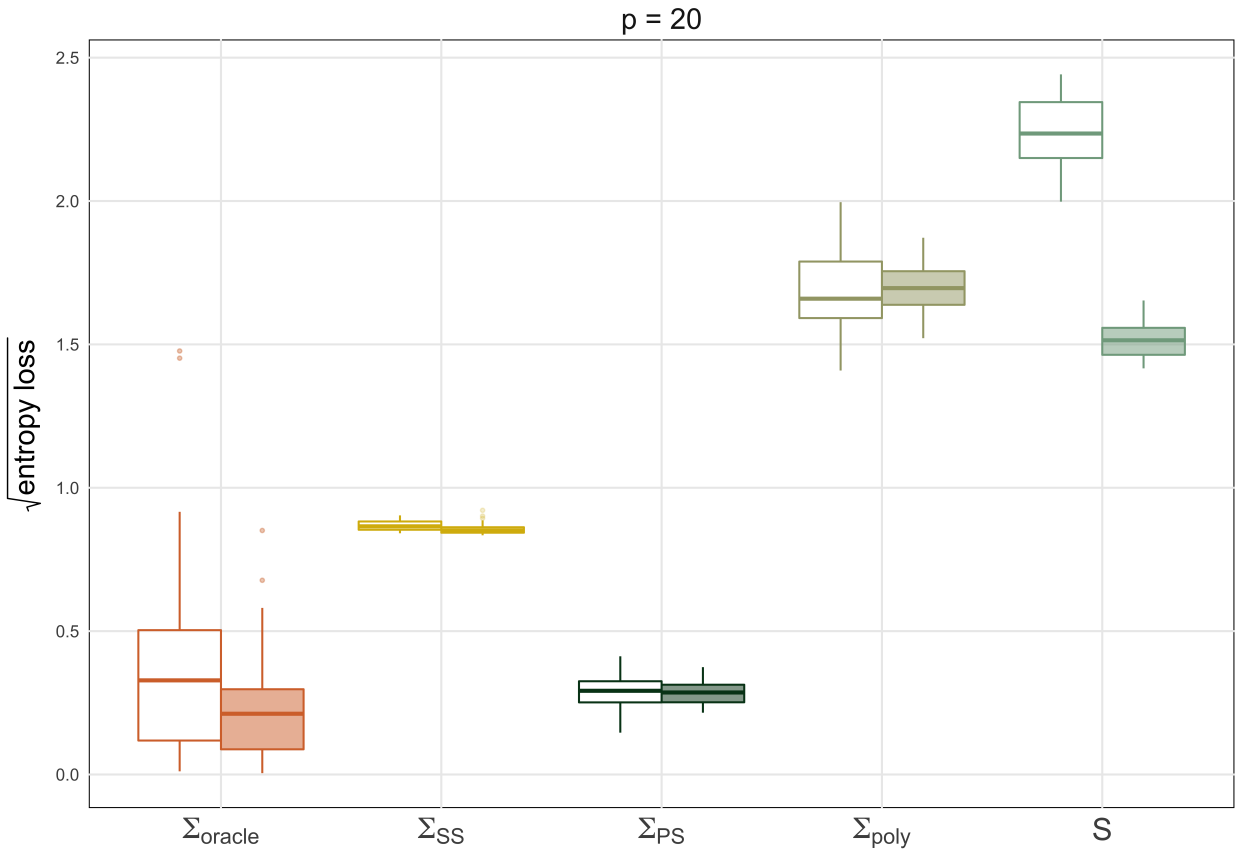


Figure 5.9: *The distribution of the square root of entropy loss* Δ_2 *under Model V for the smoothing spline estimator* $\hat{\Sigma}_{SS}$ *for 100 simulated multivariate normal samples of size* $N = 50$ *(hollow boxplots) and* $N = 100$ *(filled boxplots) with dimension* $p = 20$*. The parsimony of the compound symmetric model, having only two parameters to be estimated, lends to the marked improvement of the performance of the sample covariance when the sample size is increased from* $N = 50$ *to* 100*.*

In general, our estimators outperform the alternative estimators across the five covariance structures. This not surprising; the soft thresholding estimator assumes no ordering of the variables of the random vector, which all but one of the generating structures exhibit. The tapering estimator assumes that the absolute value of the covariance decays as l increases; only Model IV satisfies this. The parametric estimator based on the modified Cholesky decomposition assumes that ϕ can

be modeled as a univariate function of l, which does not hold for any of the models, save Model IV.

The smoothing spline estimator outperforms the P-spline estimator in cases where the underlying covariance structure cannot be modeled as a multiplicative function of l and m - namely, Model II. It also does a better job estimating the diagonal structure of Model I. The optimal difference penalty order for the P-spline under the identity covariance matrix is $d = 0$, which corresponds to a ridge penalty on the B-spline coefficients which could lead to a fitted surface which is not necessarily smooth.

Treating the differencing order as an additional tuning parameter is advantageous since searching for the optimal set of smoothing parameters is much easier when the true function belongs to the null space of the penalty. The P-spline estimator outperforms the smoothing spline estimator under Models IV and V, likely due to this advantage. While the surface of the Cholesky factor of Model IV is smooth, value of the function changes very quickly in distance from the diagonal. The local support of the B-spline basis functions aids in the P-spline estimator's ability to accommodate such fast oscillations in surface. The same can be said for Model III, which is not smooth in $l = t - s$.

Table 5.2: *Multivariate normal simulations for Model I. Estimated entropy risk is reported for the oracle estimator for each covariance structure, our smoothing spline ANOVA estimator and P-spline estimator, the parametric polynomial estimator of Pan and MacKenzie (2003), the sample covariance matrix, the tapered sample covariance matrix, and the soft thresholding estimator.*

	p	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.0135	0.0685	0.1261	0.1102	1.2047	0.5369	1.1742
	20	0.0229	0.0834	0.1713	0.1096	4.9850	1.3957	4.7796
	30	0.0196	0.1102	0.1969	0.1127	12.5517	2.8019	11.3175
$N = 100$	10	0.0105	0.0451	0.0671	0.0531	0.5685	0.2045	0.5236
	20	0.0105	0.0425	0.0965	0.0512	2.2831	0.5724	2.1358
	30	0.0139	0.0431	0.1148	0.0472	5.2770	1.2430	4.9126

Table 5.3: *Multivariate normal simulations for Model II.*

	p	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.0581	0.0689	0.3423	4.7673	1.2832	1.4644	1.1770
	20	0.0439	0.0581	1.3640	97.2334	5.1665	21.6407	39.3522
	30	0.0627	0.0811	2.6485	153.9665	12.3582	55.3674	133.9980
$N = 100$	10	0.0386	0.0457	0.2945	4.7911	0.5812	0.8335	0.5628
	20	0.0269	0.0416	1.2875	98.1989	2.3364	10.1841	10.0864
	30	0.0288	0.0367	2.4365	158.2480	5.2389	33.5207	62.5030

Table 5.4: *Multivariate normal simulations for Model III.*

	p	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.0619	0.3296	0.1065	3.0108	1.2030	1.1460	1.1467
	20	0.0695	1.1100	0.2555	62.7522	4.9824	17.2244	14.9189
	30	0.0576	2.3215	0.6242	1091.1933	12.4792	49.9135	121.7795
$N = 100$	10	0.0268	0.2904	0.0579	3.0383	0.5699	0.5545	0.5371
	20	0.0275	1.1963	0.2011	62.8960	2.2700	11.8274	9.5217
	30	0.0221	2.2811	0.3845	1105.0449	5.2234	29.1693	60.3529

Table 5.5: *Multivariate normal simulations for Model IV.*

	p	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.0217	0.3348	0.1966	0.7144	1.2218	0.7397	1.1921
	20	0.0286	0.9177	0.3499	1.4588	4.9091	1.9786	4.9206
	30	0.0283	1.5992	0.5100	2.2173	12.6114	3.7440	12.1489
$N = 100$	10	0.0125	0.3047	0.2237	0.6958	0.5570	0.3168	0.5515
	20	0.0105	0.8911	0.3704	1.4813	2.2659	0.9365	2.2474
	30	0.0134	1.5213	0.5282	2.2228	5.2106	1.9312	5.2111

Table 5.6: *Multivariate normal simulations for Model V.*

	p	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.0986	0.2769	0.2464	1.2420	1.2023	18.5222	2.9824
	20	0.2512	0.7514	0.8772	2.8557	5.0195	34.6618	13.8690
	30	0.2641	1.1776	0.9791	4.5791	12.3460	46.5437	26.1364
$N = 100$	10	0.0520	0.2416	0.1722	1.1491	0.5821	16.4081	1.7397
	20	0.0827	0.7286	0.2965	2.9080	2.2918	32.5295	5.4649
	30	0.1799	1.1813	0.4291	4.4402	5.2197	39.2914	15.4295

5.4.2 Performance with Irregularly Sampled Data

Figure 5.10 displays the distribution of entropy loss for our smoothing spline estimator when the smoothing parameters are selected using the unbiased risk estimator. Tables 5.7 - 5.11 provide summaries of these distributions as well as those for the estimator when leave-one-subject-out cross validation is used for smoothing parameter selection. Results under quadratic loss are similar to those under entropy loss and are left to Appendix C, Tables C.6 - C.10. Neither model selection perform better than the other across all of the simulation settings. This might suggest that when the estimated innovation variances are close to the true variances of the prediction residuals, using the unbiased risk estimate with the working residuals as substitute for the relative error is a reasonable

approach to modeling. Performance degradation of the estimator in the presence of missing data is highly dependent on the underlying structure of the Cholesky factor of the inverse covariance matrix. For Models I and IV, the identity matrix and the rational quadratic covariance model, performance remains fairly stable as the proportion of missing data increases. The estimator exhibits similar degrees of performance degradation under Models II, III, and V. Interestingly, these models (with the exception of Model III, which is a special case) have true varying coefficient functions which are naturally parameterized as functions of t, while the models under which the performance remain stable across increasing proportions of missing data are naturally parameterized in terms of l.

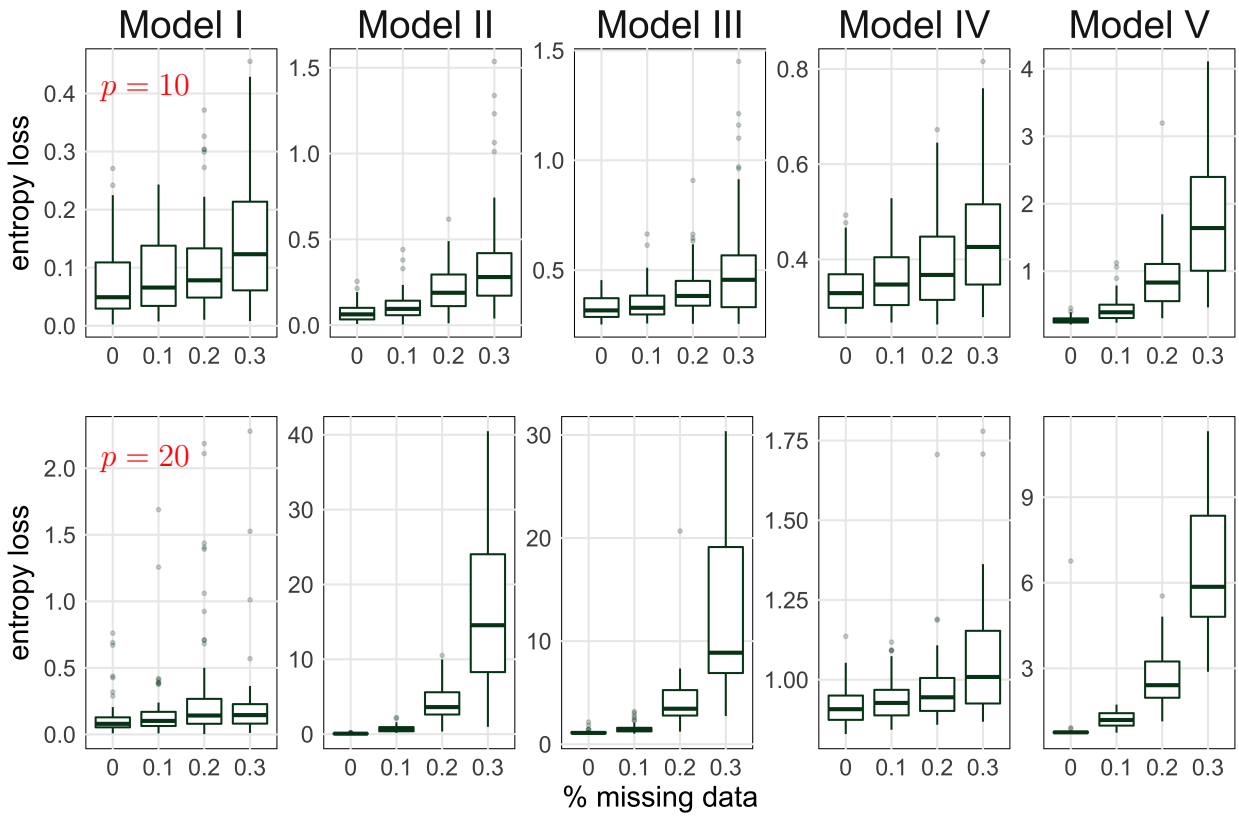


Figure 5.10: *The distribution of entropy loss for the smoothing spline estimator* $\hat{\Sigma}_{SS}$ *for 100 simulated multivariate normal samples of size* $N = 50$ *when 0%, 10%, 20%, and 30% of the data are missing for each subject for dimension* $p = 10$ *(top) and* $p = 20$ *(bottom).*

Table 5.7: *Model I: Entropy risk estimates and corresponding standard errors for the MCD smoothing spline ANOVA estimator via 100 simulated multivariate normal samples of size $N = 50$ when 0%, 10%, 20%, and 30% of the data are missing for each subject. Risk is reported for the estimator constructed using the unbiased risk estimate and leave-one-subject-out cross validation for smoothing parameter selection.*

p	% missing	$\Delta_2(\hat{\Sigma}_{SS}^{U})$		$\Delta_2(\hat{\Sigma}_{SS}^{V^*})$	
10	0.0	0.06854186	(0.0065)	0.0822183	(0.0075)
	0.1	0.08895763	(0.0080)	0.0997540	(0.0083)
	0.2	0.08474403	(0.0069)	0.1257789	(0.0110)
	0.3	0.14281452	(0.0114)	0.1552415	(0.0142)
20	0.0	0.08337738	(0.0056)	0.0924326	(0.0167)
	0.1	0.10467926	(0.0072)	0.3019903	(0.1922)
	0.2	0.13920223	(0.0076)	0.2099852	(0.0308)
	0.3	0.17160295	(0.0088)	0.3784635	(0.1054)

Table 5.8: *Model II: Entropy risk estimates and corresponding standard errors.*

p	% missing	$\Delta_2(\hat{\Sigma}_{SS}^{U})$		$\Delta_2(\hat{\Sigma}_{SS}^{V^*})$	
10	0.0	0.0689091	(0.0057)	0.0863937	(0.0070)
	0.1	0.0961388	(0.0066)	0.1396364	(0.0119)
	0.2	0.2089429	(0.0140)	0.1988000	(0.0173)
	0.3	0.2947206	(0.0212)	0.3247143	(0.0297)
20	0.0	0.0580730	(0.0042)	0.0851086	(0.0061)
	0.1	0.6508269	(0.0437)	0.6936141	(0.0366)
	0.2	3.9959421	(0.2127)	7.9307772	(2.6348)
	0.3	16.4362761	(1.3678)	24.4878411	(1.5554)

Table 5.9: *Model III: Entropy risk estimates and corresponding standard errors.*

p	% missing	$\Delta_2(\hat{\Sigma}_{SS}^{U})$		$\Delta_2(\hat{\Sigma}_{SS}^{V^*})$	
10	0.0	0.3295884	(0.0063)	0.3463639	(0.0093)
	0.1	0.3442326	(0.0079)	0.3555080	(0.0097)
	0.2	0.3922506	(0.0098)	0.4231472	(0.0138)
	0.3	0.4518739	(0.0187)	0.5270384	(0.0237)
20	0.0	1.1100351	(0.0107)	1.1312420	(0.0089)
	0.1	1.3867351	(0.0384)	1.5369483	(0.0360)
	0.2	4.4685998	(0.2608)	4.4221240	(0.2856)
	0.3	13.9195476	(1.3110)	16.5667952	(1.1101)

Table 5.10: *Model IV: Entropy risk estimates and corresponding standard errors.*

p	% missing	$\Delta_2(\hat{\Sigma}_{SS}^{U})$		$\Delta_2(\hat{\Sigma}_{SS}^{V^*})$	
10	0.0	0.3347516	(0.0056)	0.3420091	(0.0063)
	0.1	0.3561451	(0.0076)	0.3536609	(0.0079)
	0.2	0.3901020	(0.0111)	0.3884112	(0.0098)
	0.3	0.4395183	(0.0139)	0.4399004	(0.0162)
20	0.0	0.9176583	(0.0083)	0.9345338	(0.0074)
	0.1	0.9316105	(0.0101)	0.9592996	(0.0116)
	0.2	0.9620128	(0.0090)	1.0192813	(0.0201)
	0.3	1.0339355	(0.0123)	1.0986877	(0.0680)

Table 5.11: *Model V: Entropy risk estimates and corresponding standard errors.*

p	% missing	$\Delta_2(\hat{\Sigma}^U_{SS})$		$\Delta_2(\hat{\Sigma}^{V^*}_{SS})$	
10	0.0	0.2768874	(0.0054)	0.2855551	(0.0090)
	0.1	0.4139307	(0.0160)	0.4290270	(0.0161)
	0.2	0.8698641	(0.0448)	0.9289941	(0.0586)
	0.3	1.8588993	(0.1172)	2.1368920	(0.1284)
20	0.0	0.7514261	(0.0053)	0.7609570	(0.0063)
	0.1	1.2295533	(0.0522)	1.1317517	(0.0294)
	0.2	2.5715989	(0.0976)	2.4974678	(0.1081)
	0.3	7.4723499	(0.3235)	6.8275522	(0.3006)

Chapter 6: Data Analysis

Kenward (1987) reported an experiment designed to investigate the impact of the control of intestinal parasites in cattle. The grazing season runs from spring to autumn, during which cattle can potentially ingest roundworm larvae which develop from eggs deposited around the pasture from feces of previously infected cattle. Once infected, the animal is deprived of nutrients and immune resistance to disease is suppressed which can significantly impact animal growth. Monitoring the effect of a treatment for the disease requires repeated weight measurements on animals over the grazing season.

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. Weights were recorded to the nearest kilogram, and measurement times were common across animals. The longitudinal dataset is balanced, as there were no missing observations for any of the experimental units. Observed weights are shown in Figure 6.1.

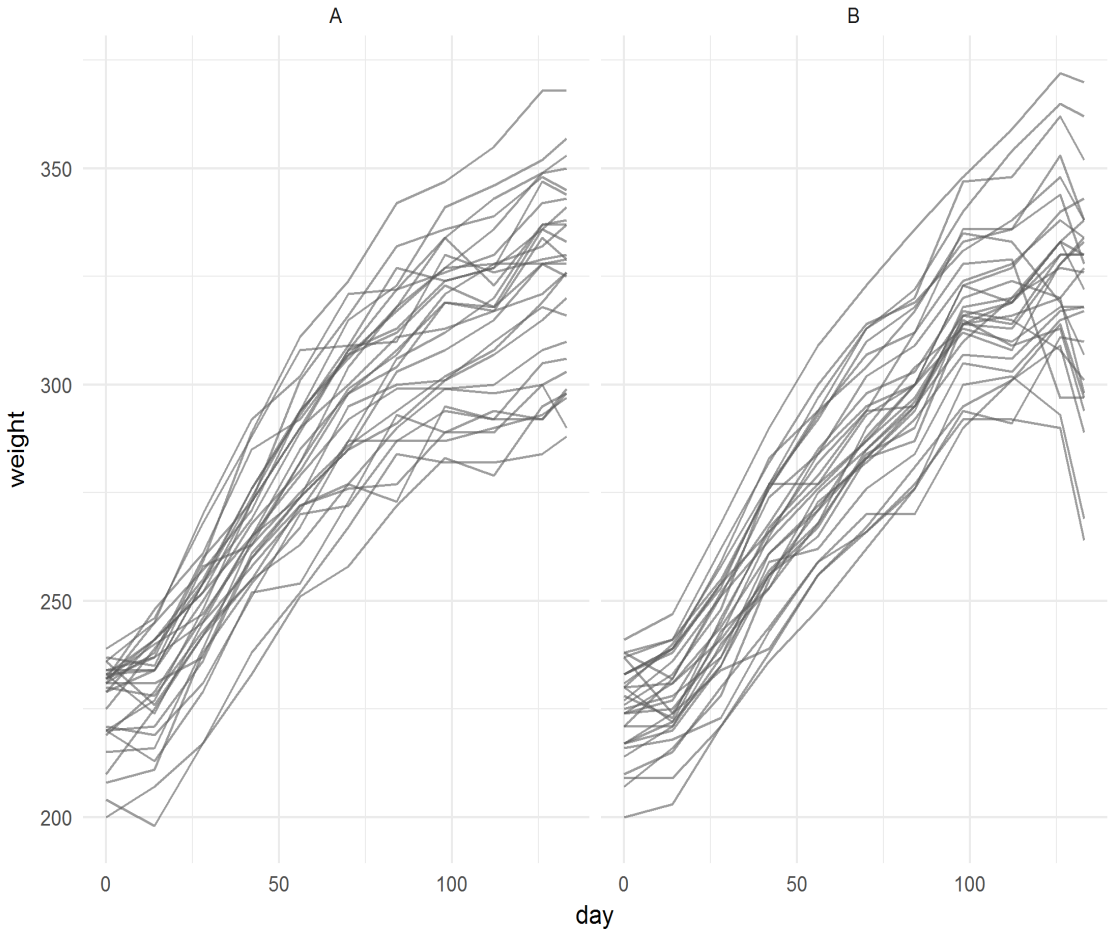


Figure 6.1: *Subject-specific weight curves over time for treatment groups A and B.*

We see an upward trend in weights over time, with variance in weights increasing over time for both groups. Treatment group B demonstrates a sharp decrease in the final weight measurement. 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 Pan and Pan (2017), Zhang et al. (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 6.1 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 6.1 is also supported by the sample correlations given in Table 6.1; correlations within the subdiagonals are not constant and increase over time, a secondary indication that a stationary covariance is not appropriate for the data. Table 6.2 gives the sample generalised autoregressive parameters and the innovation variances, which are plotted in Figure 6.2a and Figure 6.2b respectively.

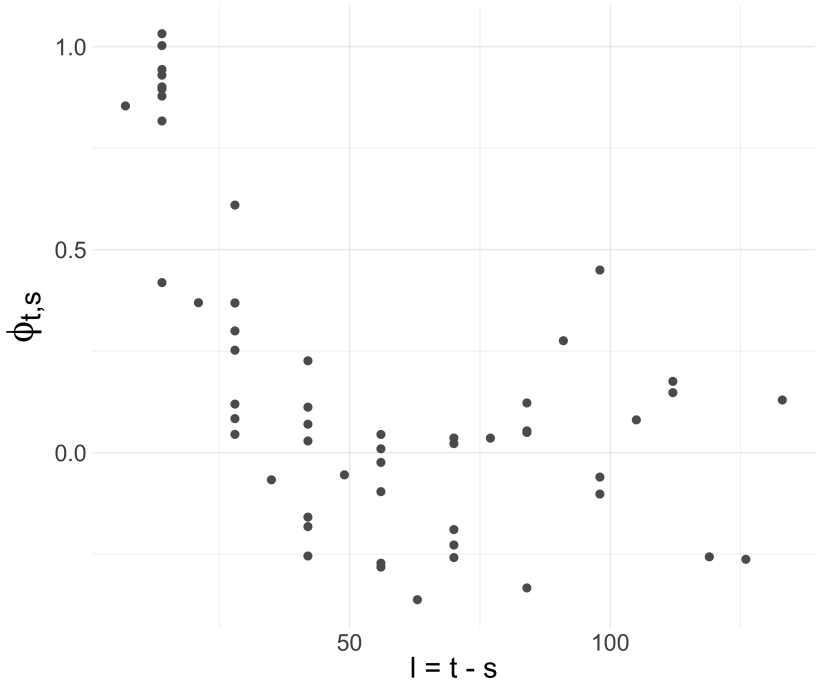
	day										
	0	14	28	42	56	70	84	98	112	126	133
0	1.00										
14	0.82	1.00									
28	0.76	0.91	1.00								
42	0.65	0.86	0.93	1.00							
56	0.63	0.83	0.89	0.93	1.00						
70	0.58	0.75	0.85	0.90	0.94	1.00					
84	0.51	0.64	0.75	0.80	0.85	0.92	1.00				
98	0.52	0.68	0.77	0.82	0.88	0.93	0.92	1.00			
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 6.1: *Cattle data: treatment group A sample correlations.*

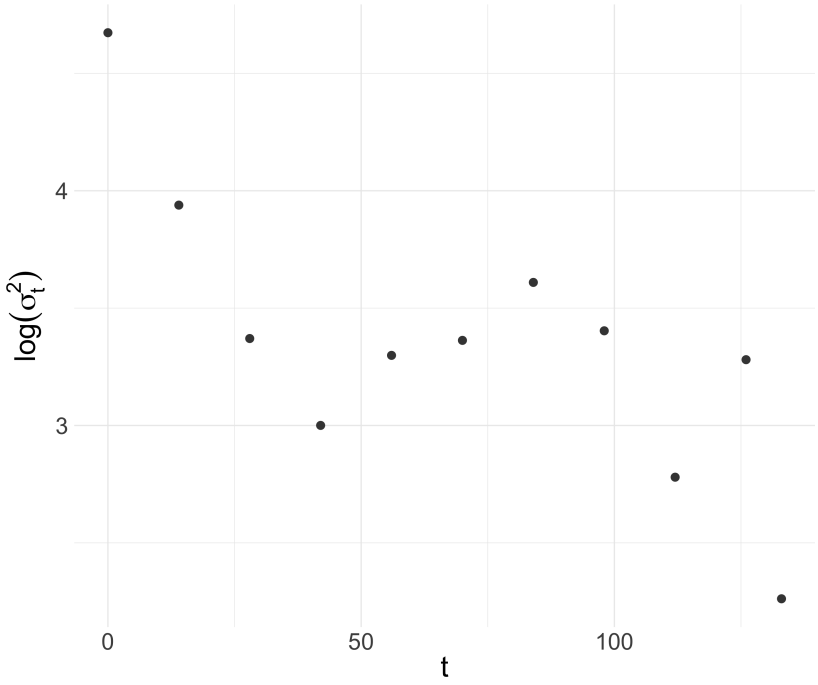
		day											
		0	14	28	42	56	70	84	98	112	126	133	
day	0	1											4.673
	14	1.00	1										3.939
	28	0.04	0.90	1									3.370
	42	-0.25	0.25	0.88	1								3.000
	56	-0.02	0.07	0.12	0.90	1							3.299
	70	0.04	-0.28	0.11	0.37	0.82	1						3.363
	84	0.12	-0.23	0.04	-0.16	0.08	1.03	1					3.610
	98	-0.06	0.05	0.02	-0.27	0.23	0.61	0.42	1				3.403
	112	0.18	-0.10	0.05	-0.26	-0.10	0.03	0.30	0.93	1			2.780
	126	-0.26	0.15	0.45	-0.33	-0.19	0.01	-0.18	0.37	0.94	1		3.280
	133	0.13	-0.26	0.08	0.28	0.04	-0.36	-0.05	-0.07	0.37	0.85	1	2.262

$\log(\hat{\sigma}_t^2)$

Table 6.2: *Cattle data: treatment group A sample generalized autoregressive parameters (below the main diagonal) and log sample innovation variances (rightmost column).*



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



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

Figure 6.2: *Empirical estimates of the parameters of the Cholesky decomposition of the sample covariance matrix.*

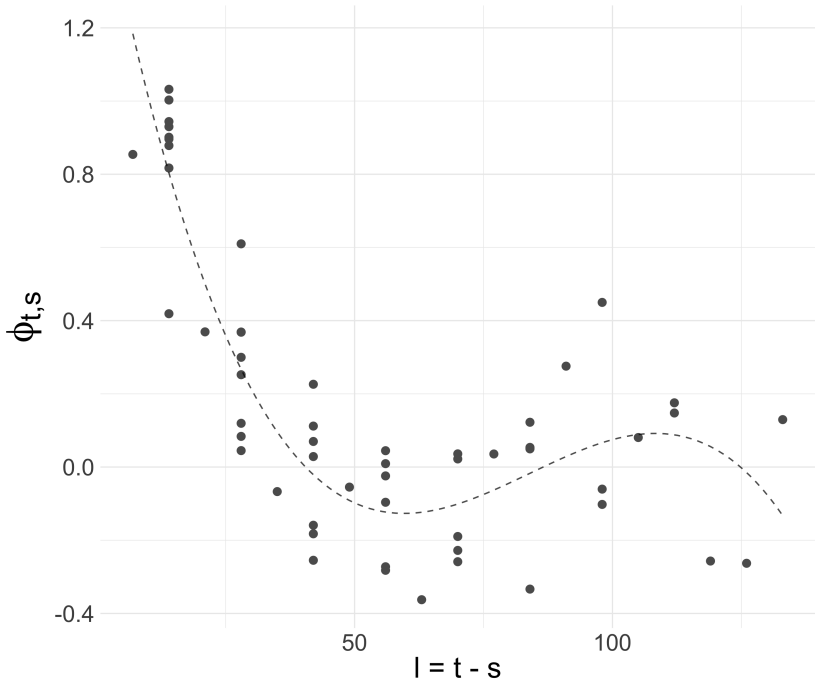
Analyzing the sample regressogram (Figure 6.2a) and sample innovation variogram (Figure 6.2b), Pourahmadi (1999) suggested that both sample generalized autoregressive parameters and the logarithms of the innovation variances can be characterized in terms of cubic functions of the lag only. They model

$$\begin{aligned} \phi_{ts} &= x'_{ts}\beta, \\ \log\left(\sigma_t^2\right) &= z'_t\gamma, \end{aligned} \tag{6.1}$$

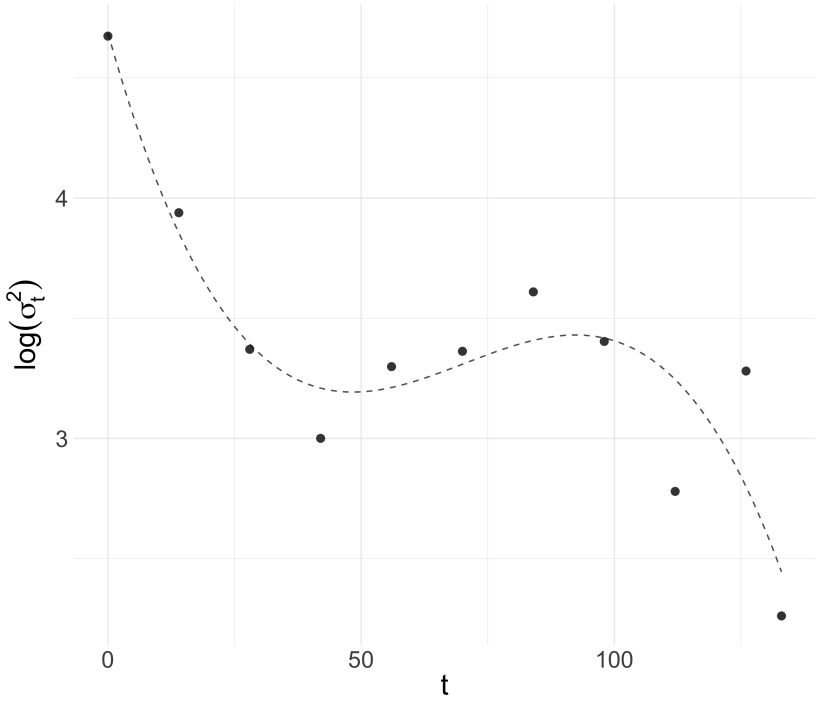
for $t = t_2, \ldots, t_{11}$ where

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

They estimate β and γ via maximum likelihood. Figure 6.3 shows the estimated cubic polynomials corresponding to Model (6.1).



(a) *Smoothed sample regressogram.*



(b) *Smoothed sample log innovation variances.*

Figure 6.3: *Cubic polynomomials fitted to the sample regressogram and log innovation variances for the cattle data from treatment group A.*

Before estimating the covariance structure, we need to center the data using an adequate estimate of the mean weight trajectories. To account for any between-subject variability, we adopt an approach akin to the dynamical conditionally linear mixed model presented in Pourahmadi and Daniels (2002):

$$Y_i = f(t_i) + Z_i b_i + \epsilon_i^*, \tag{6.2}$$

where Y_i is the $p_i \times 1$ 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 $t_i = \left(t_{i1}, \ldots, t_{i,p_i}\right)'$ is the $p_i \times 1$ vector of measurement times for subject i. We specify the random term $Z_i b_i$ as an intercept only, letting $Z_i = (1, \ldots, 1)'$ so that

$$Z_i b_i = \alpha_i 1_{p_i}.$$

The random effects α_i correspond to subject-specific shifts which are assumed to be independent and identically distributed $N(0, \sigma_\alpha^2)$ random variables. 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, $i = 1, \ldots, N$. 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 let f belong to the Hilbert space

$$\mathcal{C}^2 = \left\{ f : \ f, \ f' \text{ absolutely continuous}, \ \int \left(f''(x)\right)^2 \ dx < \infty \right\},$$

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

$$\sum_{i=1}^{N} \sum_{i=1}^{p_i} \left(y_{ij} - f(t_{ij}) - \alpha_i\right)^2 + \alpha' \Sigma_\alpha^{-1} \alpha + \lambda J(f), \tag{6.3}$$

where $\text{Cov}(\alpha) = \Sigma_\alpha = \sigma_\alpha^2 \text{I}$. The variance of the random effects σ_α^{-2} is viewed as an additional smoothing parameter and estimated alongside λ. Figure 6.4 shows the corresponding fitted mean curves.

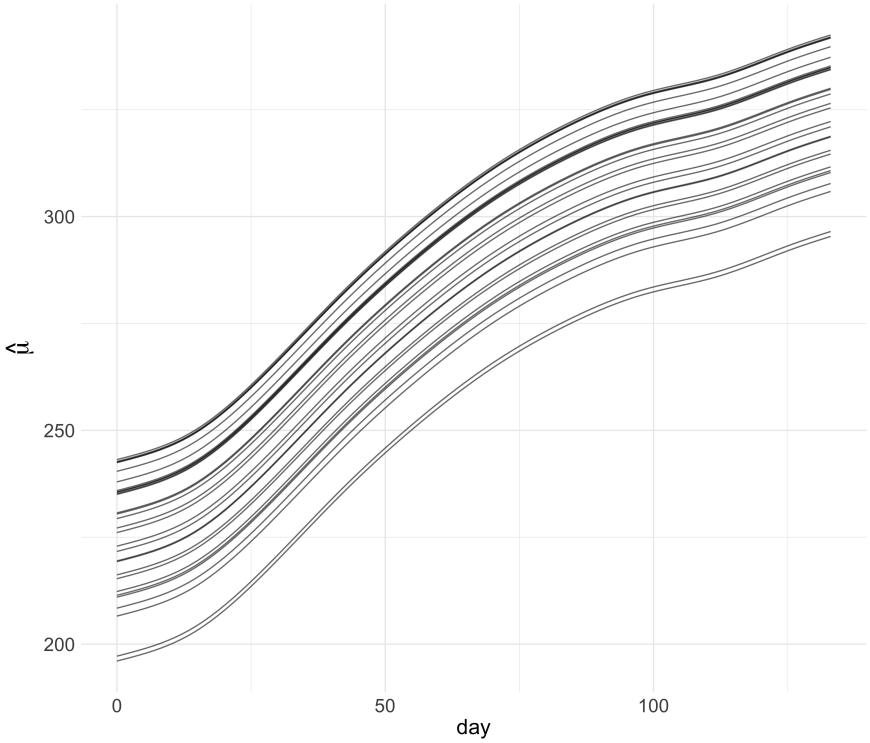


Figure 6.4: *Subject-specific fitted weight trajectories for cattle in treatment group A.*

Centering the data using the fitted mean, the residuals

$$\epsilon^*(t_{ij}) = y(t_{ij}) - (f(t_{ij}) + \alpha_i) \tag{6.4}$$

serve as the data for estimating the functions defining the Cholesky factor and innovation variances. We model

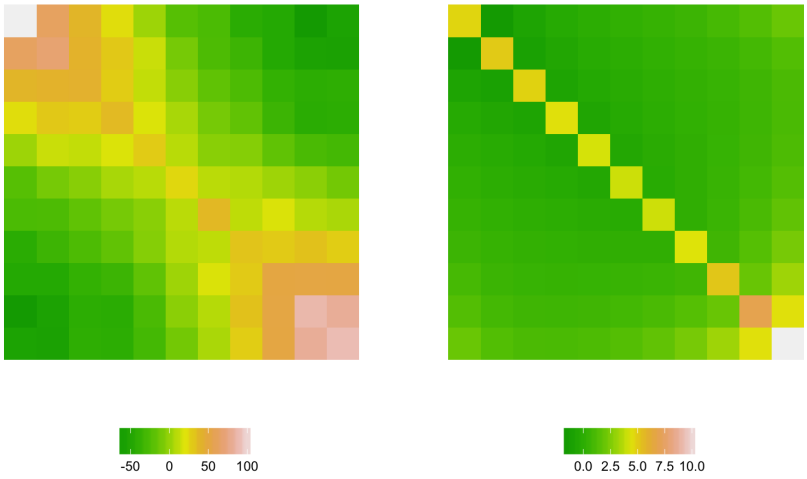
$$\epsilon^*(t_{ij}) = \sum_{k<j} \tilde{\phi}(t_{ij}, t_{ik})\, \epsilon^*(t_{ik}) + \epsilon(t_{ij}) \tag{6.5}$$

where ϵ is a mean zero Gaussian process with variance $\sigma^2(t)$.

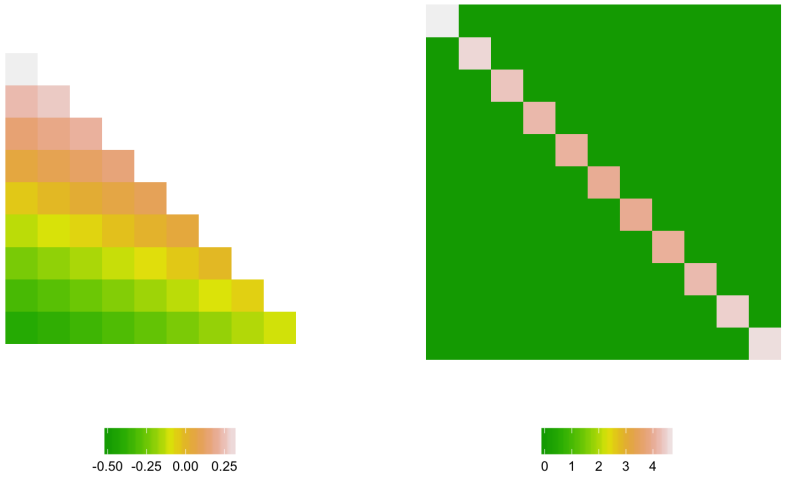
Choice of penalty is critical for convergence of the iterative estimation of $\tilde{\phi}$ and $\log(\sigma^2)$. Pan and Pan (2017) concluded that the regressogram of empirical estimates of $\tilde{\phi}_{t,s}$ show consistent behaviour over $l = t - s$ for each value of t, indicating a lack of a strong functional component of m. This is consistent Pourahmadi's choice in the specification of model (6.1) 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

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

This decomposition leads to a null space comprised of functions of l only, which is attractive because it coincides with the modeling assumptions made by Pan and Pan (2017), Huang et al. (2006), and Wu and Pourahmadi (2003) for the same data set. Figure 6.5 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 6.6 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 6.5: *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 $\hat{D}$.*

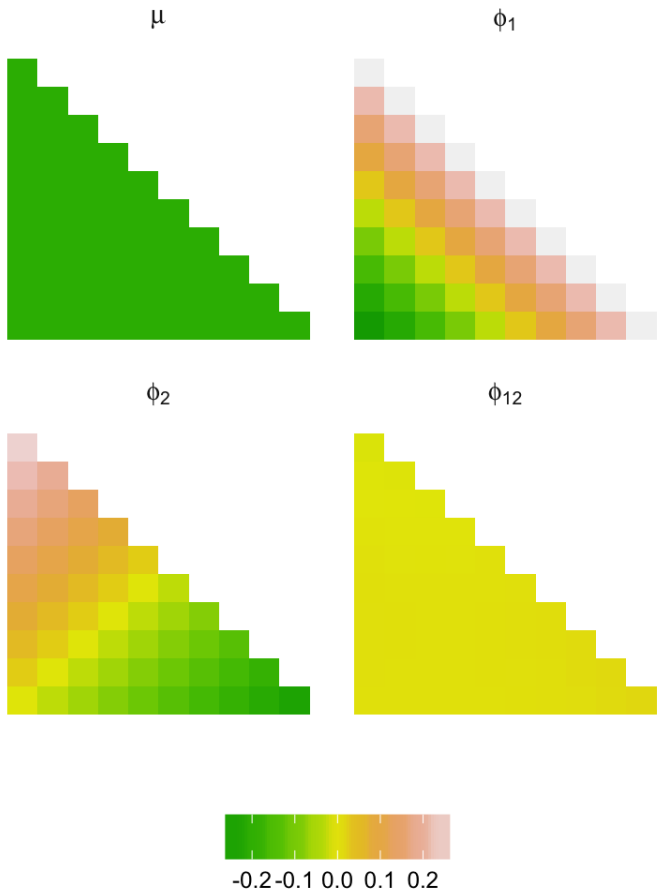


Figure 6.6: *Components of the SSANOVA decomposition of the estimated generalized autoregressive coefficient function ϕ evaluated on the grid defined by the observed time points.*

Our sole focus on covariance estimation rather than the joint estimation of the mean and covariance makes apples-to-apples comparison with other analyses of the same dataset difficult. We constructed the mean estimate for the cattle in treatment group A shown in Figure 6.4 entirely independently of the covariance estimate, which may be suboptimal compared to an iterative procedure that jointly estimates f, b, and Σ as in Pan and Pan (2017) and Pourahmadi (1999). Nevertheless, it is interesting to examine the differences between our estimates and cubic model fit shown in Figure 6.3. Modeling ϕ as a polynomial in l leaves any nonstationarity to be captured by the innovation variances. Of course, a model for the Cholesky factor having constant innovation variances

and generalized autoregressive parameters which vary in l only corresponds to a stationary process when certain conditions on the magnitude of the GARPs are satisfied (Klein, 1997; Madsen, 2007). Our estimated model instead captures the non-stationarity with both the log innovation variances as well as with ϕ_2, the functional component corresponding to the main effect of m. The size of the functional components (in terms of the squared functional norm), however, does indicate a certain degree of concordance with the model 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), and the squared norm of the interaction term, as clearly indicated by Figure 6.6, is negligible in comparison to the main effects.

Chapter 7: Concluding Remarks and Future Work

The previous discussion proposes a flexible framework for estimating the covariance matrix for longitudinal data. By modeling the Cholesky decomposition of the covariance matrix, we reframe covariance estimation as the estimation of a varying coefficient model, which allows for unconstrained estimation as well as a statistically intuitive interpretation of the elements of a covariance matrix. The varying coefficient model for the Cholesky decomposition naturally accommodates irregularly-spaced longitudinal data and allows varying within-subject sample sizes without requiring imputation of missing observations. The overall framework inherits the flexibility of the varying coefficient model, which allows us to leverage any of the tools classically used for nonparametric regression problems in the context of covariance estimation.

Estimation of the varying coefficient model is performed using bivariate smoothing using penalties which are motivated by the prevalent tendency to specify stationary models for the covariance matrix. Penalties enforce regularization of the fitted function so that under heavy penalization, the fitted components of the Cholesky factor correspond to covariance matrices which are close to stationary. We demonstrate the estimation procedure with two proposed representations of the varying coefficient function and the innovation variance function. A smoothing spline ANOVA model for the generalized autoregressive varying coefficient and the innovation variance function allow the fitted functions to be decomposed into their stationary and nonstationary functional components. We propose an alternative functional representation for $(\phi, \log \sigma^2)$ using tensor product

B-splines; smoothness is achieved by applying penalties to discrete differences of the vector of basis coefficients. The discrete penalties, which are constructed independently of the basis, offer flexibility over the smoothing spline penalties and require little computational complexity to implement.

The choice of basis is important when the unknown functions parameterizing the varying coefficient model are better represented by one or the other. Simulation studies reveal the advantages and disadvantages of our smoothing spline estimator and our P-spline estimator. The simulations illustrate the relative performance of both estimators compared to alternative estimators proposed in the longitudinal data literature.

We apply our method to data generated from a longitudinal experiment examining the effectiveness of two treatments for intestinal parasites in cattle as measured by subject body weight over time. For a single treatment group, our nonparametric estimator echoes some of the modeling assumptions made to specify parametric models in previous analysis of the same data.

Minimizing computational demand is an obvious motivator for future extensions of our work. The smoothing spline estimator circumvents the need for knot selection since it is constructed using a basis function for each of the unique within-subject pairs of measurement times. This is suitable when there is a fixed set of measurement times and unbalanced date arise due to missing observations. For the case that there is little overlap in measurement times across subjects so that these times are "nearly" unique for each subject, the size of the set $|V|$ can be so large that the dimension of the kernel matrix K_n as defined in (3.23) presents serious computational problems. An infinite dimensional Hilbert space is not necessarily required for representing the unknown function to be estimated, since the penalty effectively enforces a low dimensional model space. Efficient approximation can be carried by using a subset of the elements in V to represent the

unknown function. Algorithm 1 can directly accommodate such a low dimensional representation. Kim and Gu (2004) provide detailed discussion of the efficient approximation of $\mathcal{H}$.

The versatility of this framework leaves many paths open for further modeling exploration. In particular, an obvious example is the classification of observations into one of K groups by quadratic discriminant analysis. A lot of recent attention has been directed toward the problem of estimating separate $p \times p$ covariance matrices $\Sigma_1, \ldots, \Sigma_K$ for K separate groups, where the number of groups K and the dimension p are both potentially large. Often, there is not enough data to estimate separate Σ_i well for each group. For example, problems in financial management including portfolio selection can be reduced to the prediction of a sequence of large $p \times p$ covariance matrices (Tsay, 2005). Our procedure to covariance estimation encourages exploration of this problem; the regression model associated with the Cholesky decomposition (2.24) can incorporate additional group-specific covariates.

The construction of the penalty for the P-spline estimator is convenient and easily appended to the log likelihood. This allows us to easily use both shrinkage and smoothing for covariance estimation, and combining shrinkage and smoothing may produce better estimates than using shrinkage and smoothing alone. While adding additional penalty parameters can introduce added computational requirements, the connection between nonparametric regression modes and mixed models presents a way to mitigate this complexity. Smoothing parameters are interpreted as variances of random effects, so model estimation and smoothing parameter selection can be performed simultaneously using the stable and efficient algorithms and software that are available for mixed models. Restricted maximum likelihood (REML) has proven to be very useful as a model selection tool, often producing smoother fits than generalized cross validation due to its better resistance to over-fitting (Ruppert and Carroll, 2003).

Recently Eilers (1999) pointed how to interpret P-splines as a mixed model. Lee and Durbán (2011) proposed the use of P-splines within a mixed modeling framework to estimate multidimensional functions which can be decomposed into their functional components as with smoothing spline ANOVA models. This approach adds attractiveness to the interpretability of the models, and it allows for computationally convenient model fitting and selection. Application of the mixed model framework presented in Lee and Durbán (2011) to estimation of ϕ is attractive, because it not only provides an avenue for stable smoothing parameter selection, but it also permits the decomposition of the tensor product into functional components as in the SSANOVA model presented in Chapter 3. Direct application of this approach, however, is inaccessible due to the deconstruction of the marginal B-spline bases to adjust for the triangular domain of the autoregressive varying coefficient. Figure 4.6 illustrates how to "trim" the pairs of B-splines that don't overlap with the domain of ϕ, which lies on the triangle $0 < s < t < 1$. Trimming the basis inhibits identifiability of functional components, though in the case that an additive model is appropriate, this trimming is unnecessary and REML may be employed for model estimation.

Alternatively, bivariate B-splines inherit several of the appealing properties of univariate B-splines and are applicable in various modeling problems, particularly for those involving non-rectangular domains. They have been used extensively in the field of graphics for the construction of smooth surfaces over irregular domains, but thus have far received little attention in the field of statistics. However, a recent paper by Zhou and Pan (2014) employs a mixed effects model for the functional principal components as bivariate splines on triangulations for data observed on an irregular grid. The application of their ideas to covariance estimation presents a promising approach to estimation of the Cholesky decomposition via bivariate smoothing.

Appendix A: Chapter 2 Appendix

A.1 Proof of Theorem 3.2.1

Proof. The function space $\mathcal{H}$ is decomposed into $\mathcal{H}_0$ and $\mathcal{H}_1$. $\mathcal{H}_1$ 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

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

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 $\boldsymbol{v}_j \in V$.

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:

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

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 (3.18) depends only on $\sum_{i=1}^{\mathcal{N}_0} d_i \nu_i(\boldsymbol{v}) + \sum_{\boldsymbol{v}_j \in V} c_j K_1(\boldsymbol{v}_j, \boldsymbol{v})$. On the other hand, the penalty is given by

$$
\begin{aligned}
||P_1 \phi||^2 &= || \sum_{\boldsymbol{v}_j \in V} c_j K_1(\boldsymbol{v}_j, \cdot) + \rho(\cdot) ||_{\mathcal{H}}^2, \\
&= || \sum_{\boldsymbol{v}_j \in V} c_j K_1(\boldsymbol{v}_j, \cdot) ||_{\mathcal{H}}^2 + ||\rho(\cdot)||_{\mathcal{H}}^2.
\end{aligned}
$$

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 3.2.1.

□

Appendix B: Chapter 4 Appendix

B.1 Connecting the Finite Difference Penalty to B-spline Derivatives

The evaluation of the i^{th} B-spline using the recursive relation (4.1) can be derived from their definition as divided differences of truncated power functions.

Definition B.1.1. Let $t = \{t_i\}$ denote a non-decreasing sequence. The i^{th} B-spline of order k which corresponds to the knot sequence t is defined by

$$B_{i,k,t}(x) = (t_{i+k} - t_i)[t_i, \ldots, t_{i+k}](\cdot - x)_+^{k-1} \tag{B.1}$$

The placeholder notation, $(\cdot - x)_+^{k-1}$, is used to indicate that the k^{th} divided difference of the truncated power function $g(t) = (t-x)_+^{k-1}$ is obtained by fixing x and applying the divided difference to $g(t)$ as a function of t alone. Henceforth, we will write B_{ik} rather than $B_{i,k,t}$ when the knot sequence can be inferred from surrounding context.

The definition of B_i as a divided difference is necessary to bridge the expression for its derivative to the differences of its coefficients. The derivative of the truncated power function $g(x) = (t-x)_+^{k-1}$ is given by

$$\frac{\partial}{\partial x} g(x) = \frac{\partial}{\partial x}(t-x)_+^{k-1} = -(k-1)(t-x)_+^{k-2}.$$

Substituting B.1 into the recursive relation (4.1), we may write the derivative of the i^{th} B-spline of order k as follows:

$$\begin{aligned} B'_{i,k}(x) &= \Big[[t_{i+1}, \ldots, t_{i+k}] - [t_i, \ldots, t_{i+k-1}] \Big] \frac{\partial}{\partial x} (\cdot - x)_+^{k-1} \\ &= -(k-1) \Big[[t_{i+1}, \ldots, t_{i+k}] - [t_i, \ldots, t_{i+k-1}] \Big] (\cdot - x)_+^{k-2} \\ &= -(k-1) \left[-\frac{B_{i+1,k-1}(x)}{(t_{i+k} - t_{i+1})} + \frac{B_{i,k-1}(x)}{(t_{i+k-1} - t_i)} \right] \end{aligned}$$

This allows us to write

$$\begin{aligned} \frac{\partial}{\partial x} \left[\sum_i \theta_i B_i \right] &= \sum_i \theta_i B'_{i,k} \\ &= \sum_i (k-1) \frac{\theta_i - \theta_{i-1}}{t_{i+k-1} - t_i} B_{i,k-1}. \end{aligned} \tag{B.2}$$

Note that the limits on the previous summation in B.2 are left unspecified; the formula is written for bi-infinite sums, and their application to finite sums is accessible after they are written formally as bi-infinite sums by augmenting the appropriate zero terms. However, if we are interested in a particular interval over the domain, say $[t_r, t_s]$, then for $x \in [t_r, t_s]$, then

$$\frac{\partial}{\partial x} \left[\sum_i \theta_i B_{i,k}(x) \right] = \sum_{r-k+2}^{s-1} (k-1) \frac{\theta_i - \theta_{i-1}}{t_{i+k-1} - t_i} B_{i,k-1}(x)$$

since $B_{i,k-1}(x) = 0$ for all $i \notin \{r-k+2, \ldots, s-1\}$ when $t_r \leq x \leq t_s$. Applying B.2 j times gives us that the j^{th} derivative of $f = \sum_i \theta_i B_{ik}$ has form

$$\frac{\partial^j}{\partial x^j} \left[\sum_i \theta_i B_{i,k}(x) \right] = \sum_i \theta_i^{(j+1)} B_{i,k-j} \tag{B.3}$$

$$\theta_i^{(j+1)} \equiv \begin{cases} \theta_i, & j = 0 \\ \frac{\theta_i^{(j)} - \theta_{i-1}^{(j)}}{\left(t_{i+k-j} - t_i\right)/(k-j)}, & j \geq 1 \end{cases} \tag{B.4}$$

Proof. We proceed by induction on j. We have already shown the case for $j = 1$ in the derivation of B.2. Assume that the statement holds for some $j^* > 1$, so that we have

$$\frac{\partial^{j^*}}{\partial x^{j^*}}\left[\sum_i \theta_i B_{i,k}(x)\right] = \sum_i \frac{\theta_i^{(j^*)} - \theta_{i-1}^{(j^*)}}{\left(t_{i+k-j^*} - t_i\right)/\left(k - j^*\right)} B_{i,k-j^*}(x).$$

Then the $(j^*+1)^{st}$ derivative is given by

$$\begin{aligned}
\frac{\partial^{j^*+1}}{\partial x^{j^*+1}}\left[\sum_i \theta_i B_{i,k}\right] &= \sum_i \frac{\theta_i^{(j^*)} - \theta_{i-1}^{(j^*)}}{\left(t_{i+k-j^*} - t_i\right)/\left(k - j^*\right)} B'_{i,k-j^*} \\
&= \sum_i \theta_i^{(j^*)} B'_{i,k-j^*} \\
&= \sum_i \theta_i^{(j^*)} \left(k - (j^*+1)\right)\left[\frac{B_{i,k-(j^*+1)}}{t_{i+k-(j^*+1)} - t_i} - \frac{B_{i+1,k-(j^*+1)}}{t_{i+k-(j^*+1)+1} - t_{i+1}}\right] \\
&= \sum_i \frac{\theta_i^{(j^*)} - \theta_{i-1}^{(j^*)}}{\left(t_{i+k-(j^*+1)} - t_i\right)/\left(k - (j^*+1)\right)} B_{i,k-(j^*+1)} \\
&= \sum_i \theta_i^{(j^*+1)} B_{i,k-(j^*+1)}
\end{aligned}$$

□

The choice to write $k - j$ as a divisor in the denominator lends to the interpretation of B.3 as a difference quotient, with the quantity

$$\frac{t_{i+k-j} - t_i}{k - j}$$

representing a mean mesh length of sorts on the interval $[t_i, t_{i+k-j}]$. We note that the case where t contains replicated knots leads to division by zero. This is, however, a trivial situation, since for $t_i = t_{i+k-j}$, we have $B_i = 0$, and we take $\frac{0}{0} = 0$.

Appendix C: Chapter 5 Appendix

C.1 Quadratic Risk Estimates for Simulation with Complete Data

Table C.1: *Multivariate normal simulations for model I. Estimated quadratic risk is reported for our smoothing spline ANOVA estimator and P-spline estimator, the oracle estimator for each covariance structure, the parametric polynomial estimator of Pan and MacKenzie (2003), the sample covariance matrix, the tapered sample covariance matrix, and the soft thresholding estimator.*

	p	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.00267	0.0016	0.0052	0.0912	0.3901	0.3864	0.3874
	20	0.00459	0.0010	0.0043	0.0757	0.8371	0.7710	0.7716
	30	0.00386	0.0026	0.0036	0.1109	1.2857	1.1937	1.2074
$N = 100$	10	0.00209	0.0005	0.0010	0.0426	0.2116	0.1676	0.1720
	20	0.00212	0.0003	0.0011	0.0376	0.4255	0.3902	0.3970
	30	0.00276	0.0002	0.0011	0.0313	0.5984	0.5790	0.5842

Table C.2: *Multivariate normal simulation-estimated quadratic risk for model II.*

	p	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.0483	0.0623	0.0792	7.0137	0.6269	0.8108	0.5770
	20	0.4317	0.7972	1.2456	852.2787	2.7659	30.8197	36.1492
	30	6.7921	12.8700	7.2129	4849.8925	21.0228	365.0301	1804.9695
$N = 100$	10	0.0280	0.0254	0.0525	7.0482	0.2683	0.4351	0.2665
	20	0.2625	0.2877	0.8153	861.3937	1.3347	5.5170	7.3283
	30	2.6619	2.7399	6.9793	5075.4782	8.4769	66.9461	420.2973

Table C.3: *Multivariate normal simulation-estimated quadratic risk for model III.*

	p	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.0697	0.0656	0.0665	3.4849	0.4977	0.6678	0.5858
	20	0.4706	1.0095	0.9146	426.0848	2.0716	4.8213	8.4099
	30	5.3699	10.8782	8.1124	5061.3563	16.5536	779.2829	1181.3770
$N = 100$	10	0.0328	0.0486	0.0363	3.5437	0.2437	0.2929	0.2791
	20	0.1958	0.6260	0.3783	416.1285	1.0193	1.5353	5.1553
	30	2.2121	5.9367	3.4576	5082.1367	7.9582	14.2394	253.4296

Table C.4: *Multivariate normal simulation-estimated quadratic risk for model IV.*

	p	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.0053	0.0144	0.0196	0.2575	0.4420	0.4628	0.4620
	20	0.0073	0.0449	0.0154	0.4384	0.7951	0.9184	0.9177
	30	0.0072	0.0893	0.0189	0.6539	1.3363	1.3014	1.3013
$N = 100$	10	0.0031	0.0112	0.0186	0.2098	0.2136	0.2299	0.2295
	20	0.0027	0.0420	0.0143	0.4877	0.4509	0.4311	0.4307
	30	0.0035	0.0792	0.0181	0.6616	0.6263	0.6598	0.6589

Table C.5: *Multivariate normal simulation-estimated quadratic risk for model V.*

N	p	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10 0.1610		0.3621	0.2456	1.3738	0.8484	1.6174	0.8963
	20	0.5236	0.9911	0.8206	2.8419	1.7324	3.0233	1.6375
	30	0.4632	1.5352	1.1507	4.1877	2.5484	5.1546	2.6727
$N = 100$	10	0.0813	0.3091	0.2678	1.2439	0.4175	1.0431	0.4922
	20	0.1522	0.9734	0.4111	2.7280	0.7896	2.1932	0.8461
	30	0.3656	1.6032	0.7701	3.8905	1.2577	3.5722	1.3270

C.2 Quadratic Risk Estimates for Simulation with Irregularly Sampled Data

Table C.6: *Model 1: Quadratic risk estimates and corresponding standard errors for the MCD smoothing spline ANOVA estimator via 100 simulated multivariate normal samples of size $N = 50$ when 0%, 10%, 20%, and 30% of the data are missing for each subject. Risk is reported for the estimator constructed using the unbiased risk estimate and leave-one-subject-out cross validation for smoothing parameter selection.*

p	% missing	$\Delta_1(\hat{\Sigma}^U_{SS})$		$\Delta_1(\hat{\Sigma}^{V^*}_{SS})$	
10	0.0	0.001625283	(3e-040)	0.00242142	(5e-040)
	0.1	0.002667487	(4e-040)	0.00340902	(6e-040)
	0.2	0.002203362	(4e-040)	0.00481581	(7e-040)
	0.3	0.005959094	(9e-040)	0.00791520	(0.0016)
20	0.0	0.000865565	(1e-040)	0.00265909	(0.0018)
	0.1	0.001350105	(2e-040)	0.24942590	(0.2471)
	0.2	0.002791360	(3e-040)	0.01027696	(0.0032)
	0.3	0.004419142	(6e-040)	0.09231505	(0.0516)

Table C.7: *Model 2: Quadratic risk estimates and corresponding standard errors.*

p	% missing	$\Delta_1(\hat{\Sigma}^U_{SS})$		$\Delta_1(\hat{\Sigma}^{V^*}_{SS})$	
10	0.0	0.0450916	(0.0082)	0.0601659	(0.0096)
	0.1	0.0696728	(0.0100)	0.1512636	(0.0289)
	0.2	0.2300287	(0.0335)	0.2343197	(0.0398)
	0.3	0.4409229	(0.0661)	0.6346628	(0.1247)
20	0.0	0.4590734	(0.0705)	0.6819051	(0.1176)
	0.1	19.4089837	(2.0563)	20.8552036	(1.5583)
	0.2	268.9477374	(20.7521)	3969.3959755	(3513.7089)
	0.3	2437.4762290	(305.7227)	5001.5651163	(603.1301)

Table C.8: *Model 3: Quadratic risk estimates and corresponding standard errors.*

p	% missing	$\Delta_1(\hat{\Sigma}^U_{SS})$		$\Delta_1(\hat{\Sigma}^{V^*}_{SS})$	
10	0.0	0.0650014	(0.0055)	0.0682312	(0.0059)
	0.1	0.0770316	(0.0081)	0.0892940	(0.0118)
	0.2	0.1140654	(0.0142)	0.2008099	(0.0280)
	0.3	0.3315869	(0.0677)	0.3268610	(0.0495)
20	0.0	1.0422739	(0.1994)	1.2132111	(0.2173)
	0.1	11.9788732	(1.7077)	18.5305750	(1.5563)
	0.2	232.1002465	(23.7789)	280.9434501	(42.1525)
	0.3	1667.1547183	(263.3001)	2601.3353420	(338.6449)

Table C.9: *Model 4: Quadratic risk estimates and corresponding standard errors.*

p	% missing	$\Delta_1(\hat{\Sigma}^U_{SS})$		$\Delta_1(\hat{\Sigma}^{V^*}_{SS})$	
10	0.0	0.01436606	(7e-040)	0.01655013	(0.0013)
	0.1	0.01684656	(8e-040)	0.01893500	(0.0022)
	0.2	0.02374962	(0.0023)	0.02433408	(0.0020)
	0.3	0.03204756	(0.0028)	0.03424552	(0.0044)
20	0.0	0.04488566	(9e-040)	0.04670697	(9e-040)
	0.1	0.04654451	(0.0012)	0.05029391	(0.0015)
	0.2	0.05132972	(0.0013)	0.06053346	(0.0038)
	0.3	0.06230931	(0.0021)	0.10699654	(0.0459)

Table C.10: *Model 5: Quadratic risk estimates and corresponding standard errors.*

p	% missing	$\Delta_1(\hat{\Sigma}_{SS}^{U})$		$\Delta_1(\hat{\Sigma}_{SS}^{V^*})$	
10	0.0	0.3621065	(0.0091)	0.3623509	(0.0128)
	0.1	0.6778957	(0.0457)	0.7067101	(0.0426)
	0.2	2.1262957	(0.1590)	2.4381408	(0.2292)
	0.3	6.8051314	(0.6256)	8.2414439	(0.7087)
20	0.0	0.9910795	(0.0138)	1.0334928	(0.0099)
	0.1	1.7214964	(0.1028)	1.5051130	(0.0577)
	0.2	5.3527162	(0.3290)	5.1871496	(0.3852)
	0.3	29.6617541	(2.0158)	25.1766132	(1.8094)

C.3 Comprehensive Tables for Simulations with Complete Data

Table C.11: *Multivariate normal simulations for model V. Estimated entropy risk and standard errors of the loss are reported for our smoothing spline ANOVA estimator and P-spline estimator, the oracle estimator for each covariance structure, the parametric polynomial estimator of Pan and MacKenzie (2003), the sample covariance matrix, the tapered sample covariance matrix, and the soft thresholding estimator.*

Model	N	p	$\hat{\Sigma}^{ure}_{SS}$		$\hat{\Sigma}^{ure}_{PS}$		$\hat{\Sigma}_{oracle}$		$\hat{\Sigma}_{poly}$		S		S^{ω}		S^{λ}	
I	50	10	0.0685	(0.0072)	0.1261	(0.0107)	0.0135	(0.0023)	0.1102	(0.0083)	1.2047	(0.0286)	0.5369	(0.0563)	1.1742	(0.0366)
	50	20	0.0834	(0.0081)	0.1713	(0.0095)	0.0229	(0.0041)	0.1096	(0.0087)	4.9850	(0.0644)	1.3957	(0.1859)	4.7796	(0.1206)
	50	30	0.1102	(0.0229)	0.1969	(0.0118)	0.0196	(0.0034)	0.1127	(0.0108)	12.5517	(0.1322)	2.8019	(0.4332)	11.3175	(0.3556)
	100	10	0.0451	(0.0035)	0.0671	(0.0042)	0.0105	(0.0015)	0.0531	(0.0038)	0.5685	(0.0151)	0.2045	(0.0235)	0.5236	(0.0176)
	100	20	0.0425	(0.0062)	0.0965	(0.0048)	0.0105	(0.0020)	0.0512	(0.0031)	2.2831	(0.0285)	0.5724	(0.0744)	2.1358	(0.0606)
	100	30	0.0431	(0.0044)	0.1148	(0.0062)	0.0139	(0.0021)	0.0472	(0.0033)	5.2770	(0.0472)	1.2430	(0.1569)	4.9126	(0.1204)
II	50	10	0.0689	(0.0069)	0.3423	(0.0082)	0.0581	(0.0055)	4.7673	(0.0919)	1.2832	(0.0334)	1.4644	(0.0475)	1.1770	(0.0346)
	50	20	0.0581	(0.0080)	1.3640	(0.0158)	0.0439	(0.0051)	97.2334	(2.4537)	5.1665	(0.0610)	21.6407	(1.2914)	39.3522	(8.1602)
	50	30	0.0811	(0.0075)	2.6485	(0.0472)	0.0627	(0.0063)	153.9665	(7.9453)	12.3582	(0.1070)	55.3674	(3.8362)	133.9980	(19.2003)
	100	10	0.0457	(0.0050)	0.2945	(0.0059)	0.0386	(0.0034)	4.7911	(0.0638)	0.5812	(0.0134)	0.8335	(0.0293)	0.5628	(0.0154)
	100	20	0.0416	(0.0038)	1.2875	(0.0100)	0.0269	(0.0027)	98.1989	(2.0835)	2.3364	(0.0316)	10.1841	(0.8276)	10.0864	(1.1183)
	100	30	0.0367	(0.0033)	2.4365	(0.0293)	0.0288	(0.0031)	158.2480	(7.2097)	5.2389	(0.0475)	33.5207	(0.9390)	62.5030	(14.7791)
III	50	10	0.3296	(0.0091)	0.1065	(0.0090)	0.0619	(0.0079)	3.0108	(0.0709)	1.2030	(0.0312)	1.1460	(0.0472)	1.1467	(0.0341)
	50	20	1.1100	(0.0100)	0.2555	(0.0109)	0.0695	(0.0075)	62.7522	(2.1710)	4.9824	(0.0689)	17.2244	(0.6234)	14.9189	(2.7042)
	50	30	2.3215	(0.0132)	0.6242	(0.0390)	0.0576	(0.0071)	218.2387	(31.2219)	12.4792	(0.1182)	49.9135	(7.7026)	121.7795	(18.3978)
	100	10	0.2904	(0.0045)	0.0579	(0.0050)	0.0268	(0.0027)	3.0383	(0.0559)	0.5699	(0.0142)	0.5545	(0.0162)	0.5371	(0.0130)
	100	20	1.1963	(0.1239)	0.2011	(0.0057)	0.0275	(0.0036)	62.8960	(1.1460)	2.2700	(0.0306)	11.8274	(0.7008)	9.5217	(1.0164)
	100	30	2.2811	(0.0079)	0.3845	(0.0169)	0.0221	(0.0024)	221.0090	(21.8998)	5.2234	(0.0462)	29.1693	(0.6585)	60.3529	(14.2471)
IV	50	10	0.3348	(0.0085)	0.1966	(0.0118)	0.0217	(0.0049)	0.7144	(0.0141)	1.2218	(0.0319)	0.7397	(0.0436)	1.1921	(0.0317)
	50	20	0.9177	(0.0054)	0.3499	(0.0174)	0.0286	(0.0046)	1.4588	(0.0179)	4.9091	(0.0676)	1.9786	(0.1650)	4.9206	(0.0612)
	50	30	1.5992	(0.0154)	0.5100	(0.0152)	0.0283	(0.0044)	2.2173	(0.0238)	12.6114	(0.1179)	3.7440	(0.3991)	12.1489	(0.1908)
	100	10	0.3047	(0.0047)	0.2237	(0.0125)	0.0125	(0.0025)	0.6958	(0.0080)	0.5570	(0.0130)	0.3168	(0.0142)	0.5515	(0.0147)
	100	20	0.8911	(0.0036)	0.3704	(0.0185)	0.0105	(0.0017)	1.4813	(0.0140)	2.2659	(0.0305)	0.9365	(0.0686)	2.2474	(0.0334)
	100	30	1.5213	(0.0029)	0.5282	(0.0163)	0.0134	(0.0022)	2.2228	(0.0141)	5.2106	(0.0473)	1.9312	(0.1746)	5.2111	(0.0584)
V	50	10	0.2769	(0.0068)	0.2464	(0.0108)	0.0986	(0.0200)	1.2420	(0.0294)	1.2023	(0.0318)	18.5222	(0.6731)	2.9824	(0.3820)
	50	20	0.7514	(0.0042)	0.8772	(0.0128)	0.2512	(0.0580)	2.8557	(0.0646)	5.0195	(0.0695)	34.6618	(0.6202)	13.8690	(0.8916)
	50	30	1.1776	(0.0051)	0.9791	(0.0125)	0.2641	(0.0474)	4.5791	(0.0914)	12.3460	(0.1112)	46.5437	(0.7836)	26.1364	(0.3248)
	100	10	0.2416	(0.0039)	0.1722	(0.0049)	0.0520	(0.0090)	1.1491	(0.0202)	0.5821	(0.0111)	16.4081	(0.4280)	1.7397	(0.0363)
	100	20	0.7286	(0.0028)	0.2965	(0.0046)	0.0827	(0.0170)	2.9080	(0.0383)	2.2918	(0.0244)	32.5295	(0.5786)	5.4649	(0.5497)
	100	30	1.1813	(0.0051)	0.4291	(0.0065)	0.1799	(0.0420)	4.4402	(0.0655)	5.2197	(0.0465)	39.2914	(0.2195)	15.4295	(0.8464)

Table C.12: *Multivariate normal simulations for model V. Estimated quadratic risk and standard errors of the loss are reported for our smoothing spline ANOVA estimator and P-spline estimator, the oracle estimator for each covariance structure, the parametric polynomial estimator of Pan and MacKenzie (2003), the sample covariance matrix, the tapered sample covariance matrix, and the soft thresholding estimator.*

Model	N	p	$\hat{\Sigma}^{ure}_{SS}$		$\hat{\Sigma}^{ure}_{PS}$		$\hat{\Sigma}_{oracle}$		$\hat{\Sigma}_{poly}$		S		S^{ω}		S^{λ}	
I	50	10	0.0016	(3e-040)	0.0052	(0.0010)	0.0267	(0.0045)	0.0912	(0.0103)	0.3901	(0.0247)	0.3864	(0.0221)	0.3874	(0.0224)
	50	20	0.0010	(2e-040)	0.0043	(6e-040)	0.0459	(0.0083)	0.0757	(0.0098)	0.8371	(0.0325)	0.7710	(0.0392)	0.7716	(0.0386)
	50	30	0.0026	(0.0018)	0.0036	(6e-040)	0.0386	(0.0065)	0.1109	(0.0152)	1.2857	(0.0498)	1.1937	(0.0472)	1.2074	(0.0472)
	100	10	0.0005	(1e-040)	0.0010	(1e-040)	0.0209	(0.0031)	0.0426	(0.0051)	0.2116	(0.0124)	0.1676	(0.0090)	0.1720	(0.0099)
	100	20	0.0003	(1e-040)	0.0011	(1e-040)	0.0212	(0.0042)	0.0376	(0.0042)	0.4255	(0.0161)	0.3902	(0.0164)	0.3970	(0.0170)
	100	30	0.0002	(1e-040)	0.0011	(1e-040)	0.0276	(0.0041)	0.0313	(0.0033)	0.5984	(0.0262)	0.5790	(0.0211)	0.5842	(0.0208)
II	50	10	0.0451	(0.0070)	0.0623	(0.0043)	0.0792	(0.0083)	7.0137	(0.3452)	0.6269	(0.0363)	0.8108	(0.0690)	0.5770	(0.0377)
	50	20	0.4591	(0.1388)	1.2456	(0.1778)	0.4317	(0.0809)	852.2787	(38.4308)	2.7659	(0.2037)	30.8197	(15.7299)	36.1492	(9.3235)
	50	30	6.7921	(1.5850)	12.8700	(1.4200)	7.2129	(1.2710)	4849.8925	(901.174)	21.0228	(2.2821)	365.0301	(178.7437)	1804.9695	(435.1357)
	100	10	0.0254	(0.0044)	0.0525	(0.0033)	0.0580	(0.0071)	7.0482	(0.2405)	0.2683	(0.0164)	0.4351	(0.0279)	0.2665	(0.0166)
	100	20	0.2877	(0.0477)	0.8153	(0.1501)	0.2625	(0.0377)	861.3937	(34.1825)	1.3347	(0.1086)	5.5170	(0.6241)	7.3283	(1.4927)
	100	30	2.7399	(0.4745)	6.9793	(0.9114)	3.6619	(0.7715)	5075.4782	(908.7174)	8.4769	(0.7058)	66.9461	(6.0353)	420.2973	(119.1735)
III	50	10	0.0650	(0.0053)	0.0665	(0.0033)	0.0697	(0.0102)	3.4849	(0.2297)	0.4977	(0.0265)	0.6678	(0.0645)	0.5858	(0.0365)
	50	20	1.0423	(0.1420)	0.9146	(0.1113)	0.4706	(0.0731)	426.0848	(26.4453)	2.0716	(0.1360)	4.8213	(1.1130)	8.4099	(1.3497)
	50	30	10.8782	(1.1771)	8.1124	(1.2342)	5.3699	(0.8475)	5061.3563	(572.4879)	16.5536	(1.8098)	779.2829	(714.9847)	1181.3770	(327.7712)
	100	10	0.0486	(0.0040)	0.0363	(0.0047)	0.0328	(0.0040)	3.5437	(0.1839)	0.2437	(0.0130)	0.2929	(0.0196)	0.2791	(0.0170)
	100	20	0.6260	(0.0200)	0.3783	(0.0823)	0.1958	(0.0308)	416.1285	(12.8666)	1.0193	(0.0701)	1.5353	(0.1560)	5.1553	(1.0771)
	100	30	5.9367	(0.7791)	3.4576	(0.7345)	2.2121	(0.3658)	5082.1367	(377.1631)	7.9582	(0.8381)	14.2394	(1.7202)	253.4296	(75.1683)
IV	50	10	0.0144	(0.0010)	0.0196	(0.0039)	0.0053	(0.0012)	0.2575	(0.0340)	0.4420	(0.0293)	0.4628	(0.0365)	0.4620	(0.0363)
	50	20	0.0449	(6e-040)	0.0154	(0.0024)	0.0073	(0.0012)	0.4384	(0.0416)	0.7951	(0.0447)	0.9184	(0.0397)	0.9177	(0.0395)
	50	30	0.0893	(0.0022)	0.0189	(0.0030)	0.0072	(0.0011)	0.6539	(0.0557)	1.3363	(0.0485)	1.3014	(0.0462)	1.3013	(0.0453)
	100	10	0.0112	(5e-040)	0.0186	(0.0029)	0.0031	(6e-040)	0.2098	(0.0185)	0.2136	(0.0109)	0.2299	(0.0134)	0.2295	(0.0133)
	100	20	0.0420	(4e-040)	0.0143	(0.0014)	0.0027	(4e-040)	0.4877	(0.0325)	0.4509	(0.0167)	0.4311	(0.0159)	0.4307	(0.0158)
	100	30	0.0792	(4e-040)	0.0181	(0.0020)	0.0035	(6e-040)	0.6616	(0.0327)	0.6263	(0.0215)	0.6598	(0.0207)	0.6589	(0.0207)
V	50	10	0.3621	(0.0123)	0.2456	(0.0206)	0.1610	(0.0332)	1.3738	(0.0999)	0.8484	(0.0549)	1.6174	(0.1133)	0.8963	(0.0554)
	50	20	0.9911	(0.0102)	0.8206	(0.0213)	0.5236	(0.1373)	2.8419	(0.1751)	1.7324	(0.0802)	3.0233	(0.1872)	1.6375	(0.0889)
	50	30	1.5352	(0.0088)	1.1507	(0.0176)	0.4632	(0.0755)	4.1877	(0.2390)	2.5484	(0.0975)	5.1546	(0.3173)	2.6727	(0.1067)
	100	10	0.3091	(0.0047)	0.2678	(0.0112)	0.0813	(0.0133)	1.2439	(0.0664)	0.4175	(0.0258)	1.0431	(0.0556)	0.4922	(0.0273)
	100	20	0.9734	(0.0075)	0.4111	(0.0084)	0.1522	(0.0331)	2.7280	(0.1010)	0.7896	(0.0306)	2.1932	(0.0929)	0.8461	(0.0355)
	100	30	1.6032	(0.0088)	0.7701	(0.0098)	0.3656	(0.0968)	3.8905	(0.1447)	1.2577	(0.0466)	3.5722	(0.1457)	1.3270	(0.0411)

Bibliography

Anderson, T. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *The Annals of Statistics*, 135–141.

Anderson, T. W. (Ed.) (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. and C. Thomas-Agnan (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

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

Bickel, P. J., E. Levina, et al. (2008). Covariance regularization by thresholding. *The Annals of Statistics 36*(6), 2577–2604.

Boente, G. and R. Fraiman (2000). Kernel-based functional principal components. *Statistics & probability letters 48*(4), 335–345.

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

Carroll, R. J. and D. Ruppert (1988). *Transformation and weighting in regression*, Volume 30. CRC Press.

Champion, C. J. (2003). Empirical bayesian estimation of normal variances and covariances. *Journal of multivariate analysis 87*(1), 60–79.

Cheng, S. H. and N. J. Higham (1998). A modified cholesky algorithm based on a symmetric indefinite factorization. *SIAM Journal on Matrix Analysis and Applications 19*(4), 1097–1110.

Chiang, C.-T., J. A. Rice, and C. O. Wu (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association 96*(454), 605–619.

Chiu, T. Y., T. Leonard, and K.-W. Tsui (1996). The matrix-logarithmic covariance model. *Journal of the American Statistical Association 91*(433), 198–210.

Cox, D., E. Koh, G. Wahba, and B. S. Yandell (1988). Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *The Annals of Statistics*, 113–119.

Dahlhaus, R. et al. (1997). Fitting time series models to nonstationary processes. *The annals of Statistics 25*(1), 1–37.

Dahmen, W., C. A. Micchelli, and H.-P. Seidel (1992). Blossoming begets b-spline bases built better by b-patches. *Mathematics of computation 59*(199), 97–115.

Daniels, M. J. and R. E. Kass (1999). Nonconjugate bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association 94*(448), 1254–1263.

De Boor, C., C. De Boor, E.-U. Mathématicien, C. De Boor, and C. De Boor (1978). *A practical guide to splines*, Volume 27. Springer-Verlag New York.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, 157–175.

Dennis Jr, J. E. and R. B. Schnabel (1996). *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM.

Dey, D. K., S. K. Ghosh, and B. K. Mallick (2000). *Generalized linear models: A Bayesian perspective*. CRC Press.

Dierckx, P. (1995). *Curve and surface fitting with splines*. Oxford University Press.

Eilers, P. (1999). Discussion of "the analysis of designed experiments and longitudinal data by using smoothing splines" by verbyla et al. *Appl. Statist 48*, 306–307.

Eilers, P. H., I. D. Currie, and M. Durbán (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis 50*(1), 61–76.

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

Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics 20*(3), 339–350.

Fan, J. and W. Zhang (1999). Statistical estimation in varying coefficient models. *Annals of Statistics*, 1491–1518.

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

Friedman, J. H. and B. W. Silverman (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics 31*(1), 3–21.

Gabriel, K. (1962). Ante-dependence analysis of an ordered set of variables. *The Annals of Mathematical Statistics*, 201–212.

Gill, P. E., W. Murray, and M. H. Wright (1981). Practical optimization.

Golub, G. H. and C. F. Van Loan (2012). *Matrix computations*, Volume 3. JHU Press.

Gu, C. (2002). Smoothing spline anova models.

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

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

Haff, L. (1980). Empirical bayes estimation of the multivariate normal covariance matrix. *The Annals of Statistics*, 586–597.

Hastie, T. and R. Tibshirani (1990). *Generalized additive models*. Wiley Online Library.

Hoover, D. R., J. A. Rice, C. O. Wu, and L.-P. Yang (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika 85*(4), 809–822.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology 24*(6), 417.

Huang, J. Z., M. Chen, M. Maadooliat, and M. Pourahmadi (2012). A cautionary note on generalized linear models for covariance of unbalanced longitudinal data. *Journal of Statistical Planning and Inference 142*(3), 743–751.

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

Jennrich, R. I. and M. D. Schluchter (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 805–820.

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.

Kim, Y.-J. and C. Gu (2004). Smoothing spline gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(2), 337–356.

Kimeldorf, G. and G. Wahba (1971). Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications 33*(1), 82–95.

Kitagawa, G. and W. Gersch (1985). A smoothness priors time-varying ar coefficient modeling of nonstationary covariance time series. *IEEE Transactions on Automatic Control 30*(1), 48–56.

Klein, J. L. (1997). *Statistical visions in time: a history of time series analysis, 1662-1938*. Cambridge University Press.

Kooperberg, C. and C. J. Stone (1991). A study of logspline density estimation. *Computational Statistics & Data Analysis 12*(3), 327–347.

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

Lee, D.-J. and M. Durbán (2011). P-spline anova-type interaction models for spatio-temporal smoothing. *Statistical Modelling 11*(1), 49–69.

Levina, E., A. Rothman, and J. Zhu (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. and Y. Wang (2004). Hypothesis testing in smoothing spline models. *Journal of Statistical computation and simulation 74*(8), 581–597.

Madsen, H. (2007). *Time series analysis*. CRC Press.

Marx, B. D. and P. H. Eilers (1999). Generalized linear regression on sampled signals and curves: a p-spline approach. *Technometrics 41*(1), 1–13.

McCullagh, P. and J. Nelder (1989). *Generalized linear models* (2nd ed.). London: Chapman and Hall.

McCulloch, C. E. and J. M. Neuhaus (2001). *Generalized linear mixed models*. Wiley Online Library.

Noh, H. S. and B. U. Park (2010). Sparse varying coefficient models for longitudinal data. *Statistica Sinica*, 1183–1202.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical science*, 502–518.

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

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

Pinheiro, J. C. and D. M. Bates (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and computing 6*(3), 289–296.

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

Ramsay, J. O. (2006). *Functional data analysis*. Wiley Online Library.

Ramsay, J. O. and B. W. Silverman (2007). *Applied functional data analysis: methods and case studies*. Springer.

Rice, J. A. and B. W. Silverman (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, 233–243.

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

Ruppert, D., W. M. and R. J. Carroll (2003). *Semiparametric regression*. Cambridge University Press Cambridge:.

Seidel, H.-P. (1991). Symmetric recursive algorithms for surfaces: B-patches and the de boor algorithm for polynomials over triangles. *Constr. Approx 7*, 257–279.

Şentürk, D., L. S. Dalrymple, S. M. Mohammed, G. A. Kaysen, and D. V. Nguyen (2013). Modeling time-varying effects with generalized and unsynchronized longitudinal data. *Statistics in medicine 32*(17), 2971–2987.

Şentürk, D. and H.-G. Müller (2008). Generalized varying coefficient models for longitudinal data. *Biometrika 95*(3), 653–666.

Smith, M. and R. Kohn (2002). Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association 97*(460), 1141–1153.

Stein, C. (1975). Estimation of a covariance matrix, rietz lecture. In *39th Annual Meeting IMS, Atlanta, GA, 1975*.

Tsay, R. S. (2005). *Analysis of financial time series*, Volume 543. John Wiley & Sons.

Verbyla, A. P. (1993). Modelling variance heterogeneity: residual maximum likelihood and diagnostics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 493–508.

Wahba, G. (1990). *Spline models for observational data*, Volume 59. Siam.

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

Wand, M. and J. Ormerod (2008). On semiparametric regression with o'sullivan penalized splines. *Australian & New Zealand Journal of Statistics 50*(2), 179–198.

Wang, B. (2014). *CVTuningCov: Regularized Estimators of Covariance Matrices with CV Tuning*. R package version 1.0.

Wang, Y. (1997). Grkpack fitting smoothing spline anova models for exponential families. *Communications in Statistics-Simulation and Computation 26*(2), 765–782.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association 99*(467), 673–686.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

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

Wu, W. B. and M. Pourahmadi (2009). Banding sample autocovariance matrices of stationary processes. *Statistica Sinica*, 1755–1768.

Xiang, D. and G. Wahba (1996). A generalized approximate cross validation for smoothing splines with non-gaussian data. *Statistica Sinica*, 675–692.

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

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

Yang, R. and J. O. Berger (1994). Estimation of a covariance matrix using the reference prior. *The Annals of Statistics*, 1195–1211.

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

Zeger, S. L. and P. J. Diggle (1994). Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics*, 689–699.

Zhang, H. H. and Y. Lin (2006). Component selection and smoothing for nonparametric regression in exponential families. *Statistica Sinica*, 1021–1041.

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

Zhou, L. and H. Pan (2014). Principal component analysis of two-dimensional functional data. *Journal of Computational and Graphical Statistics 23*(3), 779–801.

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