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, B.A., M.S.

Graduate Program in Department of Statistics

The Ohio State University

2018

Dissertation Committee:

Yoonkyung Lee, Advisor

Katherine A. Calder

Sebastian Kurtek

Abstract

This is the abstract of the thesis and shows how the world's problems can be solved by having a cow.

New 2010 version:

> In reality, after you have actually had a cow, and written it up as your dissertation, remember that the dissertation or thesis abstract should be less than 500 words. There is no requirement for an external abstract, but an environment exists for it if this changes back. See the old instructions below for details.

Old 1996 version:

> In reality, after you have actually had a cow, and written it up as your dissertation, remember that the dissertation or thesis abstract should be less than 350 words. Two copies of the external version of the abstract must be submitted separately to the Graduate School. The environment `externalabstract` can be used to generate the required external abstract pages.

This is dedicated to the one I love . . . la la la . . .

Acknowledgments

I thank everyone who has ever had a cow. . . .

In reality, this is the only page of the dissertation which the author has full control of. You can write anything you want here, and no one can tell you it's wrong (except if the margins don't line up!!!!).

Vita

January 0, 1800	Born - Cowtown, USA
1900	B.S. Cow Science
1950	M.S. Cow-Dairy Science
1985-present	Graduate Teaching Associate, Holstein University.

Publications

Research Publications

B. Simpson "Milking a Cow". *Journal of Dairy Science*, 00(2):277–287, Feb. 1900.

Fields of Study

Major Field: Department of Statistics

Table of Contents

Page

Abstract ii

Dedication iii

Acknowledgments iv

Vita v

List of Tables ix

List of Figures xii

1. Introduction 1

2. Covariance estimation: a review 5

2.1 Structured parametric covariances 7
2.2 Shrinking the sample covariance matrix 12
2.3 Matrix decompositions 21
2.3.1 The variance-correlation decomposition 21
2.3.2 Gaussian graphical models 22
2.3.3 The spectral decomposition 25
2.3.4 The Cholesky decomposition 26
2.4 Generalized linear models for covariances 29
2.4.1 Linear models for covariance 30
2.4.2 Log-linear covariance models 32
2.4.3 The Cholesky decomposition as a generalized linear model 33

3. A reproducing kernel Hilbert space estimation framework for covariance estimation . . 41

3.1 A smoothing spline ANOVA model for the generalized autoregressive coefficients 45
3.2 Smoothing parameter selection 59
3.3 A smoothing spline model for the innovation variances 67
3.4 Smoothing parameter selection for exponential families 70

4. A P-spline model for the Cholesky decomposition 73

4.1 Tensor product B-splines for multidimensional smoothing 73
4.2 Difference penalties 81
4.3 The P-spline estimator of the generalized autoregressive coefficient function . . 91
4.4 Smoothing parameter selection 96

5. Simulation studies 101

5.1 Loss functions and corresponding risk measures 104
5.2 Alternative estimators 106
5.3 Data generation procedures 110
5.4 Results 112
5.4.1 Simulations with complete data 112
5.4.2 Performance with irregularly sampled data 119

6. Data analysis 123

7. Concluding remarks and future work 136

Appendices 138

A. Chapter 2 Appendix 138

A.1 Proof of Theorem 3.1.1 138

B. Chapter 4 140

B.1 Connecting the finite difference penalty to B-spline derivatives 140

C. Chapter 5 Appendix 143

C.1 Quadratic risk estimates for simulation study 1 143

C.2 Quadratic risk estimates for simulation study 2 145
C.3 Comprehensive tables for study 1 147

List of Tables

Table **Page**

2.1 *Ideal shape of repeated measurements.* 6

2.2 *Autoregressive coefficients and prediction error variances of successive regressions.* 29

2.3 *Ideal shape of repeated measurements.* 38

3.1 *Construction of the tensor product cubic spline subspace from marginal subspaces* $\mathcal{H}_{[1]}$, $\mathcal{H}_{[2]}$ 51

3.2 *Tensor product cubic spline subspace reproducing kernels and inner products* . . . 52

5.1 *Covariance models used for data generation in the simulation study.* 102

5.2 *Multivariate normal simulations for Model I. Estimated entropy 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.* 117

5.3 *Multivariate normal simulations for model II.* 117

5.4 *Multivariate normal simulations for model III.* 118

5.5 *Multivariate normal simulations for model IV.* 118

5.6 *Multivariate normal simulations for model V.* 118

5.7 *Model 1: 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.* 120

5.8 *Model 2: Entropy risk estimates and corresponding standard errors.* 120

5.9 *Model 3: Entropy risk estimates and corresponding standard errors.* 121

5.10 *Model 4: Entropy risk estimates and corresponding standard errors.* 121

5.11 *Model 5: Entropy risk estimates and corresponding standard errors.* 122

6.1 *Cattle data: treatment group A sample correlations.* 125

6.2 *Cattle data: treatment group A sample generalized autoregressive parameters (below the main diagonal) and log sample innovation variances (rightmost column).* . 126

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.* 143

C.2 *Multivariate normal simulation-estimated quadratic risk for model II.* 143

C.3 *Multivariate normal simulation-estimated quadratic risk for model III.* 144

C.4 *Multivariate normal simulation-estimated quadratic risk for model IV.* 144

C.5 *Multivariate normal simulation-estimated quadratic risk for model V.* 144

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.* 145

C.7 *Model 2: Quadratic risk estimates and corresponding standard errors.* 145

C.8 *Model 3: Quadratic risk estimates and corresponding standard errors.* 146

C.9 *Model 4: Quadratic risk estimates and corresponding standard errors.* 146

C.10 *Model 5: Quadratic risk estimates and corresponding standard errors.* 147

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.* 148

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.* . 149

List of Figures

Figure **Page**

4.1 *On the left: a single, isolated B-spline basis function, and on the right: several overlapping B-splines.* 77

4.2 *A set of parabolic B-splines corresponding to knot sequence* $\left\{\frac{1}{6}, \frac{1}{2}, \frac{2}{3}, 1\right\}$ 78

4.3 *Tensor product of two cubic B-splines* 80

4.4 *A subset of a full bivariate basis of cubic B-splines* 81

4.5 *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 fitted curve in the upper left plot is the most "wiggly" of any of the fits, as the penalty plays the weakest roll in the fitted coefficients there. 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.* 87

4.6 *P-spline smoothing of 10 observations using 60 B-spline basis functions.* 88

4.7 *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* $\lambda \to \infty$*.* 90

4.8 *Nine cubic B-spline tensor products with heavy linear column penalty and heavy linear row penalty* 92

4.9 $\frac{l}{2} < m < 1 - \frac{l}{2}, \quad 0 < l < 1$. 95

4.10 *The limiting behaviour of the trace of the smoothing matrix A_λ as the smoothing parameter increases for the P-spline fit to the 10 observations using 60 B-spline basis functions, shown in Figure 4.6. For weakly enforced smoothing, the effective dimension is equal to the number of observations, and as $\lambda \to \infty$, the effective dimension approaches the order of the difference penalty.* 99

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).* 104

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 of shows the true covariance structure, and each row beneath corresponds to each of the estimators.* 113

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.* . . . 114

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.* 115

6.1 *Subject-specific weight curves over time for treatment groups A and B.* 124

6.2 *Empirical estimates of the parameters of the Cholesky decomposition of the sample covariance matrix.* 127

6.3 *Cubic polynomomials fitted to the sample regressogram and log innovation variances for the cattle data from treatment group A.* 129

6.4 *Subject-specific fitted weight trajectories for cattle in treatment group A.* 131

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 $\hat{D}$.* 133

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.* 134

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 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 either 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 incorporate the arsenal of techniques for penalized regression for the unintuitive 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 incoherency 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. We propose estimation of a covariance matrix through its Cholesky decomposition which 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. We will highlight a number of approaches to parsimonious covariance modeling, but our attention will be delegated to recent progress in parsimonious covariance models for longitudinal data. The review will conclude with the presentation of matrix factorizations of the covariance matrix, translating covariance estimation into a generalized linear modeling problem, focusing on the Cholesky factorization, the parameters of which can be viewed as the parameters of a specific regression model. Chapters 3 and 4 present this regression as a bivariate smoothing problem and outline two approaches to model specification and estimation. 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 wrap up our thoughts on our current work and remark on potential 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, and ensuring that the estimated parameters simultaneously satisfy the structural constraints on the elements of a covariance matrix. The number of parameters of a $p \times p$ covariance matrix $\Sigma = (\sigma_{ij})$ grows quadratically in the dimension, and these parameters must satisfy the positive-definiteness constraint

$$y'\Sigma y = \sum_{i,j=1}^{p} y_i y_j \sigma_{ij} \geq 0. \tag{2.1}$$

for all $y = (y_1, \ldots, y_p)' \in \Re^p$. 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. Perhaps even more importantly, however, is that its construction produces a positive definite estimate, so that the daunting 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 $t_i = (t_{i1}, t_{i2}, \ldots, t_{ip})'$. Construction of the sample covariance matrix requires rectangular data; Table table:ideal-repeated-measurements shows the ideal shape of a (rectangular) longitudinal data set. Unfortunately, longitudinal studies frequently produce non-rectangular data, having observation times which are not evenly-spaced and common across all subjects. Constructing a sample estimate of the covariance matrix with such data is a much less straightforward endeavor.

Table 2.1: *Ideal shape of repeated measurements.*

		Occasion					
		1	2	...	t	...	m
Unit	1	y_{11}	y_{12}	...	y_{1t}	...	y_{1m}
	2	y_{21}	y_{22}	...	y_{2t}	...	y_{2m}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	i	y_{i1}	y_{i2}	...	y_{it}	...	y_{im}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	N	y_{N1}	y_{N2}	...	y_{Nt}	...	y_{Nm}

These drawbacks have accelerated numerous initiatives detouring the pitfalls on the most obvious route to a covariance estimate. At the root of most of these approaches employ methods to confine the parameter space to that of a lower dimension to aid in ensuring positive definiteness and eliminating some of the the $p\left(p+1\right)/2$ parameters to be estimated. 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.

2.1 Structured parametric covariances

In the applied statistics literature, particularly for repeated measure data, it is quite common to pick a stationary covariance matrix for the covariance structure. This is an approach due to the computational simplicity associated with model fitting. Software packages implementing fitting procedures for a growing number of simple models are readily accessible. Typical choices for the covariance matrix are parsimonious models which depend on a small number of parameters, and in the following section, we review a selection of these which are frequently encountered in the literature.

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 covariance matrix are given by

$$\sigma_{ij} = \begin{cases} \rho, & i \neq j, \\ \sigma^2, & i = j, \end{cases} \tag{2.3}$$

where σ_{ij} denotes the (i, j) element of Σ. 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.

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

$$y_t = \begin{cases} \mu_t + \epsilon_t, & t = 1, \\ \mu_t + \rho\left(y_{t-1} - \mu_{t-1}\right) + \epsilon_t, & t > 1, \end{cases} \tag{2.4}$$

where $|\rho| < 1$, and the innovations $\{\epsilon_t\}$ are independently distributed according to $N\left(0, \sigma_t^2\right)$ with $\sigma_1^2 = \sigma^2/\left(1 - \rho^2\right)$, and $\sigma_t^2 = \sigma^2$ for $t = 2, \ldots, p$. The corresponding dependence components of the covariance structure are monotonically decreasing in $l = |i - j|$; specifically,

$$\sigma_{ij} = \begin{cases} \rho^{|i-j|}, & i \neq j, \\ \sigma^2, & i = j, \end{cases} \tag{2.5}$$

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

$$y_t = \begin{cases} \mu_t + \epsilon_t, & t = 1, \\ \mu_t + \sum_{j=1}^{p^*} \phi_j \left(y_{t-j} - \mu_{t-j}\right) + \epsilon_t, & t > 1, \end{cases}$$

where $p^* = \min\left(p, t - 1\right)$, and the $\{\epsilon_t\}$ are independent mean zero Normal random variables. The variance of $\{\epsilon_t\}$ is constant for $t > p$, and for $t \leq p$, 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 $|i - j|$.

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

$$y_t = \sum_{j=0}^{q} \theta_j \epsilon_{t-j}, \tag{2.6}$$

where the $\{\epsilon_t\}$ are independently and identically distributed mean zero Normal random variables with variance σ^2. This model corresponds to covariance matrix having elements defined as follows:

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

Thus, variances are constant and correlations between y_t and y_{t-l} vanish beyond a finite, constant lag l. Here $\rho_1, \ldots, \rho_q$are arbitrary parameters subject only to positive definiteness constraints. This model generalizes to a q^{th}-order Toeplitz model, which specifies

$$\sigma_{ij} = \begin{cases} \rho_{i-j} & |i-j| \leq q, \\ 0 & |i-j| > q, \\ \sigma^2 & i = j, \end{cases} \tag{2.7}$$

or covariance matrix of the form

$$\begin{bmatrix} m_0 & m_1 & m_2 & \ldots & m_{p-1} \\ m_1 & m_0 & m_1 & \ldots & m_{p-2} \\ m_2 & m_1 & m_0 & \ldots & m_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{p-1} & m_{p-2} & m_{p-3} & \ldots & m_0 \end{bmatrix}, \tag{2.8}$$

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

In turn, one can further generalize to a q^{th}-order banded model by specifying that the covariances on off-diagonals of the correlation matrix beyond the q^{th} off-diagonal are zero, and otherwise not imposing any structural restrictions on the remaining elements of the covariance matrix

beyond those required for positive definiteness. The tradeoff of the additional flexibility of the general banded model over the MA and Toeplitz models is that the number of parameters in a general q-banded covariance structure is $O(n)$ rather than $O(1)$.

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 $O(n)$ parameters, while specifying linear or quadratic dependence on time leads to more parsimonious heterogeneous 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:

$$y_t = \mu_t + \sum_{j=1}^{t} \epsilon_j, \quad t = 1, \ldots, p, \tag{2.9}$$

where the ϵ_t are independent mean zero Normal random variables with variance σ_ϵ^2. The variance of the process increases linearly in time, and the correlation between y_t and y_{t-l} also increases, but nonlinearly, in time:

$$\sigma_{ij} = \begin{cases} \sqrt{i/j} & i \neq j \\ j\sigma_\epsilon^2 & i = j, \end{cases} \tag{2.10}$$

This model is applicable to longitudinal data only when data 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 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 is given by

$$y_i = X_i\beta + Z_i\gamma_i + \epsilon_i, \quad i = 1, \ldots, p, \tag{2.11}$$

where 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 $N\left(0, \sigma^2 \mathrm{I}_{n_i}\right)$. The G_i are usually assumed to be equal, so the covariance matrix of y_i is taken to be $\Sigma_i = Z_i G Z_i' + \sigma^2 \mathrm{I}_{n_i}$. Special cases include the linear random coefficients (RCL) and quadratic random coefficients (RCQ) models. In the linear case, $Z_i = \left[1_{m_i}, \left(t_{i1}, \ldots, t_{i,m_i}\right)'\right]$ and

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

In the quadratic case, $Z_i = \left[1_{m_i}, \left(t_{i1}, \ldots, t_{i,m_i}\right)', \left(t_{i1}^2, \ldots, t_{i,m_i}^2\right)'\right]$. It is worth noting that when $Z_i = 1_{m_i}$, the random coefficient model corresponds to the compound symmetric model 2.5. The

covariance structure for a subject having measurements $y_1, \ldots, y_{m_i}$ taken at equally spaced measurement times $t_1 = 1, \ldots, t_{m_i} = m_i$ is given by

$$\sigma_{ij} = \begin{cases} \frac{\sigma_{00}+\sigma_{01}(i+j)+\sigma_{11}ij}{\sqrt{\sigma^2+\sigma_{00}+2i\sigma_{01}+\sigma_{11}i^2}\sqrt{\sigma^2+\sigma_{00}+2j\sigma_{01}+j^2\sigma_{11}}} & i \neq j \\ \sigma^2 + \sigma_{00} + 2\sigma_{01}j + \sigma_{11}j^2 & i = j, \end{cases} \tag{2.12}$$

These models are can 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.

The previous list highlights a number of major parametric covariance specifications, but it is far from an exhaustive list of parametric covariance structures - we will later reference structures which we have not discussed here, such as antedependence models. For additional models for repeated measures data, seeJennrich and Schluchter (1986), for example.

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 a challenge 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. Where parametric models are one extreme, exhibiting low variance but

potentially high bias, the other extreme is the sample covariance matrix: unbiased but trading stability for its flexibility. In between these poles lie a broad class of estimators which seek to balance the stability of parametric models with the flexibility of the sample covariance matrix.

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 greater than the dimension, $N >> p$ (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), Champion (2003)

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. He considered estimators of the form

$$\hat{\Sigma} = \Sigma(S) = P\Phi(\lambda)P', \tag{2.13}$$

where $\lambda = (\lambda_1, \ldots, \lambda_p)'$, $\lambda_1 > \cdots > \lambda_p$ are the ordered eigenvalues of S, P is the orthogonal matrix whose i^{th} column is the normalized eigenvector of S corresponding to λ_i, and $\Phi(\lambda) = diag(\phi_1, \ldots, \phi_p)$ is the diagonal matrix where $\phi_j(\lambda)$ is an estimate of the j^{th} largest eigenvalue of Σ. Letting $\phi_j(\lambda) = \lambda_j$ corresponds to the usual unbiased estimator S. It is known that λ_1 and λ_p are biased low and high, respectively, so Stein chooses $\Phi(\lambda)$ 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 $\phi_j = \frac{N\lambda_j}{\alpha_j}$, where

$$\alpha_j(\lambda) = N - p + 2\lambda_j \sum_{i \neq j} \frac{1}{\lambda_j - \lambda_i}. \tag{2.14}$$

The Stein estimators ϕ_j differ from the sample eigenvalues when 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 not satisfy $\phi_1 < \cdots < \phi_p$.

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 $p >> N$, 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

$$\hat{\Sigma} = \alpha_1 I + \alpha_2 S, \tag{2.15}$$

where α_1, α_2 are chosen to optimize the Frobenius norm of $\hat{\Sigma} - S$ or the slightly modified Frobenius norm:

$$L\left(\hat{\Sigma}, \Sigma\right) = p^{-1}||\hat{\Sigma} - \Sigma||^2 = p^{-1}\text{tr}\left(\hat{\Sigma} - \Sigma\right)^2.$$

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

$$\begin{aligned} \mu &= \text{tr}\left(\Sigma\right)/p, \\ \alpha^2 &= ||\Sigma - \mu I||^2, \\ \beta^2 &= ||S - \Sigma||^2, \\ \delta^2 &= ||S - \mu I||^2. \end{aligned} \tag{2.16}$$

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 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.

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 $p >> N$. 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 assuming 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 be still be large. Moving average (MA) and autoregressive (AR) 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

$$B_k\left(S\right) = \left[s_{ij} 1\left(|i-j| \leq k\right)\right] \tag{2.17}$$

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

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

which is applicable if, 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

$$\hat{\Sigma} = R * S \tag{2.18}$$

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. However, several have used the same concept on the lower triangular matrix of the Cholesky decomposition of Σ^{-1}, including Wu and Pourahmadi (2003), Huang et al. (2006), Levina et al. (2008). Banding the Cholesky factor mitigates the need for the tapering matrix to be positive definite, since the parameters of the reparameterization are completely free while still guaranteeing that the estimate is positive definite. Detailed discussion follows in Section 2.3.4.

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:

$$S^{\omega} = \left[\omega_{ij}^{k} s_{ij}\right], \tag{2.19}$$

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. Without loss of generality, we assume that $k_h = k/2$ is even. The weights may be rewritten as

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

This expression of the weights makes it clear 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 banding estimator of Bickel and Levina (2008).

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 sample estimates to zero can improve the quality of estimators. 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. For $\lambda > 0$, a thresholding operator $\int_\lambda (z) : \Re \to \Re$ satisfies

- $\int_\lambda (z) \leq z$;
- $\int_\lambda (z) = 0$ for $|z| \leq \lambda$;
- $|\int_\lambda (z) - z| \leq \lambda$

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, including the soft-thresholding estimator given by

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

where σ^*_{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. Their generalized thresholding estimator $\int_\lambda (z)$ is the solution to

$$\int_\lambda (z) = \arg\min_\sigma \left[\frac{1}{2}\left(\sigma - z\right)^2 + J\left(\sigma\right)\right], \tag{2.20}$$

where J penalizes the size of the elements of the estimated matrix. Soft thresholding results from minimizing 2.20 using the lasso penalty, $J_\lambda = \lambda|\sigma|$, which corresponds to thresholding rule

$$\int_{\lambda}(\sigma) = \text{sign}(\sigma)(\sigma - \lambda)_{+}. \tag{2.21}$$

For detailed discussion of the connection between penalty functions and the resulting thresholding rules, see Antoniadis and Fan (2001). These estimators are simple to compute compared to competitor estimates like the penalized likelihood with LASSO penalty, 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.

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.

The performance of any regularized estimator depends heavily on the quality of tuning parameter selection. The Frobenius is a natural measure of the accuracy of an estimator; it quantifies the sum over the unique elements of Σ of the the first term in 2.20,

$$||\hat{\Sigma}^{\lambda} - \Sigma||^2 = \left(\sum_{i,j} \left(\hat{\sigma}_{ij}^{\lambda} - \sigma_{ij} \right)^2 \right)^{1/2} \tag{2.22}$$

If Σ were available, one would choose the value of the tuning parameter λ which minimizes (2.22). In practice, one tries to first approximate the risk, or

$$E_{\Sigma} \left[||\hat{\Sigma}^{\lambda} - \Sigma||^2 \right],$$

and then choose the optimal value of λ. As in regression methods, cross validation and a number of its variants have become popular choices for tuning parameter selection in covariance estimation, though unanimous agreement on which precise procedure is optimal is fleeting. K-fold cross validation requires first splitting the data into folds $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$. The value of the tuning parameter is selected to minimize

$$\mathbf{CV}_F(\lambda) = \arg\min_{\lambda} K^{-1} \sum_{k=1}^{K} ||\hat{\Sigma}^{(-k)} - \tilde{\Sigma}^{(k)}||_F^2, \tag{2.23}$$

where $\tilde{\Sigma}^{(k)}$ is the unregularized estimator based on based on $\mathcal{D}_k$, and $\hat{\Sigma}^{(-k)}$ is the regularized estimator under consideration based on the data after holding $\mathcal{D}_k$ out. Using this approach, the size of the training data set is approximately $(K-1)N/K$, and the size of the validation set is approximately N/K (though these quantities are only relevant when subjects have equal numbers of observations). For linear models, it has been shown that cross validation is asymptotically consistent is the ratio of the validation data set size over the training set size goes to 1. See Shao (1993). This result motivates the reverse cross validation criterion, which is defined as follows:

$$\mathbf{rCV}_F(\lambda) = \arg\min_{\lambda} K^{-1} \sum_{k=1}^{K} ||\hat{\Sigma}^{(k)} - \tilde{\Sigma}^{(-k)}||_F^2, \tag{2.24}$$

where $\tilde{\Sigma}^{(-k)}$ is the unregularized estimator based on based on the data after holding out $\mathcal{D}_k$, and $\hat{\Sigma}^{(k)}$ is the regularized estimator under consideration based on $\mathcal{D}_k$.

2.3 Matrix decompositions

The positive definite constraint poses a challenge in most covariance estimation settings. In the following section, we demonstrate the role of a selection matrix decompositions in removing it from the estimation procedure altogether. These decompositions decompose the covariance matrix into its variance and dependence components, a primary reason why they have been used in covariance modeling successfully. They are closely connected to the use of generalized linear models for covariance estimation, and 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 is perhaps the most familiar of the following three parameterizations, which parameterizes the covariance matrix according to

$$\Sigma = DRD, \tag{2.25}$$

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 to be nonnegative, 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 (see (Zimmerman and Núñez-Antón, 1997), (Diggle, 2002)).

2.3.2 Gaussian graphical models

The marginal (pairwise) dependence among the entries of a random vector are captured by the off-diagonal entries of Σ or the entries of the correlation matrix $R = (\rho_{ij})$. However, the conditional dependencies can be found in the off-diagonal entries of the precision matrix $\Sigma^{-1} = (\sigma^{ij})$. More precisely, for Y a mean zero normal random vector with a positive-definite covariance matrix, if the (i, j) component of the precision matrix is zero, then given the other variables, y_i and y_j are conditionally independent (Anderson (1984)).

Gaussian graphical models are a common way of representing the conditional independence structure in a p-dimensional random vector Y, with the nodes of the graph corresponding to variables. The absence of an edge between variables i and j, or a zero in the (i, j) position of the inverse covariance matrix indicates that the two variables are conditionally independent. The entries of the variance-correlation decomposition of the precision matrix

$$\Sigma^{-1} = \left(\sigma^{ij}\right) = \tilde{D}\tilde{R}\tilde{D} \tag{2.26}$$

can be interpreted as certain coefficients of a regression model, which assumes no natural ordering of the p variables corresponding to the columns of the covariance matrix. This lack of assumed structure among the dimensions of the matrix make it a less natural choice for modeling the covariance of longitudinal data. However, we've included it as part of this discussion because these models share a number of similarities to the Cholesky decomposition, which plays a central role in our contribution to the work in this area.

A number of regression-based approaches to modeling the precision structure have spawned from the work of Meinhausen and Buhlmann (2006). Their method is based on solving p separate LASSO regression problems. The entries of $\left(\tilde{R}, \tilde{D}\right)$ have direct statistical interpretations in terms of partial correlations, and variance of predicting a variable given the rest. Regression calculations can be used to show that the partial correlation coefficient between y_i and y_j after removing the linear effect of the $p-2$ remaining variables is given by

$$\tilde{\rho}_{ij} = -\frac{\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}}. \tag{2.27}$$

The partial variance of y_i after removing the linear effect of the remaining $p-$ variables is given by

$$\tilde{d}_{ii}^2 = \frac{1}{\sigma^{ii}}. \tag{2.28}$$

To connect these parameters to those of a regression model, consider partitioning random vector $Y = (y_1, \ldots, y_p)'$ into two components $(Y_1', Y_2')'$ of dimensions p_1 and p_2, and similarly partitioning its covariance and precision matrices:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \tag{2.29}$$

Let $\Phi_{2|1}$ denote the $p_2 \times p_1$ matrix of regression coefficients resulting from the least squares regression of Y_2 on Y_1, and let $e_{2|1} = Y_2 - \Phi_{2|1} Y_1$ denote the corresponding vector of residuals. The

regression coefficients $\Phi_{2|1}$ and residuals $e_{2|1}$ are obtained from restricting $e_{2|1}$ to be uncorrelated with Y_1:

$$\begin{aligned} \Phi_{2|1} &= \Sigma_{21}\Sigma_{11}^{-1} \\ &= -\left(\Sigma^{22}\right)^{-1}\Sigma^{21} \end{aligned} \tag{2.30}$$

$$\begin{aligned} Cov\left(e_{2|1}\right) &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \\ &= \Sigma_{22|1} = \left(\Sigma^{22}\right)^{-1}. \end{aligned} \tag{2.31}$$

If we let $p_2 = 1$, then one can establish the relationship between elements of the inverse covariance matrix and these regression coefficients and conditional covariances. When $Y_1 = Y_{-(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_p)'$ and Y_2 corresponds to a single y_i, $\Sigma_{22|1}$, a scalar, is referred to as the *partial variance* of y_i given the other variables. Denote the linear least squares predictor of y_i based on $Y_{-(i)}$ by y_i^* and $\epsilon_i^* = y_i - y_i^*$ with prediction variance $Var\left(\epsilon_i^*\right) = d^{*2}_i$. Then

$$y_i = \sum_{j \neq i} \beta_{ij} y_j + \epsilon_i^*,$$

where (2.31) and (2.32) give

$$\begin{aligned} \beta_{ij} &= -\frac{\sigma^{ij}}{\sigma^{ii}}, \quad j \neq i \\ d^{*2}_i &= Var\left(y_i | y_j\right) = \frac{1}{\sigma_{ii}}, \quad j \neq i, \;\; i = 1, \ldots, p \end{aligned} \tag{2.32}$$

Thus, the unconstrained regression coefficient of the j^{th} variable when we regressing y_i on the rest of the variables is given by the (i, j) entry of the inverse covariance matrix. The partial correlation between y_i and y_j can be defined if we consider the case where $p_2 = 2$. Letting $Y_2 = (y_i, y_j)'$, $i \neq j$ and $Y_1 = Y_{-(ij)}$ contain the remaining $p - 2$ variables, the covariance of (y_i, y_j) after removing the linear effects of $\{y_k : k \neq i, j\}$ is given by

$$\Sigma_{22|1} = \begin{bmatrix} \sigma^{ii} & \sigma^{ij} \\ \sigma^{ji} & \sigma^{jj} \end{bmatrix}^{-1}$$
$$= \frac{1}{\sigma^{ii}\sigma^{jj} - (\sigma^{ij})^2} \begin{bmatrix} \sigma^{jj} & -\sigma^{ij} \\ -\sigma^{ij} & \sigma^{ii} \end{bmatrix}$$

The regression coefficients (2.32) can be written in terms of the partial correlation between y_i and y_j:

$$\rho^*_{ij} = -\frac{\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{ij}}}. \tag{2.33}$$

Rewriting the β_{ij}, we have

$$\beta_{ij} = \rho^*_{ij}\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}, \tag{2.34}$$

which shows that the sparsity of the inverse covariance matrix mirrors that of the matrix of partial correlations. This parallel motivates estimation of the inverse covariance matrix by fitting a sequence of penalized regression models, notably the approach taken by Peng et al. (2012) which imposes a Lasso penalty on the off-diagonal elements of the partial correlation matrix.

2.3.3 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

$$\Sigma = P\Lambda P' = \sum_{i=1}^{p} \lambda_i e_i e_i', \tag{2.35}$$

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:

$$\log \Sigma = P \log \Lambda P' = \sum_{i=1}^{p} \log (\lambda_i) \, e_i e_i', \tag{2.36}$$

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.

2.3.4 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

$$\Sigma = CC', \tag{2.37}$$

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; see 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 $D^{1/2} = diag\,(c_{11}, \ldots, c_{pp})$, the standard Cholesky decomposition 2.37 can be written

$$\Sigma = CD^{-1/2}D^{1/2}D^{1/2}D^{-1/2}C' = LDL', \tag{2.38}$$

where $L = D^{-1/2}C$. This is commonly referred to as the modified Cholesky decomposition (MCD) of Σ. We can also write the modified Cholesky decomposition of the inverse covariance matrix:

$$D = T\Sigma T', \quad \Sigma^{-1} = T'D^{-1}T, \tag{2.39}$$

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 $y_{t-1}, \ldots, y_1$. Standard regression machinery gives us that there exist unique scalars ϕ_{tj} so that

$$y_t = \begin{cases} \epsilon_t, & t = 1 \\ \sum_{j=1}^{t-1} \phi_{t,j} y_j + \epsilon_t, & t = 2, \ldots, p, \end{cases} \tag{2.40}$$

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 (2.3.4) 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 $\epsilon = (y_1, \ldots, y_p)'$ denote the vector of uncorrelated prediction residuals with

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

Then model (2.3.4) can be written

$$\epsilon = TY, \tag{2.41}$$

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: $\sigma_t^2 = var(\epsilon_t)$.

$$\begin{bmatrix} 1 & & & & \\ -\phi_{21} & 1 & & & \\ -\phi_{31} & -\phi_{32} & 1 & & \\ \vdots & & & \ddots & \\ -\phi_{p1} & -\phi_{p2} & \ldots & -\phi_{p,p-1} & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{bmatrix} \tag{2.42}$$

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_{ij}s are regression coefficients, for any unstructured covariance matrix, these and the log innovation variances are unconstrained. The regression coefficients of the model in () are referred to as the *generalized autoregressive parameters* (GARP) and *innovation variances* (IV) (Pourahmadi (1999), Pourahmadi (2000)). The powerful implication of the parallel regression framework of decomposition (2.39) 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_{ij}, 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	$\cdots$	y_{m-1}	y_m
1					
ϕ_{21}	1				
ϕ_{31}	ϕ_{32}	1			
$\vdots$	$\vdots$		$\ddots$		
$\vdots$	$\vdots$			$\ddots$	
ϕ_{p1}	ϕ_{p2}	$\cdots$	$\cdots$	$\phi_{p,p-1}$	1
σ_1^2	σ_1^2	$\cdots$	$\cdots$	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 has lead to large classes of models including nonparametric and generalized additive models, Bayesian GLM, and generalized linear mixed models (see (Hastie and Tibshirani, 1990), (Dey et al., 2000), (McCulloch and Neuhaus, 2001)), and 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 Ω^{-1}. Dempster (1972) who 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:

$$\Sigma = \sum_{i=1}^{q} \alpha_q U_q \tag{2.43}$$

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. In, for q large enough, any covariance matrix admits representation of the form (2.43), since one can decompose every covariance matrix as

$$\Sigma = \sum_{i=1}^{p}\sum_{j=1}^{p} \sigma_{ij} U_{ij}, \tag{2.44}$$

where U_{ij} is an $p \times p$ matrix with a 1 in the (i, j) position, and zeros everywhere else. The linear model (2.43) can be viewed as modeling the link-transformed covariance $g\left(\Sigma\right) = \sum_{i=1}^{q} \alpha_q U_q$, where $g\left(\cdot\right)$ is the identity link. Despite the convenience of 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 (2.43) and the covariance parameterization given in (2.44) 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 observations 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

$$\sigma_{ij} = 0 \quad \text{for } |i-j| > q, \tag{2.45}$$

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

$$\sigma^{ij} = 0 \quad \text{for } |i-j| > p. \tag{2.46}$$

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 said to be p^{th}-order antedependent or $AD(p)$ Gabriel (1962) if y_t and y_{t+s+1} are independent given the intervening values $y_{t+1}, \ldots, y_{t+s}$ for $t = 1, \ldots, p?s?1$ and all $s \geq p$. A random vector $Y = (y_1, \ldots, y_p)$ is $AD(p)$ if and only if its covariance matrix satisfies (2.46). 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 (2.43) 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

$$\Sigma = P \Lambda P', \tag{2.47}$$

its matrix logarithm, denoted $\log \Sigma$, and defined by $log\Sigma = P \log \Lambda P'$ 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:

$$g(\Sigma) = \log \Sigma = \sum_{i=1}^{q} \alpha_i U_i, \tag{2.48}$$

where the U_is are as before in 2.43 and the α_is are now unconstrained. The α_is, however, now lack statistical interpretation since $g(A) = \log A$ is a highly nonlinear operation. But for diagonal Σ, $\log \Sigma = \text{diag}(\sigma_{11}, \ldots, \sigma_{pp})$, and model 2.48 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 issued 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 hybrid link constructed from the modified Cholesky decomposition of Σ^{-1} given in 2.49 combines ideas in Edgeworth (1892), Gabriel (1962), Anderson (1973), Dempster (1972), Chiu et al. (1996),

and Zimmerman and Núñez-Antón (1997). It 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. For an unstructured covariance matrix Σ, the nonredundant entries of the components $(T, \log D)$ of the modified Cholesky decompostion 2.39 can be written as the entries of

$$g(\Sigma) = 2I - T - T' + \log D. \tag{2.49}$$

These entries are unconstrained, allowing them to be modeled using any desired technique, including parametric, semi- and nonparametric, and Bayesian approaches. Including covariates in any proposed model for these components can be done so seamlessly. As in the usual GLM setting for estimation of the mean, one can elicit parametric models for ϕ_{tj} and $\log \sigma_t^2$. For example, one might model the nonredundant entries of T, say, linearly as in model 2.43 and those of $\log D$ as in, say, model 2.48, letting

$$\begin{aligned} \phi_{tj} &= x_{tj}'\beta, \\ \log \sigma_t^2 &= z_t'\gamma, \end{aligned} \tag{2.50}$$

where x_{tj} and z_t denote $q \times 1$ and $p \times 1$ vectors of known covariates, and $\beta = (\beta_1, \ldots, \beta_q)'$ and $\gamma = (\gamma_1, \ldots, \gamma_p)'$ are the parameters relating these covariates to the innovation variances and the dependence among the elements of Y. Covariates most frequently used in the analysis of real longitudinal data sets are low order polynomials of lag and time, modeling

$$\begin{aligned} z_{jk}' &= \left(1, t_j - t_k, (t_j - t_k)^2, \ldots, (t_j - t_k)^{p-1}\right)' \\ z_i' &= \left(1, t, \ldots, t^{q-1}\right)' \end{aligned} \tag{2.51}$$

Pourahmadi (1999), Pourahmadi (2000), and Pan and Mackenzie (2003) prescribe methods for identifying models such as model 2.50 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 $\log \sigma_{ij}^2$, and the generalized autoregressive parameters ϕ_{ij}. 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 is reduces a potentially high dimensional problem to something much more computationally feasible; if one models the innovation variances $\sigma^2(t)$ 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. An $p \times p$ Toeplitz matrix Σ is a matrix with elements σ_{ij} such that $\sigma_{ij} = \sigma_{|i-j|}$ i.e. a matrix of the form (2.8), having entries which are constant on each subdiagonal.

Nany have alternatively proposed nonparametric and semiparametric techniques approaches to avoid bias incurred with model misspecification. 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 is given by

$$
\begin{aligned}
-2\ell\left(Y_1, \ldots, Y_N, \Sigma\right) &= \sum_{i=1}^{N} \left(\log|\Sigma| + Y_i'\Sigma^{-1}Y_i'\right) \\
&= N\log|D| + N\text{tr}\Sigma^{-1}S \\
&= N\log|D| + N\text{tr}D^{-1}TST',
\end{aligned}
\tag{2.52}
$$

where $S = N^{-1}\sum_{i=1}^{N} Y_i Y_i'$. The negative log likelihood (2.52) is quadratic in T for fixed D, so the MLE for the ϕ_{ij} has closed form. Similarly, the MLE for D for fixed T has closed form. See Pourahmadi (2000). While the MLE is flexible and thus exhibits low bias, this advantage can be offset with high variance, so to balance the tradeoff between bias and variance, shrinkage or regularization may be applied to estimates to improve stability of estimators.

The fact that the entries of T are unconstrained makes the Cholesky decomposition ideal for nonparametric estimation and regularization methods. 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 (2.3.4). A similar procedure by Dahlhaus et al. (1997) uses varying coefficient regression models for each subdiagonal of T:

$$
y_t = \sum_{j=1}^{t-1} f_{j,p}\left(t/p\right) y_{t-j} + \sigma_p\left(t/p\right)
$$

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. Their procedure is not

capable of handling missing or irregular data. Huang et al. (2007) jointly model the mean and covariance matrix of longitudinal data using basis function expansions. They treat the subdiagonals of T as smooth functions, approximated by B-splines and carry out estimation maximum (normal) likelihood. Their method permits subject-specific observations times, but assumes that observation times lie on some notion of a regular grid. They treat within-subject gaps in measurements as missing data and which they handle using the E-M algorithm. Regularization is achieved through the choice of k, the number of nonzero subdiagonals, and the total number of basis functions used to approximate the k smoothed diagonals. They treat these as tuning parameters and use BIC for model selection. Due to the closer connection between entries of T and the family of regression (2.3.4), it is conceivable that T exhibits sparsity, having some of its entries could be zero or close to it. 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. Levina et al. (2008) impose a banded structure on the Cholesky factor using penalized maximum likelihood estimation. A novel penalty that they call the nexted Lasso 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.

Table 2.3 shows the ideal, rectangular shape of such data where N units (subjects, stocks, households, financial instruments, etc.) are measured repeatedly on one variable. In most longitudinal studies, the functional trajectories of the involved smooth random processes are not directly observable. Often, the observed data are noisy, sparse and irregularly spaced measurements of these trajectories. In the case that subjects don't share a common set of observation times, the notion of the discrete lag doesn't have a clear definition. In turn, it is not clear then, how one would

apply smoothing to each subdiagonal of T since this relies on data observed on a regular grid. Moreover, if one believes that the data used to inform one subdiagonal could inform subdiagonals close to it, failing to smooth in both directions fails to make use of this information. In Chapter 3, we outline a proposed framework for covariance estimation based on the Cholesky decomposition, viewing T as a continuous function in both the lag direction as well as the direction orthogonal to it. Using this approach allows us to also remove any restriction on observation times being regularly spaced and the same across subject. Henceforth, we take Y_i and $\epsilon_i = \left(\epsilon_{i1}, \ldots, \epsilon_{i,m_i}\right)'$ to be continuous processes $Y(t)$, $\epsilon(t)$ observed at discrete measurement times $t_1, \ldots, t_{m_i}$. Using a likelihood-based estimation approach alongside a functional interpretation of the GARPs permits a natural way to regularize the estimator and allow any functional characterizations of the dependency structure to be entirely data driven.

Table 2.3: *Ideal shape of repeated measurements.*

		Occasion					
		1	2	...	t	...	m
Unit	1	y_{11}	y_{12}	...	y_{1t}	...	y_{1m}
	2	y_{21}	y_{22}	...	y_{2t}	...	y_{2m}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	i	y_{i1}	y_{i2}	...	y_{it}	...	y_{im}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	N	y_{N1}	y_{N2}	...	y_{Nt}	...	y_{Nm}

Through the Cholesky decomposition, we formulate covariance estimation as a penalized regression problem and propose novel covariance penalties designed to yield natural null models presented in the literature. By transforming the axes of the design points, we express these penalties in terms of two directions: the lag component and the additive component and characterize

the solution coefficient function in terms of a functional ANOVA decomposition. Some have sidestepped the issue of high dimensionality by prescribing simple parametric models for the elements of the Cholesky decomposition. Chen et al. (2011), Pourahmadi (1999), and Pourahmadi and Daniels (2002) have elicited stationary parametric models for the generalized autoregressive coefficients, letting the GARPs depend only on the distance between two time points. To induce the structural simplicity of such stationary models with the flexibility of a nonparametric approach, we penalize all functional components but that corresponding to the lag component so that the set of null models is comprised of stationary models. Huang et al. (2007) follow the hueristic argument presented in Pourahmadi (1999) that the generalized autoregressive parameters are monotone decreasing in as lag increases and set off-diagonal elements of either the covariance matrix or the Cholesky factor corresponding to large lags to zero. Rather than shrinking element of the Cholesky factor to zero after particular value of l, we choose to enforce structure of the Cholesky factor such that the null models coincide with parsimonious models commonly used in time series analysis and with simple parametric models proposed in the nonparametric covariance estimation literature.

Modeling $\phi_{ij} = \phi(t_i, t_j)$ as a smooth bivariate function, we cast the problem of estimating a covariance matrix as the estimation of a functional varying coefficient model. The existing body of literature surrounding these models is an extensive one; see Şentürk and Müller (2008), Şentürk et al. (2013), and Noh and Park (2010). This class of models is both flexible and interpretable, making them a pragmatic modeling choice when understanding the underlying data generating mechanism is of as much importance as strong predictive capability. We employ two representations of the GARPs, which we refer to as the *generalized autoregressive coefficient function* within this frame. Chapter 3 presents a reproducing kernel Hilbert space framework for the estimation of both ϕ and σ^2. Chapter 4 an alternative representation the varying coefficient function using the penalized B-splines of Eilers and Marx (1996), which are based on the regularized estimation

of a smooth function using a B-spline basis expansion with a discrete finite difference penalty on adjacent B-spline coefficients. The connection between the simple difference penalty to the usual spline penalty on the second derivative is easy to establish using the derivative properties of the basis functions. We demonstrate their simple construction and how it facilitates an estimation framework that permits any regularization which is independent of the basis construction.

Chapter 3: A reproducing kernel Hilbert space estimation framework for covariance estimation

If we consider the Cholesky decomposition of Σ within such functional context, it is natural to extent the same notion to the elements of T and D. We take the GARPs $\{\phi_{tj}\}$ and innovation variances to be the evaluation of the smooth functions $\tilde{\phi}(t,s)$ and $\sigma^2(t)$ at observed time points, which we assume are drawn from some distribution having compact domain $\mathcal{T}$. Without loss of generality, we take $\mathcal{T} = [0,1]$. Henceforth, we view $\tilde{\phi}$ and σ^2 as a smooth continuous functions, but for ease of exposition, we let $\tilde{\phi}_{ij}$ denote the varying coefficient function evalutated at (t_i, t_j):

$$\tilde{\phi}_{tj} = \tilde{\phi}(t_i, t_j).$$

Adopting similar notation for the innovation variance function, denote $\sigma_j^2 = \sigma^2(t_j)$ where $0 \leq t_j < t_i \leq 1$ for $j < i$. This leads to varying coefficient model

$$y(t_i) = \sum_{j=1}^{i-1} \tilde{\phi}(t_i, t_j)\, y(t_j) + \sigma(t_j)\, \epsilon(t_j) \qquad i = 1, \ldots, p, \tag{3.1}$$

Our goal is now to estimate the above model, utilizing bivariate smoothing to estimate $\tilde{\phi}(t,s)$ for $0 \leq s < t \leq 1$, and one-dimensional smoothing to estimate $\sigma(t)$, $0 \leq t \leq 1$. Our proposed method for covariance estimation defines a flexible, general framework which makes all of the existing techniques for penalized regression accessible for the seemingly far different task of estimating a covariance matrix.

Our approach to estimation is constructed to provide a fully data-driven methodology for selecting the optimal covariance model (given some optimization criterion) from a expansive class of estimators ranging in complexity from that of the previously aforementioned parametric models to that of completely unstructured estimators, like the sample covariance matrix. We leverage the collection of regularization techniques that are accessible in the usual function estimation setting. By properly specifying the roughness penalty, our optimization procedure results in null models which correspond to the parametric and semiparametric models for ϕ and σ^2 discussed in Chapter 2. To facilitate the penalty specification that achieves this, we consider modeling the varying coefficient function which takes inputs

$$\begin{aligned} l &= t - s \\ m &= \frac{t+s}{2}, \end{aligned} \tag{3.2}$$

where l is the continuous analogue of the usual "lag" between time points t and s, and m is simply its orthogonal direction. We have discussed many parsimonious covariance structures which model $y(t)$ as a stationary process with covariance function which depends on time points t_i and t_j only through the Euclidean distance $||t_i - t_j||$ between them. Covariance functions taking the form $Cov(y(t_i), y(t_j)) = G(t_i, t_j) = G(||t_i - t_j||)$ can then be written as

$$Cov(y(t_i), y(t_j)) = G(l_{ij})$$

where $l_{ij} = |t_i - t_j|$. Regularizing the functional components of the Cholesky decomposition so that functions incurring large penalty correspond to functions which vary in only l and are constant in m allows us to model nonstationarity in a fully data-driven way. Our goal is to estimate

$$\phi(l, m) = \phi\left(s - t, \frac{1}{2}(s + t)\right) = \tilde{\phi}(t, s). \tag{3.3}$$

While our framework allows for estimation of the autoregressive coefficient function and the innovation variance function via any nonparametric regression setup, we focus on two primary approaches for representing ϕ and σ. First, we assume that ϕ belongs to a reproducing kernel Hilbert space, $\mathcal{H}$ and employ the smoothing spline methods of Kimeldorf and Wahba (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.

Let random vector Y follow a multivariate normal distribution with zero mean vector and covariance Σ. The loglikelihood function $\ell(Y, \Sigma)$ satisfies

$$-2\ell(Y, \Sigma) = \log|\Sigma| + Y'\Sigma Y \tag{3.4}$$

Using $T\Sigma T' = D$, we can write

$$|\Sigma| = |D| = \prod_{i=1}^{m} \sigma_i^2$$

and

$$\Sigma^{-1} = T'D^{-1}T.$$

Writing 3.4 in terms of the prediction errors and their variances of the non-redundant entries of (T, D), we have

$$\begin{aligned} -2\ell(Y, \Sigma) &= \log|D| + Y'T'D^{-1}TY \\ &= \sum_{i=1}^{m} \log\sigma_i^2 + \sum_{i=1}^{m} \frac{\epsilon_i^2}{\sigma_i^2}, \end{aligned} \tag{3.5}$$

where

$$\epsilon_i = \begin{cases} y(t_1), & i = 1, \\ y(t_i) - \sum_{j=1}^{i-1} \phi(\boldsymbol{v}_{ij}) y_j, & i = 2, \ldots, p, \end{cases} \tag{3.6}$$

where $\phi(\boldsymbol{v}_{ij}) = \phi(l_{ij}, m_{ij}) = \tilde{\phi}(t_i, t_j)$. Accommodating subject-specific sample sizes and measurement times merely requires appending an additional index to observation times. Let $Y_1, \ldots, Y_N$ denote a sample of N independent mean zero random trajectories from a multivariate normal distribution with common covariance Σ. We associate with each trajectory $Y_i = \left(y_{i1}, \ldots, y_{i,m_i}\right)'$ with a vector of potentially subject-specific observation times $\left(t_{i1}, \ldots, t_{i,m_i}\right)'$, so that the j^{th} measurement of trajectory i is modeled

$$\begin{aligned} y(t_{ij}) &= \sum_{k=1}^{j-1} \tilde{\phi}(t_{ij}, t_{ik}) y(t_{ik}) + \sigma(t_{ij}) \epsilon(t_{ij}) \\ &= \sum_{k=1}^{j-1} \phi(\boldsymbol{v}_{ijk}) y(t_{ik}) + \sigma(t_{ij}) \epsilon(t_{ij}) \end{aligned} \tag{3.7}$$

for $i = 1, \ldots, N$, $j = 2, \ldots, m_i$. Making similar ammendments to indexing, the joint log likelihood for the sample $Y_1, \ldots, Y_N$ is given by

$$-2\ell\left(Y_1, \ldots, Y_N, \phi, \sigma^2\right) = \sum_{i=1}^{N} \sum_{j=1}^{m_i} \log \sigma_{ij}^2 + \sum_{i=1}^{N} \sum_{j=1}^{m_i} \frac{\epsilon_{ij}^2}{\sigma_{ij}^2}, \tag{3.8}$$

With this, we can estimate ϕ and $\log \sigma^2$ using maximum likelihood or any of its penalized variants by appending a roughness penalty (penalties) to 3.8. Employing regularization, we take ϕ, σ^2 to minimize

$$-2\ell\left(Y_1, \ldots, Y_N, \phi, \sigma^2\right) + \lambda J(\phi) + \breve{\lambda} \breve{J}\left(\sigma^2\right), \tag{3.9}$$

where J and $\breve{J}$ are roughness penalties on ϕ and σ^2, and λ, $\breve{\lambda}$ are non-negative smoothing parameters. To jointly estimate the GARP function and the IV 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 3.8 starts with initializing $\left\{\sigma_{ij}^2\right\} = 1$ for $i = 1, \ldots, N$, $j = 1, \ldots, m_i$. For fixed σ^2, the penalized likelihood (as a function of ϕ) is given by

$$-2\ell\left(Y_1, \ldots, Y_N, \phi|\sigma^2\right) + \lambda J\left(\phi\right) = \sum_{i=1}^{N}\sum_{j=2}^{m_i}\sigma_{ij}^{-2}\left(y_{ij} - \sum_{k<j}\phi\left(\boldsymbol{v}_{ijk}\right)y_{ik}\right)^2 + \lambda J\left(\phi\right), \quad (3.10)$$

which corresponds to the usual penalized least squares functional encountered in the nonparametric function estimation literature. The first term, the residual sums of squares, encourages the fitted function's fidelity to the data. The second term penalizes the roughness of ϕ, and λ is a smoothing parameter which controls the tradeoff between the two conflicting concerns. Given ϕ^* the minimizer of 3.10 and setting $\phi = \phi^*$, we update our estimate of σ^2 by minimizing

$$-2\ell\left(Y_1, \ldots, Y_N, \sigma^2|\phi\right) + \breve{\lambda}\breve{J}\left(\sigma^2\right) = \sum_{i=1}^{N}\sum_{j=2}^{m_i}\log\sigma_{ij}^2 + \sum_{i=1}^{N}\sum_{j=1}^{m_i}\sigma_{ij}^{-2}r_{ij}^{*\,2} + \breve{\lambda}\breve{J}\left(\sigma^2\right), \quad (3.11)$$

where the $\left\{r_{ij}^{*\,2} = \left(y_{ij} - \sum_{k<j}\phi^*\left(\boldsymbol{v}_{ijk}\right)y_{ik}\right)\right\}$ denote the working residuals based on the current estimate of ϕ. This process of iteratively updating ϕ^* and σ^{2*} is repeated until convergence is achieved.

3.1 A smoothing spline ANOVA model for the generalized autoregressive coefficients

This section presents a reproducing kernel Hilbert space (RKHS) framework for estimating the generalized autoregressive coefficient function ϕ, and in later sections we apply the same approach for estimating the innovation variance function, σ^2. Specifically, we adopt the smoothing spline ANOVA models developed by Gu (2002) to connect fitted models to parsimonious models proposed in the literature for the components of the Cholesky decomposition. The flexibility of this

framework permits an entirely data-driven modeling approach through careful penalty specification and the use of already well-developed model selection methods. Though RKHS methods, and in particular smoothing spline ANOVA models, have been studied extensively for nonparametric function estimation (see Aronszajn (1950), Wahba (1990), and Berlinet and Thomas-Agnan (2011) for detailed examinations), to our knowledge they have received little attention in the context of covariance models. To demonstrate our framework, we first must establish some notation and review the relevant mathematical details of reproducing kernel Hilbert spaces.

A Hilbert space $\mathcal{H}$ of functions on a set $\mathcal{V}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is defined as a complete inner product linear space. A Hilbert space is called a reproducing kernel Hilbert space if the evaluation functional $[\boldsymbol{v}] f = f(\boldsymbol{v})$ is continuous in $\mathcal{H}$ for all $\boldsymbol{v} \in \mathcal{V}$. The Reisz Representation Theorem gives that there exists $Q \in \mathcal{H}$, the representer of the evaluation functional $[\boldsymbol{v}](\cdot)$, such that $\langle Q_{\boldsymbol{v}}, \phi \rangle_{\mathcal{H}} = \phi(\boldsymbol{v})$ for all $\phi \in \mathcal{H}$. See Gu (2013), Theorem 2.2.

The symmetric, bivariate function $Q(\boldsymbol{v}_1, \boldsymbol{v}_2) = Q_{\boldsymbol{v}_2}(\boldsymbol{v}_1) = \langle Q_{\boldsymbol{v}_1}, Q_{\boldsymbol{v}_2} \rangle_{\mathcal{H}}$ is called the reproducing kernel (RK) of $\mathcal{H}$. The RK satisfies that for every $\boldsymbol{v} \in \mathcal{V}$ and $f \in \mathcal{H}$,

I. $Q(\cdot, \boldsymbol{v}) \in \mathcal{H}$

II. $f(\boldsymbol{v}) = \langle f, Q(\cdot, v) \rangle_{\mathcal{H}}$

The first property is called the reproducing property of Q. Every reproducing kernel uniquely determines the RKHS, and in turn, every RKHS has unique reproducing kernel. See Gu (2013), Theorem 2.3. The kernel satisfies that for any $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n_1}\}$, $\{\breve{\boldsymbol{v}}_1, \ldots, \breve{\boldsymbol{v}}_{n_2}\} \in \mathcal{V}$ and $\{a_1, \ldots, a_{n_1}\}$, $\{a_1, \ldots, a'_{n_2}\} \in \Re$,

$$\langle \sum_{i=1}^{n_1} a_i Q(\cdot, \boldsymbol{v}_i), \sum_{j=1}^{n_2} a'_j Q(\cdot, \breve{\boldsymbol{v}}_j) \rangle_{\mathcal{H}}. \tag{3.12}$$

Let $\mathcal{N}_J = \{\phi : \ J(\phi) = 0\}$ denote the null space of J, and consider the decomposition

$$\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J.$$

The space $\mathcal{H}_J$ is a RKHS having $J(\phi)$ as the squared norm. The representer of any bounded linear functional can be obtained from the reproducing kernel Q. Let ψ_{ij} denote the representer for the evaluation functional, L_{ij}, i.e. ψ_{ij} satisfies

$$\langle \psi_{ij}, \phi \rangle = L_{ij}\phi, \quad \phi \in \mathcal{H}.$$

Then one may write $\psi(\boldsymbol{v}_{ij})$ as the inner product of itself with the reproducing kernel:

$$\psi_{ij}(\boldsymbol{v}) = \langle \psi_{ij}, Q\boldsymbol{v} \rangle = L_{ij}Q\boldsymbol{v} = L_{ij(\cdot)}Q(\boldsymbol{v}, \cdot) \tag{3.13}$$

where the notation $L_{ij(\cdot)}$ indicates that L_{ij} is applied to what immediately follows as a function of $(\cdot)$, so that one can obtain $\psi_{ij}(\boldsymbol{v})$ by applying L_{ij} to $Q(\boldsymbol{v}, \boldsymbol{v}^*)$, considered as a function of $\boldsymbol{v}^*$. Wahba (1990) established an explicit form for the minimizer of the penalized sums of squares

$$-2\ell_\phi + \lambda J(\phi) = \sum_{i=1}^{N}\sum_{j=2}^{m_i} \sigma_{ij}^{-2} \left(y_{ij} - \sum_{k<j} \phi(\boldsymbol{v}_{ijk})\, y_{ik} \right)^2 + \lambda J(\phi), \tag{3.14}$$

which can now be written

$$-2\ell_\phi + \lambda J(\phi) = \sum_{i=1}^{N}\sum_{j=2}^{m_i} \sigma_{ij}^{-2} \left(y_{ij} - \sum_{k<j} (L_{ijk}\phi)\, y_{ik} \right)^2 + \lambda ||P_J\phi||_{\mathcal{H}}^2, \tag{3.15}$$

where P_J is the projection operator which projects ϕ onto the subspace $\mathcal{H}_J$, and L_{ijk} denotes the evaluation functional $[\boldsymbol{v}_{ijk}]\,\phi$.

Theorem 3.1.1. *Let $\{\eta_1, \ldots, \eta_{d_0}\}$ span the null space of P_J, $\mathcal{H}_0$. Let $V = \bigcup_{i,j,k} \boldsymbol{v}_{ijk} \equiv \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{|V|}\}$ denote the set of unique within-subject pairs of observation times. Let B denote the $|V| \times d_0$ matrix*

having i^{th} column equal to η_i evaluated at the vector of observed $\boldsymbol{v} \in V$, and assume that B has full column rank. Then the minimizer ϕ_λ of 3.15 is given by

$$\phi_\lambda = \sum_{\nu=1}^{d_0} d_\nu \eta_n u + \sum_{\boldsymbol{v}_i \in V} c_i \xi_i, \tag{3.16}$$

where $\xi_i = P_J \psi_i$ is the projection of L_i, the representer for the evaluation functional corresponding to the i^{th} element of V, onto $\mathcal{H}_J$.

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

Convenient construction of a reproducing kernel Hilbert space on a domain

$$\mathcal{V} = \mathcal{V}_1 \otimes \mathcal{V}_2$$

which can be written as a product domain, is available through the tensor product of the RKHS for each of the marginal domains $\mathcal{V}_1$ and $\mathcal{V}_2$. Without loss of generality, we can let $l,\ m \in [0,1] = \mathcal{V}_1 = \mathcal{V}_2$. Given Hilbert space for the domain of l, $\mathcal{H}_{[1]}$ with reproducing kernel Q_1 and Hilbert space on the domain of m, $\mathcal{H}_{[2]}$ with reproducing kernel Q_2, the reproducing kernel $Q = Q_{[1]} Q_{[2]}$ corresponds to that of the tensor product space of $\mathcal{H}_{[1]}$ and $\mathcal{H}_{[2]}$, denoted

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

See Gu (2002), Theorem 2.6. Let $\mathcal{A}_1$, $\mathcal{A}_2$ denote the averaging operators defining ANOVA decompositions on $\mathcal{H}_{[1]}$, $\mathcal{H}_{[2]}$, respectively, where $\mathcal{H}_{0[i]}$ has RK $Q_{0[i]}$, $i = 1, 2$ and $\mathcal{H}_{1[i]}$ has RK $Q_{1[i]}$ satisfying $\mathcal{A}_1 Q_{[1]}(l, \cdot) = \mathcal{A}_2 Q_{[2]}(m, \cdot) = 0$. Then the tensor product space $\mathcal{H}$ has tensor sum decomposition

$$\begin{aligned}\mathcal{H} &= [\mathcal{H}_{0[1]} \oplus \mathcal{H}_{1[1]}] \otimes [\mathcal{H}_{0[2]} \oplus \mathcal{H}_{1[2]}] \\ &= [\mathcal{H}_{0[1]} \otimes \mathcal{H}_{0[2]}] \oplus [\mathcal{H}_{0[1]} \otimes \mathcal{H}_{1[2]}] \oplus [\mathcal{H}_{1[1]} \otimes \mathcal{H}_{0[2]}] \oplus [\mathcal{H}_{1[1]} \otimes \mathcal{H}_{1[2]}]\end{aligned} \tag{3.17}$$

If $Q_{0[i]} \propto 1$ for $i = 1, 2$, then $\mathcal{H}$ can be further simplified:

$$\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2, \tag{3.18}$$

which has reproducing kernel $Q = Q_{[1]}Q_{[2]}$.

Example 3.1.1. Tensor product cubic spline

Let the marginal domains of l and m correspond to $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively, where

$$\mathcal{H}_i = \mathcal{C}^{(m_i)} = \left\{ \phi : \int_0^1 \phi^{(m_i)} \, dv < \infty \right\},$$

which are equipped with inner product

$$\begin{aligned}\langle f, g \rangle &= \langle f, g \rangle_0 + \langle f, g \rangle_1 \\ &= \sum_{\nu=0}^{m_i - 1} M_\nu f M_\nu g + \int_0^1 f^{(m_i)}(v)\, g^{(m_i)}(v)\, dv, \quad i = 1, 2\end{aligned} \tag{3.19}$$

where the order i differential operator M_ν is defined $M_\nu \phi = \int_0^1 \phi^{(m)}(v)\, dv \;, \quad \nu = 1, \ldots, m_i$, $i = 1, 2$. Denote the norm corresponding to this inner product by

$$||f||^2 = \langle f, f \rangle = \langle f, f \rangle_0 + \langle f, f \rangle_1 = ||P_0 f||^2 + ||P_1 f||^2$$

The reproducing kernel Q can be expressed in terms of the scaled Bernoulli polynomials $\left\{ k_j(v) = \frac{1}{j!} B_j(v) \right\}$, $v \in [0, 1]$, where B_j is defined according to:

$$\begin{aligned} B_0(x) &= 1 \\ \frac{d}{dx} B_j(x) &= j B_{j-1}(x), \; j = 1, 2, \ldots \end{aligned}$$

One can verify that $\int_0^1 k_\mu^\nu dv = \delta_{\mu,\nu}$ for $\nu, \mu = 0, \ldots, m_i - 1$, where $\delta_{\mu,\nu}$ is the Kronecker delta. This implies that the k_ν, $\nu = 0, \ldots, m_i - 1$ for an orthonormal basis for $\mathcal{H}_{0[i]} = \left\{\phi : \phi^{(m_i)} = 0\right\}$ under the inner product

$$\langle f, g\rangle_0 = \sum_{\nu=0}^{m_i-1} M_\nu f M_\nu g, \quad i = 1, 2,$$

and that

$$Q_{0[i]}(v, v') = \sum_{\nu=0}^{m_i-1} k_\nu(v)\, k_\nu(v')$$

is the reproducing kernel for $\mathcal{H}_{0[i]}$. The subspaces of $\mathcal{H}_{[i]}$ which are orthogonal to $\mathcal{H}_{0[i]}$ are comprised of functions ϕ satisfying

$$\mathcal{H}_{1[i]} = \{\phi : M_\nu f = 0, \;\; \nu = 0, 1, \ldots, m_i - 1, \int_0^1 \phi^{(m_i)}\, dv < \infty\}, \quad i = 1, 2.$$

One can show that the representer for the evaluation functional $[v]\,\phi$ in $\mathcal{H}_{1[i]}$ with squared norm $\langle f, g\rangle_1 = \int_0^1 f^{(m_i)} g^{(m_i)}\, dv$ is given by the function

$$Q_{[i]_v}'(v) = k_{m_i}(v)\, k_{m_i}(v') + (-1)^{m_i - 1} k_{2m_i}(v' - v) \tag{3.20}$$

See Gu (2002) Example 2.3.3 for proof. The tensor product smoothing spline results from letting $m_1 = m_2 = 2$, so that the marginal subspaces can be written

$$\{\phi : \phi'' \in \mathcal{L}_2[0,1]\} = \{\phi : \phi \propto 1\} \oplus \{\phi : \phi \propto k_1\} \oplus \left\{\phi : \int_0^1 \phi dv = \int_0^1 \phi' dv = 0,\; \phi'' \in \mathcal{L}_2[0,1]\right\} \tag{3.21}$$

$$= \mathcal{H}_{00} \oplus \mathcal{H}_{01} \oplus \mathcal{H}_1, \tag{3.22}$$

where $\mathcal{H}_{01} \oplus \mathcal{H}_1$ forms the contrast in a one-way ANOVA decomposition with averaging operator $\mathcal{A}\phi = \int_0^1 \phi \, dv$. The corresponding reproducing kernels are

$$Q_{00}(v, v') = 1 \tag{3.23}$$

$$Q_{01}(v, v') = k_1(v) \, k_1(v') \tag{3.24}$$

$$Q_1(v, v') = k_2(v) \, k_2(v') - k_4(v - v') \, . \tag{3.25}$$

The tensor product space can be constructed with nine tensor sum terms; the construction of the tensor product space from the terms of the tensor sum. The corresponding reproducing kernels and inner products are given in Table 3.1 and Table 3.2, respectively.

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

Table 3.1: *Construction of the tensor product cubic spline subspace from marginal subspaces* $\mathcal{H}_{[1]}$, $\mathcal{H}_{[2]}$

Table 3.2: *Tensor product cubic spline subspace reproducing kernels and inner products*

Subspace	Reproducing kernel	Inner product
$\mathcal{H}_{00[1]} \otimes \mathcal{H}_{00[2]}$	1	$\left(\int_0^1 \int_0^1 f\right)\left(\int_0^1 \int_0^1 g\right)$
$\mathcal{H}_{01[1]} \otimes \mathcal{H}_{00[2]}$	$k_1(l)\, k_1(l')$	$\left(\int_0^1 \int_0^1 f'_{[1]}\right)\left(\int_0^1 \int_0^1 g'_{[1]}\right)$
$\mathcal{H}_{01[1]} \otimes \mathcal{H}_{01[2]}$	$k_1(l)\, k_1(l')\, k_1(m)\, k_1(m')$	$\left(\int_0^1 \int_0^1 f''_{[12]}\right)\left(\int_0^1 \int_0^1 g''_{[12]}\right)$
$\mathcal{H}_{1[1]} \otimes \mathcal{H}_{00[2]}$	$k_2(l)\, k_2(l') - k_4(l-l')$	$\int_0^1 \left(\int_0^1 f''_{[12]}\, dl'\right)\left(\int_0^1 g''_{[12]}\, dl'\right)\, dl$
$\mathcal{H}_{1[1]} \otimes \mathcal{H}_{01[2]}$	$[k_2(l)\, k_2(l') - k_4(l-l')]\, k_1(m)\, k_1(m')$	$\int_0^1 \left(\int_0^1 f^{(3)}_{[112]}\, dl'\right)\left(\int_0^1 g^{(3)}_{[112]}\, dl'\right)\, dl$
$\mathcal{H}_{1[1]} \otimes \mathcal{H}_{1[2]}$	$[k_2(l)\, k_2(l') - k_4(l-l')]\, [k_2(m)\, k_2(m') - k_4(m-m')]$	$\int_0^1 \int_0^1 f^{(4)}_{[1122]} g^{(4)}_{[1122]}$

For $\boldsymbol{v} \in V$ where V is a product domain, ANOVA decompositions can be characterized by

$$\mathcal{H} = \bigoplus_{\beta=0}^{g} \mathcal{H}_\beta \tag{3.26}$$

and

$$J(\phi) = \sum_{\beta=0}^{g} \theta_\beta^{-1} J_\beta(\phi_\beta), \tag{3.27}$$

where $\phi_\beta \in \mathcal{H}_\beta$, J_β is the square norm in $\mathcal{H}_\beta$, and $0 < \theta_\beta < \infty$. This gives

$$\begin{aligned}
\mathcal{H}_0 &= \mathcal{N}_J \\
\mathcal{H}_J &= \bigoplus_{\beta=1}^{g} \mathcal{H}_\beta, \text{ and} \\
Q &= \sum_{\beta=1}^{g} \theta_\beta Q_\beta,
\end{aligned}$$

where Q_β is the RK in $\mathcal{H}_\beta$. The $\{\theta_\beta\}$ are additional smoothing parameters, which are implicit in notation to follow for the sake of concise demonstration.

Let Y denote the vector of length $n_y = \sum_i M_i - N$ 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:

$$\begin{aligned}
Y &= (Y_1', Y_2', \ldots, Y_N')' \\
&= \left(y_{12}, y_{13}, \ldots, y_{1,m_1}, \ldots, y_{N,2}, y_{N,3}, \ldots, y_{N,m_N}\right)'
\end{aligned}$$

Define X_i to be the $m_i \times |V|$ matrix containing the covariates necessary for regressing each measurement $y_{i2}, \ldots, y_{i,m_i}$ on its predecessors as in model 3.7, and stack these on top of one another to obtain

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}, \tag{3.28}$$

which has dimension $n_y \times |V|$. Then the solution ϕ_λ minimizing 3.15 is the solution to the minimization problem

$$||D^{-1/2}\left(Y - X\left(Bd + Qc\right)\right)||^2 + \lambda c' Q c \tag{3.29}$$

where the (i, j) entry of the $|V| \times |V|$ matrix Q is given by $\langle P_1\xi_i, P_1\xi_j\rangle_{\mathcal{H}}$. The $|V| \times d_0$ matrix B has i-ν^{th} element equal to $\eta_\nu\left(\boldsymbol{v}_i\right)$, and we assume B to be full column rank. The diagonal matrix D holds the $n_y \times n_y$ innovation variances σ^2_{ijk}.

Example 3.1.2. Construction of X_i with complete data

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

$$X_i = \begin{bmatrix} y_{i,t_1} & 0 & 0 & 0 & & \ldots & 0 \\ 0 & y_{i,t_1} & y_{i,t_2} & 0 & 0 & \ldots & 0 \\ \vdots & & & & & & \\ 0 & \ldots & 0 & \ldots & y_{i,t_1} & \ldots & y_{i,t_{p-1}} \end{bmatrix} \tag{3.30}$$

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

Example 3.1.3. 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 $m_1 = 4$, $m_2 = 4$, 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

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

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 $v = (l, m)$ 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:

$$X_1 = \begin{bmatrix} 0 & y_{1,1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & y_{1,1} & 0 & y_{1,2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & y_{1,1} & y_{1,2} & y_{1,3} \end{bmatrix} \tag{3.31}$$

and similarly, for subject 2:

$$X_2 = \begin{bmatrix} y_{2,1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & y_{2,1} & y_{2,2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & y_{2,1} & y_{2,2} & y_{2,3} & 0 & 0 & 0 \end{bmatrix} \tag{3.32}$$

Construction of the solution $\hat{\phi}$

Differentiating $-2\ell_\phi + \lambda J(\phi)$ with respect to c and d and setting equal to zero, we have that

$$\frac{\partial}{\partial c}\left[-2\ell_\phi + \lambda J\left(\phi\right)\right] = QX'D^{-1}\left[X\left(Bd+Qc\right) - Y\right] + \lambda Qc = 0$$
$$\iff X'D^{-1}X\Big[Bd+Qc\Big] + \lambda c = X'D^{-1}Y \tag{3.33}$$

$$\frac{\partial}{\partial d}\left[-2\ell_\phi + \lambda J\left(\phi\right)\right] = B'X'D^{-1}\left[X\left(Bd+Qc\right) - Y\right] = 0$$
$$\iff -\lambda B'c = 0 \tag{3.34}$$

For fixed smoothing parameter, the solution ϕ is obtained by finding c and d which satisfy

$$Y = X\Big[Bd + \Big(Q + \lambda\left(X'D^{-1}X\right)^{-1}\Big)c\Big] \tag{3.35}$$
$$B'c = 0 \tag{3.36}$$

Letting $\tilde{Y} = D^{-1/2}Y$, $\tilde{B} = D^{-1/2}XB$, and $\tilde{Q} = D^{-1/2}XQ$, the penalized log likelihood **??** may be written

$$-2\ell_\lambda\left(c,d\right) + \lambda J\left(\phi\right) = \Big[\tilde{Y} - \tilde{B}d - \tilde{Q}c\Big]'\Big[\tilde{Y} - \tilde{B}d - \tilde{Q}c\Big] + \lambda c'Qc. \tag{3.37}$$

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

$$\begin{aligned}\tilde{B}'\tilde{B}d + \tilde{B}'\tilde{Q}c &= \tilde{B}'\tilde{Y}\\ \tilde{Q}'\tilde{B}d + \tilde{Q}'\tilde{Q}c + \lambda Qc &= \tilde{Q}'\tilde{Y},\end{aligned} \tag{3.38}$$

Some algebra yields that this is equivalent to solving the system

$$\begin{bmatrix}\tilde{B}'\tilde{B} & \tilde{B}'\tilde{Q}\\ \tilde{Q}'\tilde{B} & \tilde{Q}'\tilde{Q} + \lambda Q\end{bmatrix}\begin{bmatrix}d\\ c\end{bmatrix} = \begin{bmatrix}\tilde{B}'\tilde{Y}\\ \tilde{Q}'\tilde{Y}\end{bmatrix} \tag{3.39}$$

Fixing smoothing parameters λ and θ_β (hidden in Q and $\tilde{Q}$ if present), assuming that $\tilde{Q}$ is full column rank, 3.39 can be solved by the Cholesky decomposition of the $(n+d_0)\times(n+d_0)$ matrix followed by forward and backward substitution. See Golub and Van Loan (2012). Singularity of $\tilde{Q}$ demands special consideration. Write the Cholesky decomposition

$$\begin{bmatrix} \tilde{B}'\tilde{B} & \tilde{B}'\tilde{Q} \\ \tilde{Q}'\tilde{B} & \tilde{Q}'\tilde{Q}+\lambda Q \end{bmatrix} = \begin{bmatrix} C_1' & 0 \\ C_2' & C_3' \end{bmatrix} \begin{bmatrix} C_1 & C_2 \\ 0 & C_3 \end{bmatrix} \tag{3.40}$$

where $\tilde{B}'\tilde{B} = C_1'C_1$, $C_2 = C_1^{-T}\tilde{B}'\tilde{Q}$, and $C_3'C_3 = \lambda Q + \tilde{Q}'\left(I - \tilde{B}\left(\tilde{B}'\tilde{B}\right)^{-1}\tilde{B}'\right)\tilde{Q}$. 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

$$\tilde{C}_3 = \begin{bmatrix} H_1 & H_2 \\ 0 & \delta I \end{bmatrix}, \quad \tilde{C} = \begin{bmatrix} C_1 & C_2 \\ 0 & \tilde{C}_3 \end{bmatrix}; \tag{3.41}$$

then

$$\tilde{C}^{-1} = \begin{bmatrix} C_1^{-1} & -C_1^{-1}C_2\tilde{C}_3^{-1} \\ 0 & \tilde{C}_3^{-1} \end{bmatrix}. \tag{3.42}$$

Premultiplying 3.40 by $\tilde{C}^{-T}$, straightforward algebra gives

$$\begin{bmatrix} I & 0 \\ 0 & \tilde{C}_3^{-T}C_3^TC_3\tilde{C}_3^{-1} \end{bmatrix} \begin{bmatrix} \tilde{d} \\ \tilde{c} \end{bmatrix} = \begin{bmatrix} C_1^{-T}\tilde{B}'\tilde{Y} \\ \tilde{C}_3^{-T}\tilde{Q}'\left(I - \tilde{B}\left(\tilde{B}'\tilde{B}\right)^{-1}\tilde{B}'\right)\tilde{Y} \end{bmatrix} \tag{3.43}$$

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

$$\begin{aligned}\tilde{C}_3^{-T} C_3^T C_3 \tilde{C}_3^{-1} &= \begin{bmatrix} K' \\ L' \end{bmatrix} C_3' C_3 \begin{bmatrix} K & L \end{bmatrix} \\ &= \begin{bmatrix} K' \\ L' \end{bmatrix} H' H \begin{bmatrix} K & L \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}.\end{aligned}$$

If $L' C_3^T C_3 L = 0$, then $L' \tilde{Q}' \left(I - \tilde{B} \left(\tilde{B}' \tilde{B} \right)^{-1} \tilde{B}' \right) \tilde{Q} L = 0$, so $L' \tilde{Q}' \left(I - \tilde{B} \left(\tilde{B}' \tilde{B} \right)^{-1} \tilde{B}' \right) \tilde{Y} = 0$. Thus, the linear system has form

$$\begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{d} \\ \tilde{c}_1 \\ \tilde{c}_2 \end{bmatrix} = \begin{bmatrix} * \\ * \\ 0 \end{bmatrix}, \tag{3.44}$$

which can be solved, but with c_2 arbitrary. One may perform the Cholesky decomposition of 3.39 with pivoting, replace the trailing 0 with δI for appropriate value of δ, and proceed as if $\tilde{Q}$ were of full rank.

It follows that

$$\widehat{\tilde{Y}} = \tilde{B} d + \tilde{Q} c = \begin{bmatrix} \tilde{B} & \tilde{Q} \end{bmatrix} \tilde{C}^{-1} \tilde{C}^{-T} \begin{bmatrix} \tilde{B}' \\ \tilde{Q}' \end{bmatrix} \tilde{Y} = \tilde{A}_{\lambda, \boldsymbol{\theta}} \tilde{Y}. \tag{3.45}$$

where

$$\begin{aligned}\tilde{A}_{\lambda, \boldsymbol{\theta}} &= \begin{bmatrix} \tilde{B} & \tilde{Q} \end{bmatrix} \tilde{C}^{-1} \tilde{C}^{-T} \begin{bmatrix} \tilde{B}' \\ \tilde{Q}' \end{bmatrix} \\ &= G + (I - G) \tilde{Q} \left[\tilde{Q}' (I - G) \tilde{Q} + \lambda Q \right]^{-1} \tilde{Q}' (I - G),\end{aligned} \tag{3.46}$$

for

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

3.2 Smoothing parameter selection

By varying smoothing parameters λ and θ_β, the minimizer ϕ_λ of 3.39 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 a 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 demonstrates 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 $\tilde{Y} = D^{-1/2}Y$, $\tilde{B} = D^{-1/2}XB$, and $\tilde{Q} = D^{-1/2}XQ$ as before. Let $\tilde{\epsilon} = D^{-1/2}\epsilon$ denote the vector of length $\sum_{i=1}^{N} m_i - N$ containing the standardized prediction errors $\epsilon_{ij} \sim N(0,1)$, and write the vector of transformed means

$$\Phi = D^{-1/2}X\left[Bd + Qc\right]. \tag{3.47}$$

We can assess $\hat{\tilde{Y}}_\lambda$, an estimate of the mean of $\tilde{Y}$ based on observed data y_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, m_i$, using the loss function

$$\begin{aligned} L(\lambda) &= \sum_{i=1}^{N}\sum_{j=1}^{m_i}\left(\hat{\tilde{y}}_{ij} - E\left[\tilde{y}_{ij}\right]\right)^2 \\ &= ||\tilde{Y} - \tilde{\mu}||^2 \end{aligned} \tag{3.48}$$

where $\mu = D^{-1/2}W\Phi^*$ denotes the $\left(\sum_i m_i - N\right) \times 1$ with i^{th} element equal to the expected value of the i^{th} element of $\tilde{Y}$. Then straightforward algebra yields that

$$L(\lambda) = \mu'\left(I - \tilde{A}_{\lambda,\boldsymbol{\theta}}\right)^2\mu - 2\mu'\left(I - \tilde{A}_{\lambda,\boldsymbol{\theta}}\right)^2\tilde{A}_{\lambda,\boldsymbol{\theta}}\tilde{\epsilon} + \tilde{\epsilon}'\tilde{A}^2_{\lambda,\boldsymbol{\theta}}\tilde{\epsilon} \tag{3.49}$$

Define the unbiased risk estimate

$$U(\lambda) = \frac{1}{N}\tilde{Y}'\left(I - \tilde{A}_{\lambda,\boldsymbol{\theta}}\right)^2\tilde{Y} + \frac{2}{N}\text{tr}\tilde{A}_{\lambda,\boldsymbol{\theta}} \tag{3.50}$$

Adding and subtracting μ to the quadratic terms, one can verify with straightforward algebra that

$$\begin{aligned} U(\lambda) &= \left(\tilde{Y} - \mu + \mu - \tilde{A}_{\lambda,\boldsymbol{\theta}}\tilde{Y}\right)'\left(\tilde{Y} - \mu + \mu - \tilde{A}_{\lambda,\boldsymbol{\theta}}\tilde{Y}\right) + 2\text{tr}\tilde{A} \\ &= \left(\tilde{A}\tilde{Y} - \mu\right)'\left(\tilde{A}\tilde{Y} - \mu\right) + \tilde{\epsilon}'\tilde{\epsilon} + 2\tilde{\epsilon}'\left(I - \tilde{A}\right)\mu - 2\left(\tilde{\epsilon}'\tilde{A}\tilde{\epsilon} - \text{tr}\tilde{A}\right) \end{aligned} \tag{3.51}$$

This gives

$$U(\lambda) - L(\lambda) - \tilde{\epsilon}'\tilde{\epsilon} = 2\tilde{\epsilon}'\left(I - \tilde{A}\right)\mu - 2\left(\tilde{\epsilon}'\tilde{A}\tilde{\epsilon} - \text{tr}\tilde{A}\right), \tag{3.52}$$

which allows one to easily see that $U(\lambda)$ is unbiased for the relative loss $L(\lambda) + \tilde{\epsilon}'\tilde{\epsilon}$. Under mild conditions on the risk function

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

one can establish that U is also a consistent estimator. See Gu (2013), Chapter 3 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

$$V_{loso}(\lambda) = \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{Y}_i - \widehat{\tilde{\mu}}_i^{[-i]}\right)'\left(\tilde{Y}_i - \widehat{\tilde{\mu}}_i^{[-i]}\right) \tag{3.53}$$

where $\widehat{\tilde{\mu}}_i^{[-i]}$ is the estimate of $E\left[\tilde{Y}_i\right]$ 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 $\tilde{Y}_i^* = \left(\tilde{y}_{i1}^*, \tilde{y}_{i1}^*, \ldots, \tilde{y}_{i,m_i}^*\right)$. We may write the mean squared prediction error for the new observations as follows:

$$\begin{aligned} MSPE &= \frac{1}{N}\sum_{i=1}^{N} E\left[||\tilde{Y}_i^* - \widehat{\tilde{\mu}}_i||^2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} E\left[||\tilde{Y}_i^* - D_i^{-1/2}W_i\Phi^* + D_i^{-1/2}W_i\Phi^* - D_i^{-1/2}W_i\hat{\Phi}^*||^2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{m_i + E\left[||\tilde{\mu}_i - \widehat{\tilde{\mu}}_i^{[-i]}||^2\right]\right\} \end{aligned} \tag{3.54}$$

where $\tilde{\epsilon}_i = \tilde{Y}_i^* - D_i^{-1/2}W_i\Phi^*$. When $\{\sigma^2(t)\}$ is known, $\tilde{\epsilon}_i$ is a mean zero multivariate normal vector with $Cov(\tilde{\epsilon}_i) = I_{m_i}$, which gives the last equality. Since $\tilde{Y}_i$ and $\widehat{\tilde{\mu}}_i$ are independent, the

expected LosoCV score can be written

$$E\left[V_{loso}\left(\lambda\right)\right] = \frac{1}{N}\sum_{i=1}^{N}\left\{m_i + E\left[||\widehat{\tilde{\mu}}_i - \tilde{\mu}_i||^2\right]\right\}. \tag{3.55}$$

When N is large, we expect that $\widehat{\tilde{\mu}}_i$ should be close to $\widehat{\tilde{\mu}}_i^{[-i]}$, so $E\left[V_{loso}\left(\lambda\right)\right]$ 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 3.2.1 (Shortcut formula for LosoCV)**.** *The LosoCV score satisfies the following identity:*

$$V_{loso}\left(\lambda\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{Y}_i - \widehat{\tilde{Y}}_i\right)'\left(I_{ii} - \tilde{A}_{ii}\right)^{-T}\left(I_{ii} - \tilde{A}_{ii}\right)^{-1}\left(\tilde{Y}_i - \widehat{\tilde{Y}}_i\right),$$

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_{ii} is a $m_i \times m_i$ identity matrix.

A detailed presentation and proof can be found in Xu et al. (2012) and supplementary materials Xu and Huang (Xu and Huang). 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 $I_{ii} - \tilde{A}_{ii}$. Using the Taylor expansion of $\left(I_{ii} - \tilde{A}_{ii}\right)^{-1} \approx I_{ii} + \tilde{A}_{ii}$, we can use the following to approximate V_{loso}:

$$V^*_{loso}\left(\lambda\right) = \frac{1}{N}||\left(I - \tilde{A}_{\lambda,\boldsymbol{\theta}}\right)\tilde{Y}||^2 + \frac{2}{N}\sum_{i=1}^{N}\hat{\tilde{e}}_i'\tilde{A}_{ii}\hat{\tilde{e}}_i, \tag{3.56}$$

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 Equation 3.46 permits the straightforward evaluation of both scores $U(\lambda, \boldsymbol{\theta})$ and $V^*_{loso}(\lambda, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_g)'$ denotes the vector of smoothing parameters associated with each RK. In this section, we discuss a algorithm to minimize the unbiased risk estimate $U(\lambda, \boldsymbol{\theta})$ with respect to λ and $\boldsymbol{\theta}$ hidden in $Q = \sum_{\beta=1}^{q} \theta_\beta Q_\beta$, where the (i, j) entry of Q_β is given by $R_\beta(\boldsymbol{v}_i, \boldsymbol{v}_j)$. 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 procedue 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) (which is the basis for the algorithm which follows) and Wood (2004). 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 step 1 follows immediately from the expression for the smoothing matrix. Step 2 requires evaluating 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. With examination of U and V^* and 3.46, it is immediate that the $\theta_\beta \tilde{Q}_\beta$ are what matter in determining the minimum. Multiplying the $\tilde{Q}_\beta$ by any positive constant leaves the θ_β subject to rescaling, though the problem itself is unchanged by scale transformations. 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. In addition, optimizing with respect to $\boldsymbol{\kappa}$ 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 parameters, the outer loop minimizes U with respect to smoothing parameters via quasi-Newton iteration of Dennis Jr and Schnabel (1996), as implemented in the `nlm` function in `R`. The inner loop then minimizes ℓ_λ with fixed tuning parameters via Newton iteration. Fixing the θ_βs in $J(\phi^*) = \sum_\beta \theta_\beta^{-1} J_\beta\left(\phi_\beta^*\right)$, the outer loop with a single λ is straightforward.

Algorithm 1

Initialization:

Set $\Delta\boldsymbol{\kappa} := 0;\ \boldsymbol{\kappa}_{-} := \boldsymbol{\kappa}_0;\ V_{-} = \infty;$ (or $M_{-} = \infty$)

Iteration:

while not converged **do**

For current value $\boldsymbol{\kappa}^* = \boldsymbol{\kappa}_{-} + \Delta\boldsymbol{\kappa}$, compute $Q_\theta^* = \sum_{\beta=1}^{g} \theta_\beta^* Q_\beta$ and scale so that tr (Q_β) is fixed.

Compute $\tilde{A}_{\lambda,\boldsymbol{\theta}}(\lambda|\boldsymbol{\theta}^*) = \tilde{A}_{\lambda,\boldsymbol{\theta}}(\lambda, \exp(\boldsymbol{\kappa}^*))$.

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

Set $U_* := \min_{\lambda} Y(\lambda|\boldsymbol{\kappa}^*)$

if $U^* > U_{-}$ **then**

Set $\Delta\boldsymbol{\kappa} := \Delta\boldsymbol{\kappa}/2$

Go to (1).

else

Continue

end if

Evaluate gradient $\mathbf{g} = (\partial/\partial\boldsymbol{\kappa})\, U(\boldsymbol{\kappa}|\lambda)$

Evaluate Hessian $H = (\partial^2/\partial\boldsymbol{\kappa}\partial\boldsymbol{\kappa}')\, U(\boldsymbol{\kappa}|\lambda)$.

Calculate step $\Delta\boldsymbol{\kappa}$:

if H positive definite **then**

$\Delta\boldsymbol{\kappa} := -H^{-1}\mathbf{g}$

else

$\Delta\boldsymbol{\kappa} := -\tilde{H}^{-1}\mathbf{g}$, where $\tilde{H} = \text{diag}(\boldsymbol{\epsilon})$ is positive definite.

end if

end while

Calculate optimal model:

if $\Delta\kappa_\beta < -\gamma$, for γ large **then**

Set $\kappa_{*\beta} := -\infty$

end if

Compute $Q_\theta^* = \sum_{\beta=1}^{g} \theta^* \beta Q_\beta$;

Calculate $\begin{bmatrix} d \\ c \end{bmatrix} = \tilde{C}^{-1}\tilde{C}^{-T} \begin{bmatrix} \tilde{B}' \\ \tilde{Q}_*^{\theta'} \end{bmatrix} \tilde{Y}$

Calculation of the gradient $\boldsymbol{g}$ and Hessian H mirror the details in Gu and Wahba (1991), replacing the null basis matrix B and representer matrix Q with $D^{-1}XB$ and $D^{-1}XB$, respectively. They also present details on convergence criteria based on those suggested in Gill et al. (1981), who also present detailed discussion of the Newton method based on the Cholesky decomposition

necessary for calculating the update direction for $\boldsymbol{\kappa}$. The step in 21 returns a descent direction even when H is not positive definite by adding positive mass to the diagonal elements of H if necessary to produce $\tilde{H} = G'G$ where G is upper triangular. See Gill et al. (1981) 4.4.2.2 for details.

The unbiased risk estimate $U(\lambda, \boldsymbol{\theta})$ is fully parameterized by

$$(\lambda_1, \ldots, \lambda_q) = \left(\lambda\theta_1^{-1}, \ldots, \lambda\theta_q^{-1}\right), \tag{3.57}$$

so the smoothing parameters $(\lambda, \theta_1, \ldots, \theta_q)$ over-parameterize the score, which is the reason for scaling the trace of Q_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 $\breve{\theta}_\beta^{-1} \propto \text{tr}\left(\tilde{Q}_\beta\right)$, minimize $U(\lambda)$ with respect to λ to obtain $\breve{\phi}$.

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

The first pass allows equal opportunity for each penalty to contribute to the GCV score, 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: the penalty is of the form

$$J() = \sum_{\beta=1}^{q} \theta_\beta^{-1} \langle \phi, \phi \rangle_\beta$$

After the first pass, the initial fit $\breve{\phi}$ reveals where the structure in the true ϕ lie in terms of the components of the subspaces $\mathcal{H}_\beta$. Less penalty should be applied to terms exhibiting strong signal.

3.3 A smoothing spline model for the innovation variances

Once we have an initial estimate of the generalized autoregressive coefficient function, ϕ, we can use the model residuals to estimate the innovation variance function $\sigma^2(t)$. We use the same estimation approach as outlined in Section 3.1. Fixing $\phi = \phi^*$ for given estimate ϕ^*, the negative log likelihood of the data $Y_1, \ldots, Y_N$ is satisfies

$$-\ell\left(Y_1, \ldots, Y_N, \phi, \sigma^2\right) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{m_i}\log\sigma_{ij}^2 + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{\epsilon_{ij}^2}{\sigma_{ij}^2}; \tag{3.58}$$

where $\epsilon_{ij} = y_{ij} - \sum_{k<j}\phi_{ijk}^* y_{ik}$. Let

$$\text{RSS}(t) = \sum_{i,j:t_{ij}=t}\left(y_{ij} - \sum_{k<j}\phi_{ijk}y_{ik}\right)^2 \tag{3.59}$$

denote the squared residuals for the observations y_{ij} having corresponding measurement time $t_{ij} - t$. Then $\text{RSS}(t)/\sigma^2(t) \sim \chi^2_{df_t}$, 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 3.58 is that of a variance model with the ϵ_{ij}^2 serving as the response. This corresponds to a generalized linear model with gamma errors and known scale parameter equal to 2. Let $z_{ij} = \epsilon_{ij}^2$, and let $Z_i = \left(z_{i1}, z_{i,m_i}\right)'$ denote the vector of residuals for the i^{th} observed trajectory. The Gamma distribution is parameterized by shape parameterα and scale parameter β, where the mean of the distribution given by $\mu = \alpha\beta$. Reparameterizing the Gamma likelihood in terms of (α, μ) and dropping terms that don't involve $\mu(\cdot)$ gives

$$-\ell(z, \mu, \alpha) \propto \alpha\left[\frac{z}{\mu} + \log\mu\right] \tag{3.60}$$

$$= \alpha\left[ze^{-\eta} + \eta\right], \tag{3.61}$$

where α^{-1} is the dispersion parameter and $\eta = \log\mu$. Letting μ_{ij} denote $E[z_{ij}] = \sigma_{ij}^2$, the log likelihood of the working residuals becomes

$$-\ell\left(Z_1,\ldots,Z_N,\phi,\sigma^2\right)=\sum_{i=1}^{N}\sum_{j=1}^{m_i}\log\mu_{ij}+\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{z_{ij}}{\mu_{ij}}, \tag{3.62}$$

which we can see coincides with a Gamma dsitribution with scale parameter $\alpha = 2$. Smoothing spline ANOVA models for exponential families have been studied extensively (Wahba et al. (1995), Wang (1997), Gu (2013)). Parallel to the penalized sums of squares for ϕ (**??**), we can append a smoothness penalty to obtain the penalized likelihood for $\eta(t) = \log\sigma^2(t)$:

$$-\ell\left(Z_1,\ldots,Z_N,\phi,\sigma^2\right)+=\sum_{i=1}^{N}\sum_{j=1}^{m_i}\eta_{ij}+\sum_{i=1}^{N}\sum_{j=1}^{m_i}z_{ij}e^{-\eta_{ij}}+\lambda J(\eta), \tag{3.63}$$

noindent for $\eta\in\mathcal{H}=\oplus_{\beta=0}^{q}\mathcal{H}_\beta$, where the penalty J can be written as a square norm and decomposed as in (3.27), with

$$J(\kappa)=\langle\eta,\eta\rangle=\sum_{\beta=1}^{q}\theta_\beta^{-1}\langle\eta,\eta\rangle_\beta.$$

The $\langle\cdot,\cdot\rangle_\beta$ are inner products in $\mathcal{H}_\beta$ having reproducing kernels $Q_\beta(t,t')$. The penalty $J(\kappa)$ is an inner product in $\oplus_{\beta=0}^{q}\mathcal{H}_\beta$ with reproducing kernel $\sum_{\beta=1}^{q}\theta_\beta Q_\beta(t,t')$ and null space $\mathcal{N}_J=\mathcal{H}_0$. The first term in (3.63) serves as a measure of the goodness of fit of κ to the data, and only depends on κ through the evaluation functional $[t_{ij}]\,\kappa$. So the argument justifying the form of the minimizer in (**??**) applies, and the minimizer of the penalized likelihood has the form

$$\eta(t)=\sum_{\nu=1}^{d_0}d_\nu\kappa_\nu(t)+\sum_{i=1}^{|\mathcal{T}|}c_iQ_J(t,t_i), \tag{3.64}$$

where $\mathcal{T}=\bigcup_{j=1}^{N}\bigcup_{k=1}^{m_i}t_{jk}$ denotes the unique values of the observations times pooled across subjects, where $\{\kappa_\nu\}_{\nu=1}^{d_0}$ is a basis for the null space $\mathcal{N}_J=\mathcal{H}_0$.

Standard theory for exponential families gives us that the functional

$$
\begin{aligned}
L(\eta) &= -\sum_{i=1}^{N}\sum_{j=1}^{m_i}\left[z_{ij}\eta(t_{ij}) - b(\eta(t_{ij}))\right] \\
&= -\sum_{i=1}^{N}\sum_{j=1}^{m_i}\left[z_{ij}\eta(t_{ij}) - b(\eta(t_{ij}))\right]
\end{aligned}
\tag{3.65}
$$

is continuous and convex in $\eta \in \mathcal{H}$. We assume that the $|V| \times d_0$ matrix B which has i-ν^{th} element $\eta_\nu(\boldsymbol{v}_i)$ is full column rank, so that $L(f)$ is strictly convex in $\mathcal{H}$ and the minimizer of (3.63) uniquely exists. See Wahba et al. (1995).

For fixed λ and θ_β, which may be hidden in J, the penalized log likelihood (3.63) is convex in η, so that the minimizer can be computed via Newton iteration. Let

$$
\begin{aligned}
u_{ij} &= -z_{ij} + b'(\tilde{\eta}(t_{ij})) = -z_{ij} + \tilde{\mu}(t_{ij}), \text{ and} \\
\tilde{\omega}_{ij} &= b''(\tilde{\eta}(t_{ij})) = \tilde{v}(t_{ij}).
\end{aligned}
$$

The quadratic approximation of $-z_{ij}\eta(t_{ij}) + b(\eta(t_{ij}))$ at $\tilde{\eta}(t_{ij})$ is given by

$$
\begin{aligned}
&-y_{ij}\tilde{\eta}(t_{ij}) + b(\tilde{\eta}(t_{ij})) + \tilde{u}_{ij}\left[\eta(t_{ij}) - \tilde{\eta}(t_{ij})\right] + \frac{1}{2}\tilde{\omega}_{ij}\left[\eta(t_{ij}) - \tilde{\eta}(t_{ij})\right]^2 \\
&\qquad = \frac{1}{2}\tilde{\omega}_{ij}\left[\eta(t_{ij}) - \tilde{\eta}(t_{ij}) + \frac{\tilde{u}_{ij}}{\tilde{\omega}_{ij}}\right]^2 + C_{ij},
\end{aligned}
$$

where C_{ij} is independent of $\tilde{\eta}(t_{ij})$. The Newton iteration uses the minimizer of the penalized weights sums of squares

$$
\sum_{i=1}^{N}\sum_{j=1}^{m_i}\tilde{\omega}_{ij}\left(\tilde{y}_{ij} - \eta(t_{ij})\right)^2 + \lambda J(\eta) \tag{3.66}
$$

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

3.4 Smoothing parameter selection for exponential families

The gamma penalized log likelihood (3.64) is non-quadratic, so η_λ must be computed using iteration even for fixed smoothing parameters. A typical choice for method of smoothing parameter selection when data are generated from a distribution belonging to exponential families is performance-oriented. The follow section provides a brief overview of the the performance-oriented iteration, specifically for selecting the optimal degree of smoothing for σ^2. 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

$$\begin{aligned} \text{KL}(\eta, \eta_\lambda) &= E_\lambda \left[Z(\eta - \eta_\lambda) - (b(\eta) - b(\eta_\lambda)) \right] / a(\phi) \\ &= \left[b'(\eta)(\eta - \eta_\lambda) - (b(\eta) - b(\eta_\lambda)) \right] / a(\phi), \end{aligned} \tag{3.67}$$

For the gamma dsitribution, letting $\eta = \log \mu$, the KL distance simplifies to

$$-\mu \left(e^{-\eta} - e^{-\eta_\lambda} \right) - (\eta - \eta_\lambda).$$

The KL distance is not symmetric, so sometimes people opt for its symmetrized version:

$$\begin{aligned} \text{SKL}(\eta, \eta_\lambda) &= \text{KL}(\eta, \eta_\lambda) + \text{KL}(\eta_\lambda, \eta) \\ &= \left(b'(\eta) - b'(\eta_\lambda) \right)(\eta - \eta_\lambda) / a(\phi), \\ &= (\mu - \mu_\lambda)(\eta - \eta_\lambda) / a(\phi), \end{aligned} \tag{3.68}$$

A natural choice of loss function for measuring the performance of an estimator $\eta_\lambda(t)$ of $\eta(t)$ is the symmetrized Kullback-Leibler distance averaged over the observed time points $t_{11}, \ldots, t_{N,m_N}$:

$$L(\eta, \eta_\lambda) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{m_i} \sum_{j=1}^{m_i} (\mu(t_{ij}) - \mu_\lambda(t_{ij}))(\eta(t_{ij}) - \eta_\lambda(t_{ij})), \tag{3.69}$$

For the Gamma distribution, this reduces to

$$L(\eta, \eta_\lambda) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{N} \sum_{j=1}^{m_i} \left(\frac{\mu(t_{ij})}{\mu_\lambda(t_{ij})} - \frac{\mu_\lambda(t_{ij})}{\mu(t_{ij})} - 2 \right). \tag{3.70}$$

The ideal smoothing parameters are those which minimize (3.70). The performance-oriented iteration operates on a alternative expression of the symmetrized Kullback-Leibler loss. The mean value theorem gives us that (3.70) can be written

$$L_\omega(\eta, \eta_\lambda) = L(\eta, \eta_\lambda) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{N} \sum_{j=1}^{m_i} \omega^*(t_{ij}) (\eta(t_{ij}) - \eta_\lambda(t_{ij}))^2, \tag{3.71}$$

where $\omega^*(t_{ij}) = b''(\eta^*(t_{ij}))$ and $\eta^*(t_{ij})$ is a convex combination of $\eta(t_{ij})$ and $\eta_\lambda(t_{ij})$. One can construct an unbiased risk estimate under the weighted loss, L_ω, using re-weighted observations. Letting $Z_{i\omega} = W_i Z_i$, where W_i is the $m_i \times m_i$ diagonal matrix having diagonal entries $\omega^*(t_{i1}), \ldots, \omega^*\left(t_{i,m_i}\right)$, an unbiased estimate of relative loss is given by

$$U_\omega(\lambda) = L(\eta, \eta_\lambda) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{N} \sum_{j=1}^{m_i} \omega^*(t_{ij}) (\eta(t_{ij}) - \eta_\lambda(t_{ij}))^2. \tag{3.72}$$

See Gu (2013), Theorem 5.2. To find the optimal value of the smoothing parameter, the performance-oriented iteration tracks loss $L(\eta, \eta_\lambda)$ 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. The performance-oriented iteration works on (3.70) and updates the smoothing parameters updated according to $U_\omega(\lambda)$. 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 $U_{\omega}(\lambda)$. 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 (3.66) closely approximate the penalized likelihood functional (3.65) for smoothing parameters close to $\left(\lambda^{*}, \theta_{\beta}^{*}\right)$, then the $\eta_{\lambda,\eta^{*}}$s are, in turn, hopefully close approximations to the η_{λ}s. Thus, through indirect comparison η^{*} is perceived to be better performing among the other η_{λ}s in the neighborhood.

An alternative to the performance-oriented iteration is to choose the optimal smoothing parameters by comparing candidate η_{λ}s directly; the generalized approximate cross validation (GACV) score 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 (see (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.

Chapter 4: A P-spline model for the Cholesky decomposition

The reproducing kernel Hilbert space methods presented in Chapter 3 comprise a small slice of the existing methods for smoothing noisy data. In addition to smoothing splines, regression splines Eubank (1999) and B-splines De Boor et al. (1978) have been widely used for nonparametric function estimation. We consider a second approach to smoothing ϕ and σ^2 which is based on the penalized B-splines, or P-splines, of Eilers and Marx (1996). P-splines 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. In addition to this flexibility, P-splines, which are a straightforward extension of (generalized) linear regression models, exhibit a number of attractive properties. Fitted P-splines exhibit no boundary effects, conserve moments of the data, and in the limit, approach polynomial curves. Perhaps equally important is the relatively inexpensive computation necessary for model fitting, including smoothing parameter selection.

4.1 Tensor product B-splines for multidimensional smoothing

B-splines are piecewise polynomial functions, where the piecewise polynomials are joined at certain values of the domain called knots. Given a set of knots, B-splines can be easily computed

recursively for any polynomial degree (see De Boor et al. (1978), 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 propose alternative an approach to nonparametric smoothing based on based on finite difference penalties, which is simple to compute. Their approach circumvents both the choice of knot specification as well as any complexity associated with constructing traditional penalty matrices by omitting derivatives and integrals altogether. This approach achieves smoothness in fitted functions in two ways: By purposefully overfitting the smooth coefficient vectors using a B-spline basis with a large number of equally spaced knots, one can avoid the difficulty of choosing the optimal set of knots. Augmenting the goodness of fit measure (which is, in our case, the log likelihood) with a difference penalty allows for penalized maximum likelihood estimation as in Chapter 3, which prevents overfitting and accommodates a potentially ill-conditioned fitting procedure.

Analogous to the smoothing spline representation (3.16), 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), 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

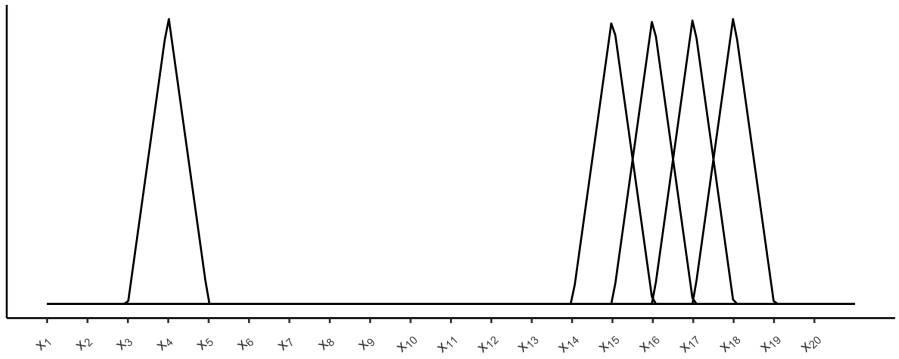
$$\begin{aligned} B_{i,1}(x) &= \begin{cases} 1, & t_i \leq x < t_{i+1} \\ 0, & otherwise \end{cases} \\ B_{i,k}(x) &= \frac{x - t_i}{t_{i+k-1} - t_i} B_{i,k-1}(x) + \frac{t_{i+k} - x}{t_{i+k} - t_{i+1}} B_{i+1,k-1}(x). \end{aligned} \tag{4.1}$$

Figure 4.1 shows two sets of B-splines; the top facet displays linear B-splines and the bottom displays B-splines of degree 2. A single isolated B-spline is shown on the left side of the axis in each panel. In Figure 4.1a, the single B-spline of degree 1 consists of two linear pieces: one piece from x_1 to x_2, and the other from x_2 to x_3, which are the knots that define its support. In the right part of Figure 4.1a, 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 4.1b, we can see that the extent to which neighboring B-splines overlap depends on the polynomial degree of the basis. These simple example illustrate some generate properties of a B-spline of degree k:

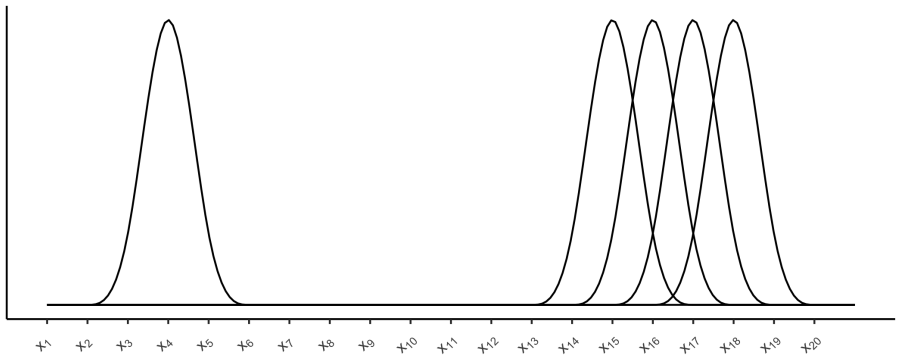
- It is constructed using $k + 1$ polynomial pieces, each of degree q,
- The polynomial pieces join at k inner knots.
- At the inner knots, its derivatives are continuous up to order $k - 1$.
- The B-spline has positive support spanned by $k + 2$ knots; everywhere else it is zero.
- With the exception of at the boundaries, it overlaps with $2k$ polynomial pieces of its neighbors.
- At a given point in the domain, v, $k + 1$ B-splines take nonzero values.

Figure 4.2 shows a set of B-spline basis functions of degree 2. Each basis function consists of three quadratic pieces, joined at two knots. At the joining points, not only the ordinates of the polynomial pieces match, but their first derivatives are also equal (but not their second derivatives). De Boor et al. (1978) presented an algorithm to compute B-splines of any degree from B-splines of lower degree. A formal definition of the i^{th} B-spline of order k for a fixed knot sequence is given in Appendix B Definition B.1.1. Additional mathematical properties of B-splines which are pertinent to the presentation of P-spline smoothing can also be found in Appendix B.

Figure 4.1: *On the left: a single, isolated B-spline basis function, and on the right: several overlapping B-splines.*



(a) *of degree 1*



(b) *of degree 2*

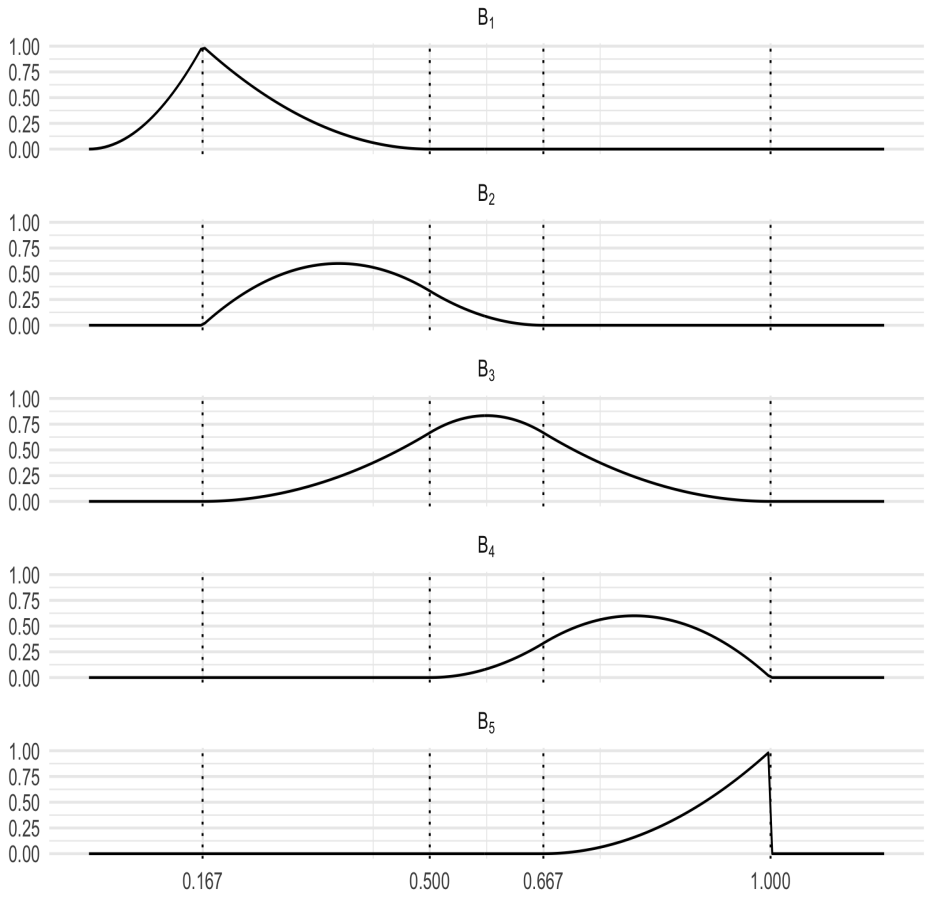


Figure 4.2: *A set of parabolic B-splines corresponding to knot sequence* $\left\{\frac{1}{6}, \frac{1}{2}, \frac{2}{3}, 1\right\}$

B-splines make attractive basis functions for nonparametric regression; a linear combination of, say, cubic B-splines gives a smooth curve. Once a B-spline basis is computed, their application is no more difficult than polynomial regression, and nearly seamless 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

$$B_{l,1}(l), \ldots, B_{l,K_1}(l) \text{ and } B_{m,1}(m), \ldots, B_{m,K_2}(m)$$

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, of course, using different numbers of knots. 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. Figure 4.4 shows a thinned tensor product basis $\{T_{kk'}\}$; a portion of the basis was omitted to eliminate overlapping of the basis functions so that the reader can identify individual tensor products. Each "hill" in Figure 4.4 is associated with an unknown coefficient θ_{ij} which determines the height of the hill. For a given knot grid, we can approximate a surface, ϕ, by

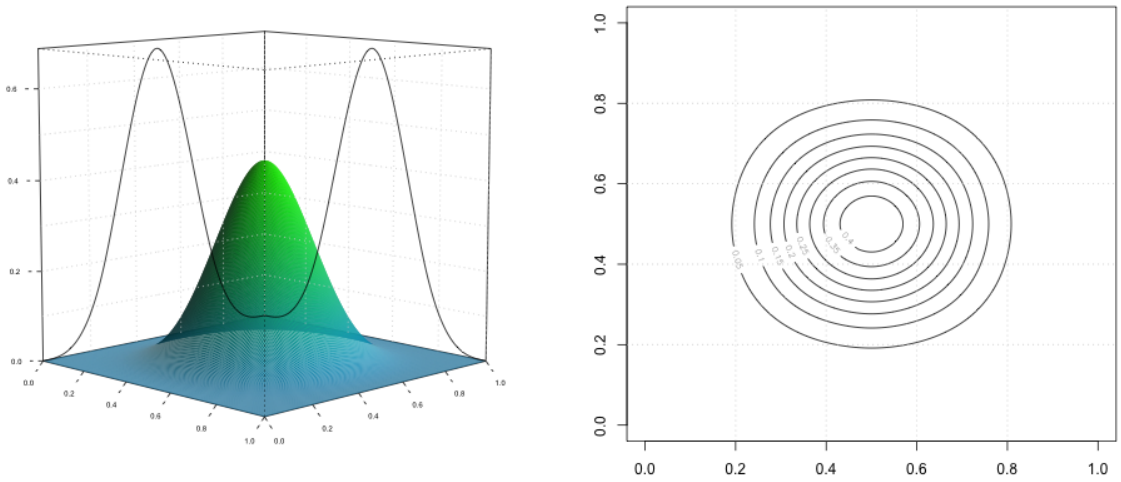
$$\phi(l,m) = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} \theta_{ij} B_{l,i}(l)\, B_{m,j}(m)\,. \tag{4.2}$$

By using rich B-spline bases for l and m alongside discrete difference penalties on the spline coefficients, we can achieve as much smoothness of the fitted function in both the l and m dimensions as desired. Fixing σ^2 as in (3.10), we define the estimator of ϕ as the minimizer of

$$-2\ell_\phi + \lambda J(\phi) = \sum_{i=1}^{N}\sum_{j=2}^{m_i} \sigma_{ij}^{-2}\left(y_{ij} - \sum_{k<j}\phi(\boldsymbol{v}_{ijk})\, y_{ik}\right)^2 + \lambda J(\phi)\,, \tag{4.3}$$

where $J(\phi)$ is a penalty on the roughness of the fitted function.

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



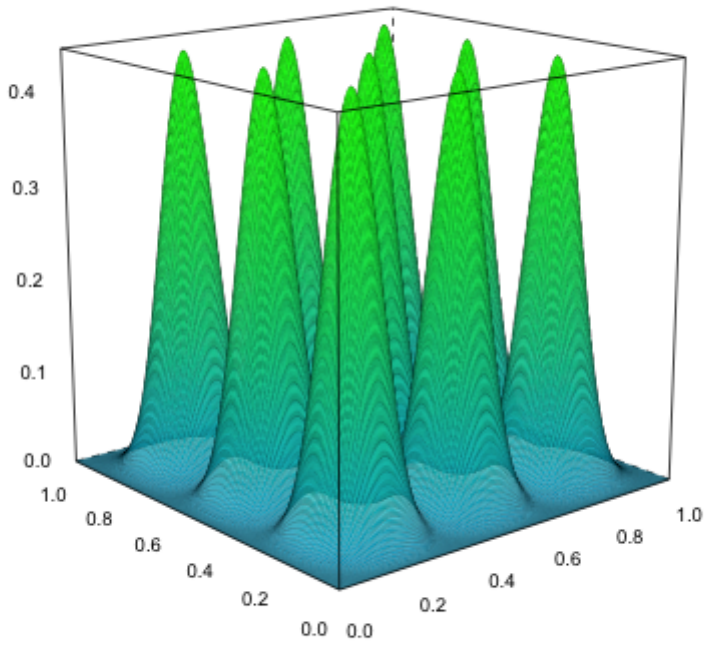


Figure 4.4: *A subset of a full bivariate basis of cubic B-splines*

4.2 Difference penalties

O'Sullivan (1986) was the first to propose using a rich B-spline basis alongside a penalty to restrict the flexibility of the fitted curve. Like Wahba (1990), he proposed appending the following penalty to the goodness of fit measure to enforce smoothness on the fitted curve:

$$J = \int_0^1 \left[\phi''\right]^2 . \tag{4.4}$$

For a function that can be written as linear combination of B-spline basis functions,

$$\phi(v) = \sum_{j=1}^{k} \theta_i B_j(v),$$

one can derive a banded matrix P using the properties of B-splines such that

$$J = \theta' P \theta$$

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

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

Wand and Ormerod (2008) extend O'Sullivan's work 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

$$\left(\phi_1, \frac{\phi_1 + \phi_2}{2}, \phi_2, \phi_2, \frac{\phi_2 + \phi_3}{2}, \phi_3, \ldots, \phi_{k+7}, \frac{\phi_{k+7} + \phi_{k+8}}{2}, \phi_{k+8}\right),$$

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

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

where $(\Delta\phi)_j = \phi_{j+1} - \phi_j$. 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 proposed replacing the derivative-based penalty (4.4) with a finite difference penalty on the B-spline coefficients. Using the properties of B-splines, it is straightforward to show that the difference penalty of order d is a good discrete approximation to the integrated square of the d^{th} derivative, so little is lost by using it in place of the derivative-based penalty. The difference penalty is simple to compute and can be handled mechanically for any order of the differences, and since it is easily appended to a goodness of fit measure (such as the log likelihood), it is feasible to evaluate the impact of different orders of the differences on the fitted model.

The d^{th} order difference penalty is given by

$$J_d(\phi) = \sum_{j=d}^{n} \left(\Delta^d \theta_j \right)^2, \tag{4.5}$$

Let D_d denote the matrix difference operator:

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

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,

$$\Delta^d \theta_j = \Delta \left(\Delta^{d-1} \theta_j \right).$$

Then, (4.5) can be written in terms of the squared norm of the difference operator applied to the vector of B-spline coefficients:

$$\begin{aligned} J_d(f) &= ||D_d\theta||^2 \\ &= \theta' P_d \theta \end{aligned} \tag{4.6}$$

where $P_d = D_d' D_d$. To examine the connection between the second-derivative penalty to the penalty on second-order differences of the B-spline coefficients, we only need to employ straightforward calculus and exploit the recursive property of the B-spline basis functions. Consider a cubic B-spline

$$f = \sum_{i=1}^{k} \theta_i B_{i,3}.$$

The traditional smoothness penalty applied to f is given by

$$\int_0^1 [f'']^2 = \int_0^1 \left\{ \sum_{j=1}^{k} \theta_j B''_{j,3} \right\}^2 .$$

The derivative properties of B-splines permits this to be written as

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

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

$$\begin{aligned} \int_0^1 [f'']^2 &= \int_0^1 \left[\left(\sum_{j=1}^{k} \Delta^2\theta_j B_j \right)^2 + 2 \sum_{j} \Delta^2\theta_j \Delta^2\theta_{j-1} B_{j,1} B_{j-1,1} \right] \\ &= \sum_{j=1}^{k} \left(\Delta^2\theta_j\right)^2 \int_0^1 B_{j,1}^2 + 2 \sum_{j=1}^{k} \Delta^2\theta_j \Delta^2\theta_{j-1} \int_0^1 B_{j,1} B_{j-1,1} \end{aligned} \tag{4.7}$$

which can be written as

$$\int_0^1 [f'']^2 = c_1 \sum_{j=2}^{k} \left(\Delta^2 \theta_j\right)^2 + c_2 \sum_{j=3}^{n} \Delta^2 \theta_j \Delta^2 \theta_{j-1} \tag{4.8}$$

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

$$\begin{aligned} c_1 &= \int_0^1 B_{j,1}^2 \\ c_2 &= \int_0^1 B_{j,1} B_{j-1,1}. \end{aligned} \tag{4.9}$$

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 4.5 and the sum of the cross products of neighboring second differences. The second term in 4.8 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 4.8 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. We can seamlessly augment the likelihood with the difference penalty to achieve smooth fitted functions without the complexity posed by the derivative-based penalty.

A smoother sequence of coefficients leads to a smoother curve, as illustrated in Figure 4.5. 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. Figure 4.6 illustrates this utility of the penalty for simulated data; there are $p = 10$ observations and 60 cubic B-splines. Unless computational

constraints are of concern, which is possible with large models, it is prudent to use even more B-splines.

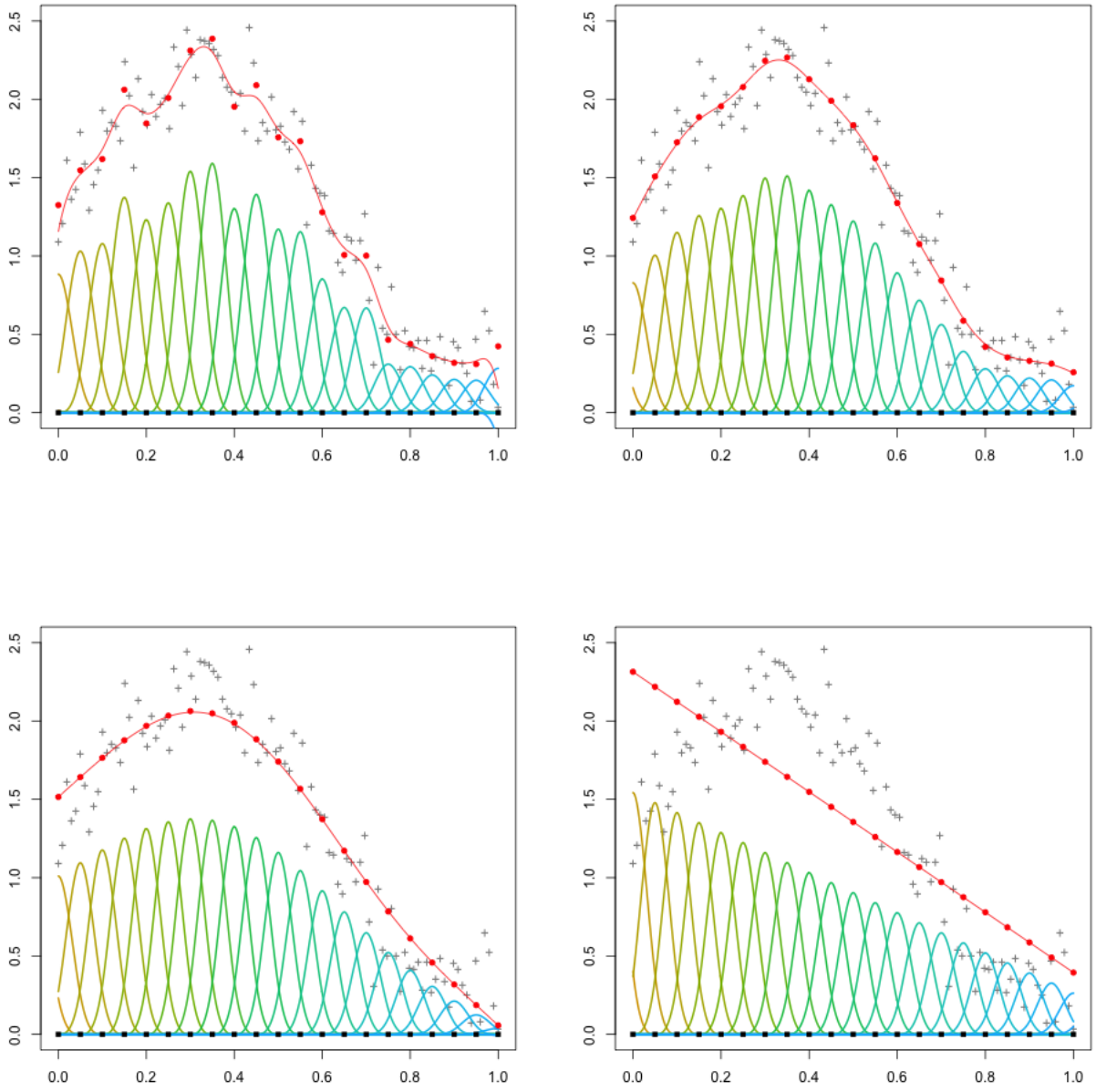


Figure 4.5: *Illustration of the impact of the second order difference penalty. The number of B-splines used is the same in each plot, with the value of the penalty parameter increasing from left to right and top to bottom across each plot. The fitted curve in the upper left plot is the most "wiggly" of any of the fits, as the penalty plays the weakest roll in the fitted coefficients there. 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.*

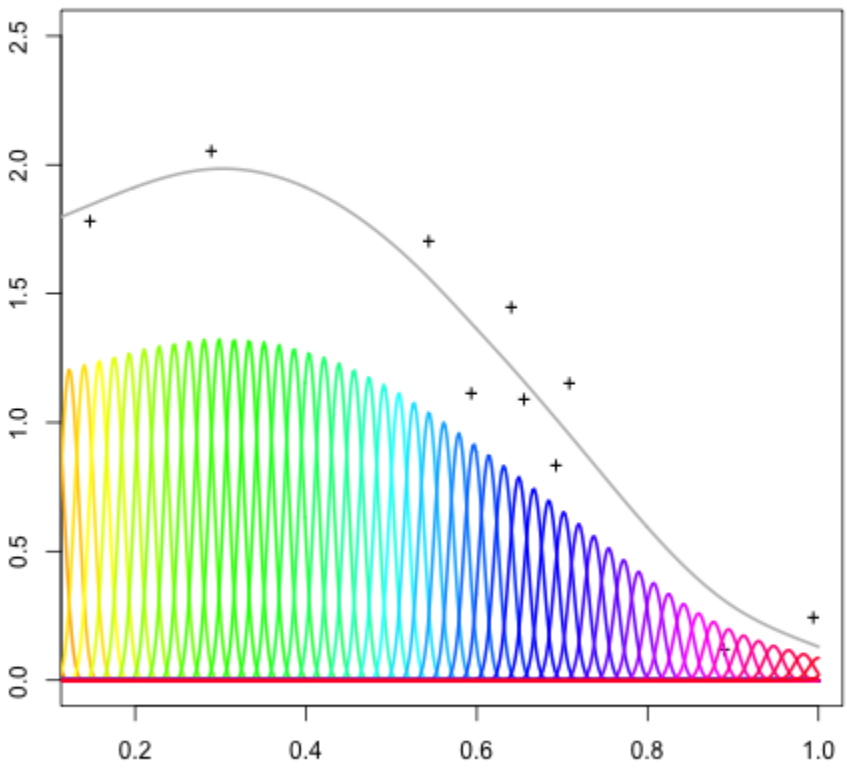
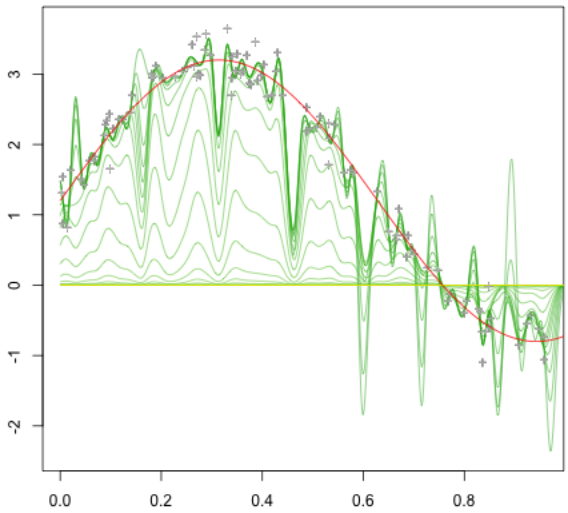


Figure 4.6: *P-spline smoothing of 10 observations using 60 B-spline basis functions.*

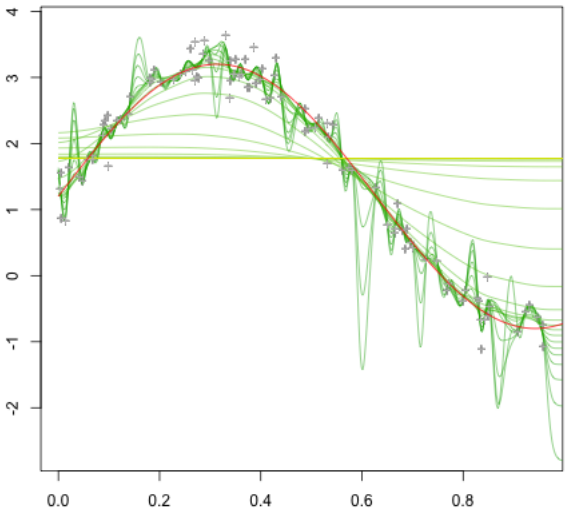
P-splines can fit polynomial data exactly. If the true underlying function to be estimation is a polynomial in l of degree k, then B-splines of degree k or higher will fit the data perfectly. The proof of this requires a bit of mathematical labor and is left to Appendix B. Additionally, the limiting P-spline fit approaches a polynomial as the smoothing parameter tends to infinity. As $\lambda \rightarrow \infty$, under a difference penalty of order d, the fitted function will approach a polynomial of degree $d - 1$ as long as the degree of the B-splines is greater than or equal to k. Figure 4.7 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 $\sum_{j=3}^{k} \left(\Delta^2 \theta_j\right)^2$ 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.

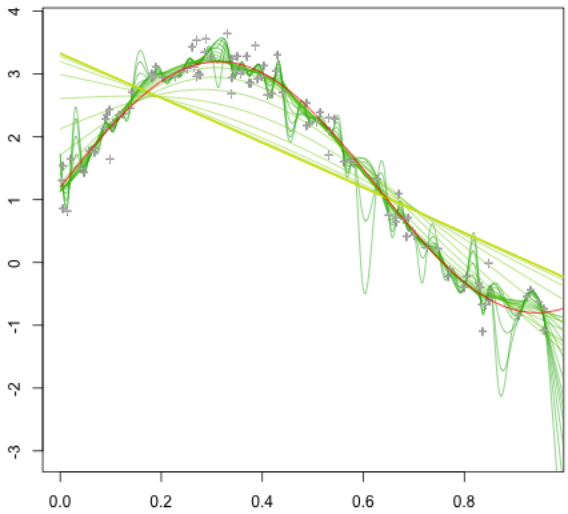
Figure 4.7: *Illustration of the impact of the order of the difference penalty. The number of B-splines used is the same in each plot, with the penalty parameter varying from across the same grid of values. The fitted curves in the upper left plot correspond to the difference penalty of order* 0*, where* $|D_0\theta|^2 = \sum_i \theta_i^2$*, analogous to ridge regression using the B-spline basis as regression covariates. The fitted curves approach polynomials of degree* $d-1$ *as* $\lambda \to \infty$*.*



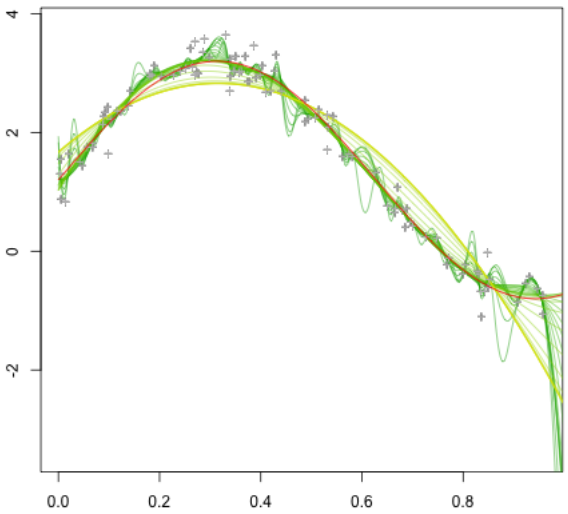
(a) $d = 0$



(b) $d = 1$



(c) $d = 2$



(d) $d = 3$

4.3 The P-spline estimator of the generalized autoregressive coefficient function

To extend the use of the difference penalty (4.5) to the bivariate setting, the only necessary modification to the one-dimensional differencing procedure is the addition of a second difference penalty, one for each variable l and m. The explicit form of the minimizer of the penalized log likelihood is available, but for exposition, we first need to establish some notation. The smooth varying coefficient function ϕ evaluated at a $n_1 \times n_2$ grid of points over the $l \times m$ plane may be written

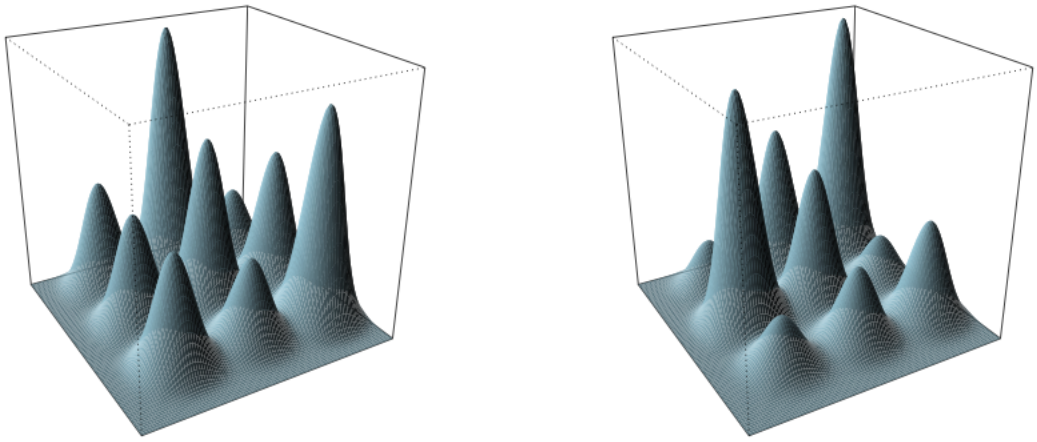
$$\boldsymbol{\phi} = B_2 \Theta B_1'$$

where B_1 is the $n_1 \times k_1$ matrix with i^{th} column equal to the i^{th} B-spline basis function for l evaluated at the grid points $l_1, \ldots l_{n_1}$, B_2 is the $n_2 \times k_2$ matrix with j^{th} column equal to the j^{th} B-spline basis function for m evaluated at the grid points $m_1, \ldots m_{n_2}$. The matrix Θ denotes the $k_1 \times k_2$ matrix of tensor product coefficients, with elements θ_{ij}. Smoothing ϕ in the l direction can be achieved by applying a difference penalty to the rows of Θ, and similarly in the m direction by applying a difference penalty to the columns. We take ϕ_λ to be the minimizer of

$$\begin{aligned} -2\ell + J_\theta(\phi) = &\sum_{i=1}^{N} \sum_{j=2}^{m_i} \sigma^{-2}(t_{ij}) \left(y_{ij} - \sum_{k=1}^{j-1} \left(\sum_{r=1}^{k_1} \sum_{c=1}^{k_2} \theta_{rc} B_r(l_{ijk}) B_c(m_{ijk}) \right) y_{ik} \right)^2 \\ &+ \lambda_1 \sum_{r=1}^{k_1} |D_{d_l} \theta_{r\cdot}|^2 + \lambda_2 \sum_{c=1}^{k_2} |D_{d_m} \theta_{\cdot c}|^2. \end{aligned} \tag{4.10}$$

where $\theta_{r\cdot}$ and $\theta_{\cdot c}$ denote the r^{th} row and c^{th} column of Θ, respectively. The first term in 4.10 imposes a difference penalty of order d_l on the rows of the coefficient matrix while the second term places a difference penalty (of possible different order d_m) on the columns. We give each

direction its own smoothing parameter to permit anisotropic smoothing; however, one could opt to use a single smoothing parameter for both directions and dodge the added work of optimizing the amount of smoothing with two separate parameters. Figure 4.8 shows a simple demonstration of the result of heavy column penalization (left) and heavy row penalization (right) under a second order difference penalty on each row and each column for large values of the smoothing parameters λ_1 and λ_2. The figure demonstrates that the limiting behavior of each row (column) is linear, but the resulting surface may exhibit slope reversals from one row (column) to the next.



(a) *heavy column penalty* (b) *heavy row penalty*

Figure 4.8: *Nine cubic B-spline tensor products with heavy linear column penalty and heavy linear row penalty*

The penalized log likelihood is quadratic in $\theta = (\theta_{11}, \ldots, \theta_{k_1,k_2})'$, and computation is quite simple if we express the coefficient matrix in "unfolded" notation. This allows us to express the

varying coefficient function at the observed within-subject pairs of observation times as in the usual multiple regression form:

$$\text{vec}\left\{\phi\left(\boldsymbol{v}\right)\right\} = B\theta$$

Stacking the columns of Θ gives the vectorized coefficient matrix $\theta = \text{vec}\left(\Theta\right)$. The $|V| \times k_1 k_2$ 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:

$$B = B_2 \square B_1 = \left(B_2 \otimes 1'_{k_2}\right) \odot \left(1'_{k_1} \otimes B_1\right). \tag{4.11}$$

The operator $\odot$ denotes the element-wise matrix product; 1_{k_1} (1_{k_2}) denotes the column vector of ones having length k_1 (k_2.) The operations in (4.11) construct B such that the i^{th} row of $B_2 \square B_1$ is the Kronecker product of the corresponding rows of B_2 and B_1. The penalty in (4.10) can also be compactly expressed:

$$\lambda_1 ||P_1\theta||^2 + \lambda_2 ||P_2\theta||^2$$

where $P_1 = I_{k_2} \otimes D'_{d_l} D_{d_l}$ and $P_2 = D'_{d_m} D_{d_m} \otimes I_{k_1}$. We define Y, the vector of stacked subject-specific vectors, as before, and the matrix X of autoregressive covariates as previously (3.29) so that (4.10) can be written in matrix form as

$$-2L + J_\theta\left(\phi\right) = \left(Y - XB\theta\right)' D^{-1}\left(Y - XB\theta\right) + \lambda_1 ||P_1\theta||^2 + \lambda_2 ||P_2\theta||^2, \tag{4.12}$$

with $\hat{\theta}$ solving the system of equations

$$\left[\left(XB\right)' D^{-1} XB + \lambda_1 P_1 + \lambda_2 P_2\right] \theta = \left(XB\right)' D^{-1} Y. \tag{4.13}$$

From (4.13), we note that the system of equations depends on basis coefficients remains fixed at $k_1 k_2$, 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_1 columns having length k_2. The vector of fitted values is given by

$$\begin{aligned} \hat{Y} &= X \left[(XB)' D^{-1} XB + \lambda_1 P_1 + \lambda_2 P_2 \right] (XB)' D^{-1} Y \\ &= AY \end{aligned} \tag{4.14}$$

where $A = X \left[(XB)' D^{-1} XB + \lambda_1 P_1 + \lambda_2 P_2 \right] (XB)' D^{-1}$ is the "smoothing" matrix, analogous to the smoothing matrix $\tilde{A}$ (3.46) 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.

This recipe for constructing a tensor product basis for ϕ is an easy and convenient way to construct a two-dimensional basis for a bivariate function with domain corresponding to the unit square. However, the domain autoregressive coefficient function, ϕ, lies on the lower triangle of the unit square $0 < s < t < 1$:

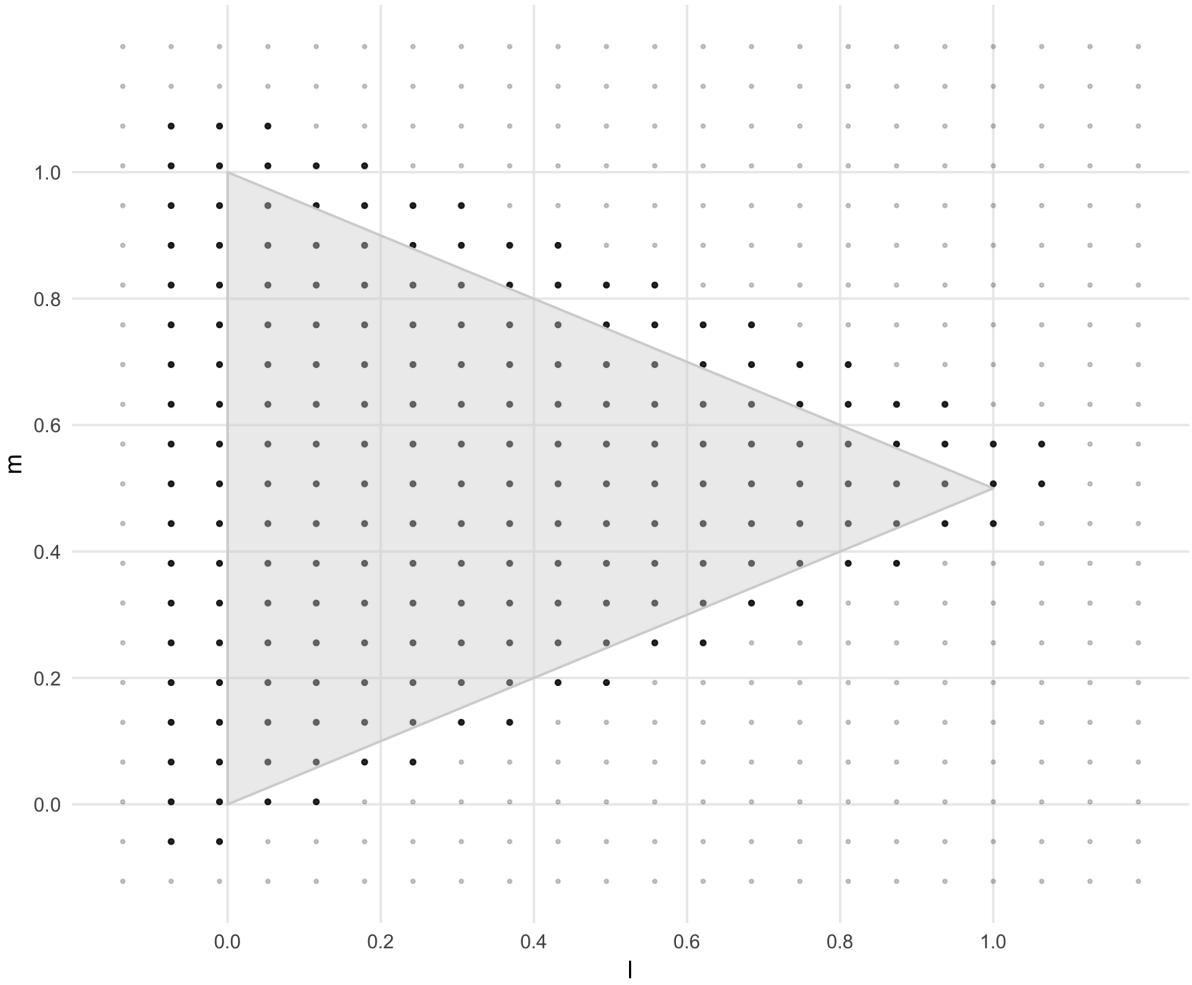


Figure 4.9: $\frac{l}{2} < m < 1 - \frac{l}{2}, \quad 0 < l < 1.$

When the tensor product basis is constructed on the regular grid defined by the cartesian product of the knots of the marginal bases B_1 and B_2, 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 (4.13) 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_1 and P_2 as needed. With the trimmed basis and penalties, optimization can be carried out as previously discussed. Alternatively, one could employ the use of multidimensional B-splines for the construction of the basis for $\boldsymbol{v} = (l, m)$. There is little about multidimensional splines in the Statistics literature - likely because the computational complexity associated with these methods, however, some in the field of computer graphics have proposed the use of their use for smoothing over arbitrary function domains, which are approximated by triangulations. See Dahmen et al. (1992) and Seidel (1991) for details.

4.4 Smoothing parameter selection

The limiting behaviour of H_λ

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

$$A_\lambda = X\big((XB)' D^{-1} XB + \lambda_1 P_1 + \lambda_2 P_2 \big)^{-1} (XB)' D^{-1} \tag{4.15}$$

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 (3.53) and its computationally efficient approximation (3.56) in Chapter 3, which rely directly on the smoothing matrix for calculation. The results in Xu et al. (2012) are basis agnostic, so we can employ the losoCV criterion for selecting P-spline smoothing parameters as in the smoothing spline setting by replacing $\tilde{A}_{\lambda,\boldsymbol{\theta}}$ with A_λ.

Summarizing the complexity of a fitted P-spline is non-trivial; it requires the simultaneous consideration of the smoothing parameters, the number of basis functions in the B-spline basis, and the order of the difference penalties. To assess model complexity, Eilers and Marx (1996) follow

Hastie and Tibshirani (1990), who use the trace of the smoothing matrix as an approximation of the *effective dimension* (ED) of a linear smoother. The effective (model) dimension is defined as

$$\text{ED} = \text{tr}\left(A_\lambda\right) = \text{tr}\Big(X\left((XB)'D^{-1}XB + \lambda_1 P_1 + \lambda_2 P_2\right)^{-1}(XB)'D^{-1}\Big) \tag{4.16}$$

The ED combines the effect of the smoothing parameter, the number of basis functions, and the differencing order, and it is easy to compute. When the number of basis functions is significantly smaller than the sample size, it is advantageous to use the cyclic property of the trace:

$$\begin{aligned}\text{tr}\left(A_\lambda\right) &= \text{tr}\Big(X\left((XB)'D^{-1}XB + \lambda_1 P_1 + \lambda_2 P_2\right)^{-1}(XB)'D^{-1}\Big)\\ &= \text{tr}\Big(\ (XB)'D^{-1}X\left((XB)'D^{-1}XB + \lambda_1 P_1 + \lambda_2 P_2\right)^{-1}\Big),\end{aligned}$$

which requires computing the trace of a $k_1k_2 \times k_1k_2$ matrix, which is computationally more economical when the total number of basis functions is smaller than the total number of observations. This approach to approximating the effective model dimension is also consistent with Ye (1998), who constructed a generalization of the concept of a model's degrees of freedom using the idea that the degrees of freedom can also be interpreted as the sum of the sensitivity of each fitted value with respect to the corresponding observed value.

Using the eigenstructure of the smoothing matrix, one can show that as the smoothing parameters tend to infinity, the effective dimension approaches $d_l + d_m$, the sum of the order of the differencing operators for l and m. Let

$$Q = (XB)'D^{-1}XB \qquad \text{and} \qquad Q_\lambda = \lambda_1 P_1 + \lambda_2 P_2.$$

Again using cyclic property of the trace, we can write

$$\begin{aligned}\text{tr}\left(A_\lambda\right) &= \text{tr}\left[\left(Q+Q_\lambda\right)^{-1}Q\right] \\ &= \text{tr}\left[Q^{1/2}\left(Q+Q_\lambda\right)^{-1}Q^{1/2}\right] \\ &= \text{tr}\left[\left(I+Q^{-1/2}Q_\lambda Q^{-1/2}\right)^{-1}\right]\end{aligned}$$

Finally we have that

$$\text{tr}\left(A_\lambda\right) = \text{tr}\left[\left(I+\lambda Q^{-1/2}Q_\lambda Q^{-1/2}\right)^{-1}\right] = \sum_{j=1}^{k_1k_2}\frac{1}{1+\lambda\gamma_j},$$

where γ_j, $j=1,\ldots,k_1k_2$ are the eigenvalues of $Q^{-1/2}Q_\lambda Q^{-1/2}$. The matrix constructed from the sum of the penalty terms, Q_λ, has exactly d_l+d_m eigenvalues equal to zero. Hence, $Q^{-1/2}Q_\lambda Q^{-1/2}$ has d_l+d_m eigenvalues equal to zero, so for large λ, only the d_l+d_m terms with $\gamma_j=0$ contribute to the sum which gives the trace of A_λ. Then

$$\lim_{\lambda\to\infty}\text{tr}\left(A_\lambda\right) = d_l+d_m.$$

This clearly shows that the effective dimension is always less than k_1k_2, the number of B-spline used in the regression basis; further, the effective dimension is always smaller than $\min\left(\sum_{i=1}^{N}m_i-N, k_1k_2\right)$. This is visually demonstrated in Figure 4.10, which displays impact of the smoothing parameter on the effective dimension of the P-spline fit to the simulated data shown in Figure 4.6. As λ increases, the effective dimension approaches the order of the difference penalty. Even for small λ, the effective dimension never exceeds the number of observations, so there are no issues when fitting P-splines with many more basis functions than observations.

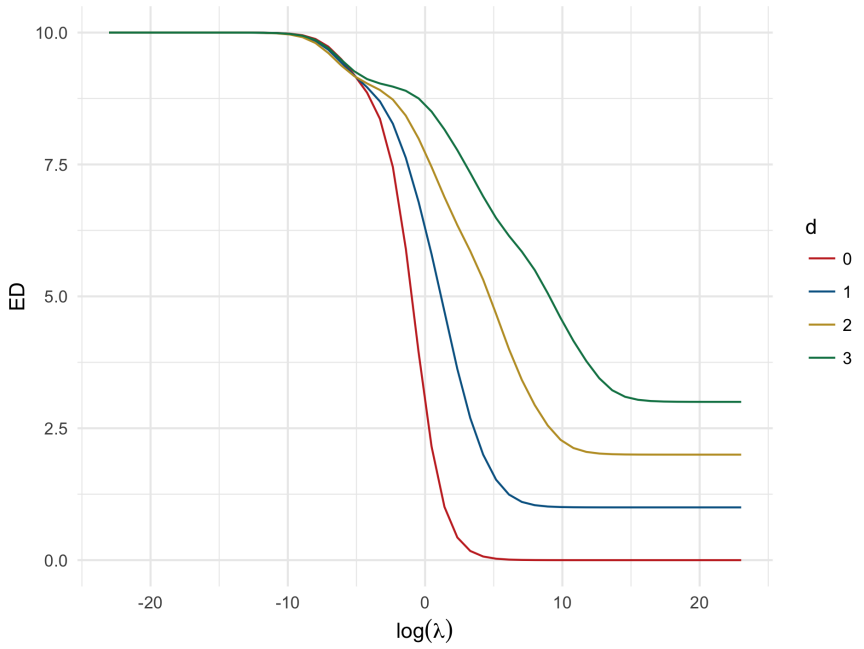


Figure 4.10: *The limiting behaviour of the trace of the smoothing matrix A_λ as the smoothing parameter increases for the P-spline fit to the 10 observations using 60 B-spline basis functions, shown in Figure 4.6. For weakly enforced smoothing, the effective dimension is equal to the number of observations, and as $\lambda \to \infty$, the effective dimension approaches the order of the difference penalty.*

The effective model dimension is closely connected to model selection criteria; Eilers and Marx (1996) propose the use of the Akaike information criterion (AIC) for selecting the optimal value of the smoothing parameters $\lambda = (\lambda_1, \lambda_2)$, which is equivalent to the unbiased risk estimator discussed in Chapter 3 under a Gaussian likelihood. For a detailed discussion of the connection between the unbiased risk estimate and AIC in the non-gaussian case, see Wood (2017), Chapter 4, Section 5. The same reference provides a detailed discussion of computational methods for minimizing the unbiased risk estimate with respect to multiple smoothing parameters. The algorithm shares the same basic structure as the one outlined in Section 3.2, with modifications on the derivatives and the

Hessians to account for the fact that the P-spline basis and penalty are constructed independently of one another.

TO DO: add short section including discussion of P-spline model for σ^2

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 does 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	
$\Sigma = \mathrm{I}$	$\phi(t,s) = 0, \quad 0 \le s < t \le 1,$ $\sigma^2(t) = 1, \quad 0 \le t \le 1.$
II: Linear varying coefficient function, constant innovation variances	
$\Sigma = T^{-1}DT'^{-1}$	$\phi(t,s) = t - \frac{1}{2}, \quad 0 \le t \le 1,$ $\sigma^2(t) = 0.1^2, \quad 0 \le t \le 1.$
III: Banded linear varying coefficient function, constant innovation variances	
$\Sigma = T^{-1}DT'^{-1}$	$\phi(t,s) = \begin{cases} t - \frac{1}{2}, & t - s \le 0.5 \\ 0, & t - s > 0.5 \end{cases},$ $\sigma^2(t) = 0.1^2, \quad 0 \le t \le 1.$
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	
$\Sigma = \sigma^2(\rho \mathrm{J} + (1-\rho)\mathrm{I}),$ $\rho = 0.7, \quad \sigma^2 = 1$	$\phi_{ts} = \frac{\rho}{1 + (t-2)\rho}, \; \begin{matrix} t = 2, \ldots, M, \\ s = 1, \ldots, t-1 \end{matrix}$ $\sigma_t^2 = \begin{cases} 1, & t = 1 \\ 1 - \frac{(t-2)\rho^2}{1+(t-2)\rho}, & t = 2, \ldots, M \end{cases}$

Figure 5.1 displays a two dimensional representation of each covariance matrix Σ and it's corresponding Cholesky factor T used in the simulation study. 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.

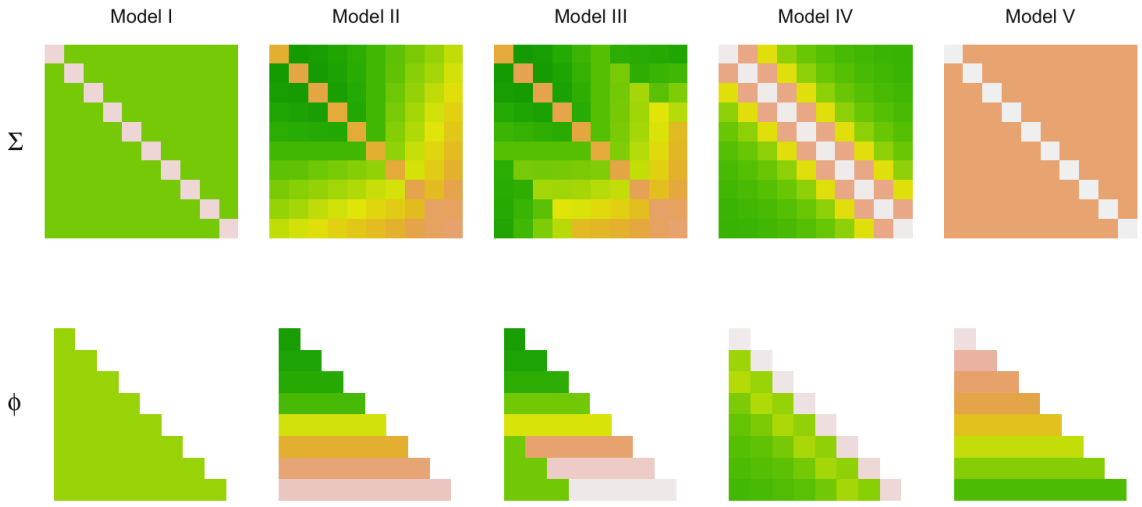


Figure 5.1: *Heatmaps of the true covariance matrices (row 1) under simulation Model I - Model V (see Table 5.1) and the function ϕ defining the corresponding Cholesky factor T (row 2).*

5.1 Loss functions and corresponding risk measures

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

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

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

Σ denotes the true covariance matrix and $\hat{\Sigma}$ is an $M \times M$ positive definite matrix. 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 $M \times M$ 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 $\frac{NS}{N+M+1}$ 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 **??** 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:

$$\phi_{jk} = z'_{jk}\gamma$$

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

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

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

where the orders of the polynomials, p 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 σ_{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

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

The weights ω_{ij}^{k} are controlled by a tuning parameter, k, which can take integer values between 0 and M. 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 in for the class of estimators based on element-wise shrinkage of the sample covariance matrix. However, based on the results of an extensive simulation study presented in Fang et al. (2016), we use $K = 10$-fold cross validation to select the tuning parameters for both the tapering estimator S^{ω} and the soft thresholding estimator S^{λ}. They 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 1, 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 λ, 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 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}, $i = 1, \ldots, N$, $j = 1, \ldots, m_i$. The oracle estimator for Model II is obtained by fitting the model

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

where ϵ_{ij} 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 (3.29) and Example 3.1.2, and the matrix B contains the basis for a linear function of t:

$$\begin{bmatrix} 1 & t_{11} \\ 1 & t_{12} \\ \vdots & \vdots \\ 1 & t_{1,m_1} \\ \vdots & \vdots \\ 1 & t_{N,1} \\ \vdots & \vdots \\ 1 & t_{N,m_N} \end{bmatrix}.$$

The estimator for $\sigma^2(t)$ is then the mean of the squared residuals:

$$\hat{\sigma^2}(t) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{m_i - 1} \sum_{j=1}^{m_i} e_{ij}^2,$$

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 $t_{ij} - t_{ik} < 0.5$:

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

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 $N\left(0, \sigma_\epsilon^2\right)$ distribution, and the b_i are independent $N\left(0, \sigma_b^2 \mathrm{I}\right)$ 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 $M = 10, 20, 30$. 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 ϕ and σ^2 at the fixed observation times. The data are then sampled according to the multivariate normal distribution with covariance matrix $\Sigma = T^{-1} D T'^{-1}$. Given covariance matrix Σ, risk estimates are obtained

from$N_{sim} = 100$ samples from an M-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 $1, 2, \ldots, M$ 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 M, the size of the set of possible measurement times

$$t_1, \ldots, t_M.$$

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

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

where $Y_1, \ldots, Y_N$ are independently and identically distributed according to an M-dimensional multivariate Normal distribution with mean zero and having covariance structure identical to one of Models I - V in 5.1. To induce sparsity, we subsample from the complete data $\{y_i\left(t_j\right)\}$, $i = 1, \ldots, N$, $j = 1, \ldots, M$, 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 of shows the true covariance structure, and each row beneath corresponds to each of the estimators.*

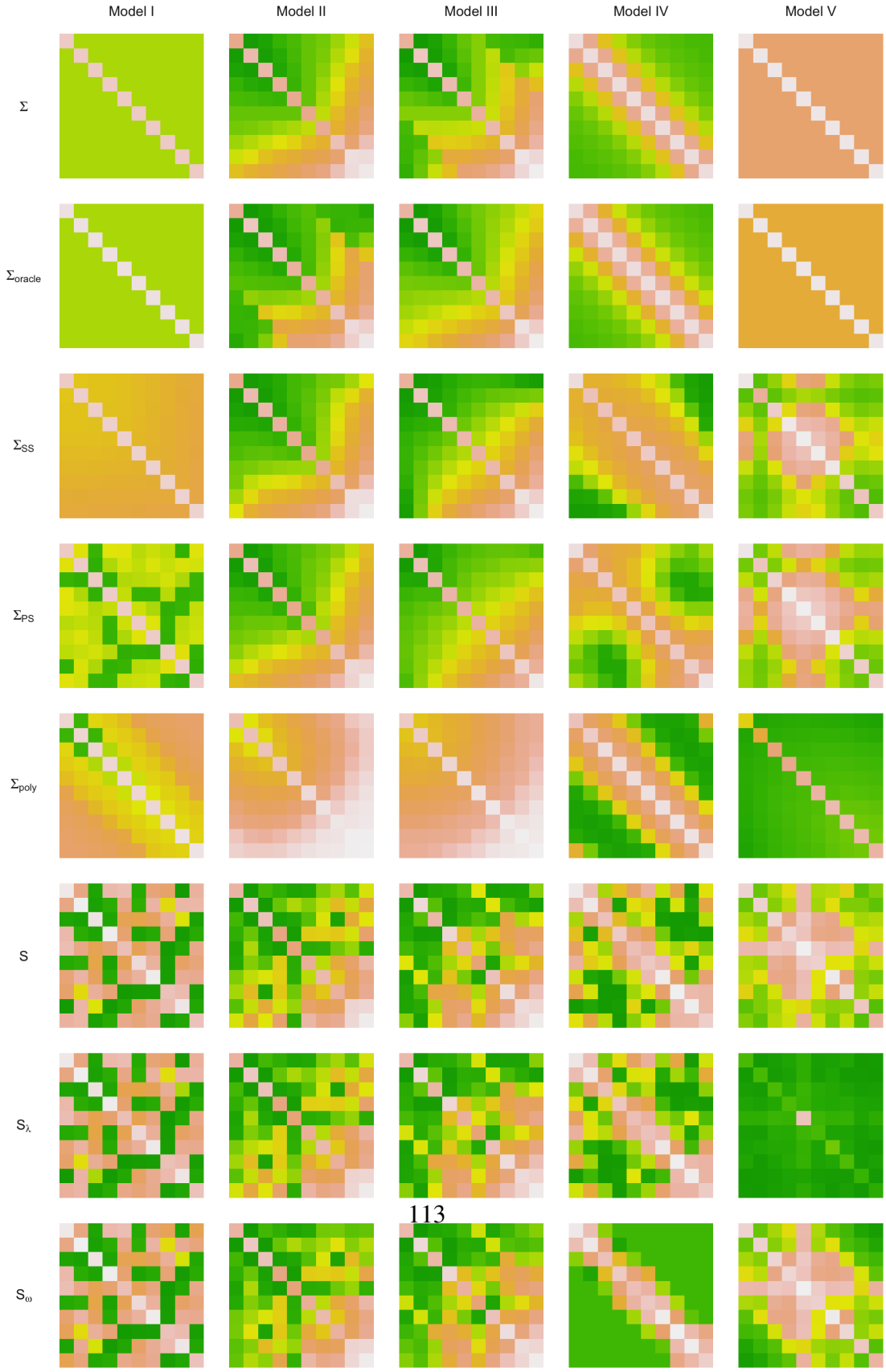


Figure 5.3: *The generalized autoregressive coefficient function ϕ which defines the elements of the true lower triangle of Cholesky factor T corresponding to Model I - Model V and estimates of the same surface for estimators based on the modified Cholesky decomposition. The true covariance structure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

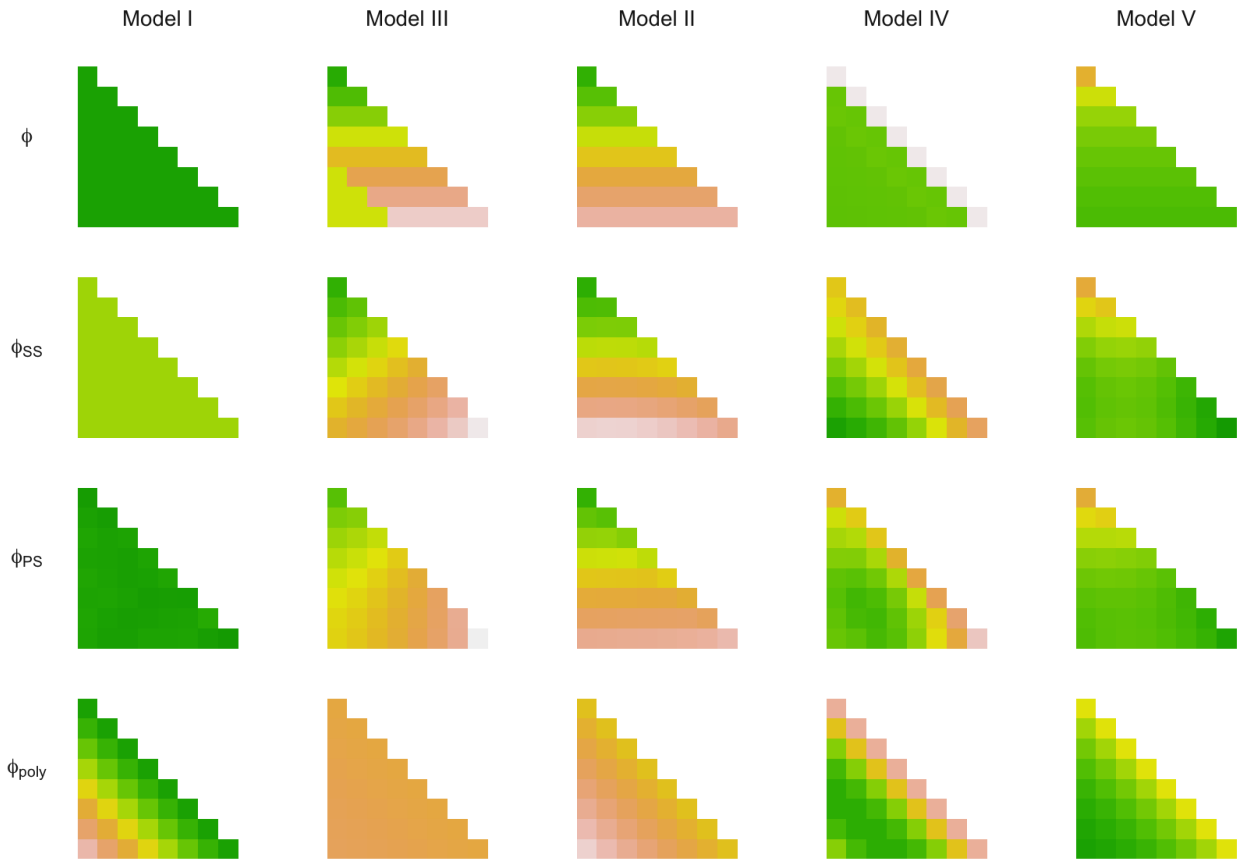
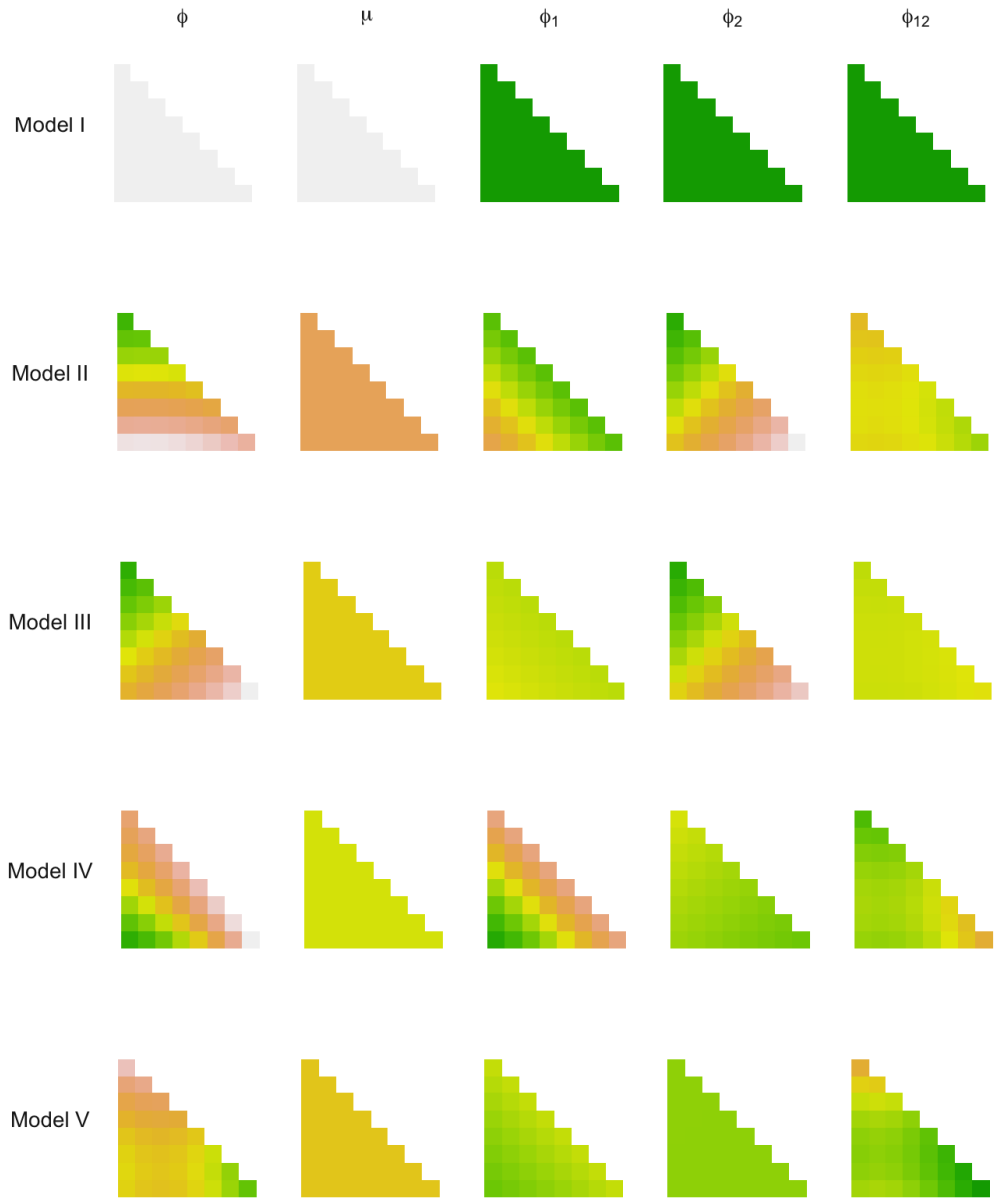


Figure 5.4: *Estimated functional components of the smoothing spline ANOVA decomposition* $\phi = \phi_1 + \phi_2 + \phi_{12}$ *for* $\hat{\Sigma}_{SS}$ *under each simulation model I - V.*



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 M.

In general, our estimators outperform the alternative estimators across the five covariance structures. This is 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. In most cases, 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 selecting the search 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 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.*

	M	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	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.*

	M	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	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.*

	M	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	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.*

	M	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	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.*

	M	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	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

Estimated risk under entropy loss is given in Tables 5.7 - 5.11. Risk estimates under quadratic loss echo in sentiment 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.

Table 5.7: *Model 1: 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.*

M	% missing	$\Delta_2(\hat{\Sigma}^U_{SS})$		$\Delta_2(\hat{\Sigma}^{V^*}_{SS})$	
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 2: Entropy risk estimates and corresponding standard errors.*

M	% missing	$\Delta_2(\hat{\Sigma}^U_{SS})$		$\Delta_2(\hat{\Sigma}^{V^*}_{SS})$	
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 3: Entropy risk estimates and corresponding standard errors.*

M	% missing	$\Delta_2(\hat{\Sigma}^U_{SS})$		$\Delta_2(\hat{\Sigma}^{V^*}_{SS})$	
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 4: Entropy risk estimates and corresponding standard errors.*

M	% missing	$\Delta_2(\hat{\Sigma}^U_{SS})$		$\Delta_2(\hat{\Sigma}^{V^*}_{SS})$	
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 5: Entropy risk estimates and corresponding standard errors.*

M	% 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 cattle weight data

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 $m = 11$ 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.

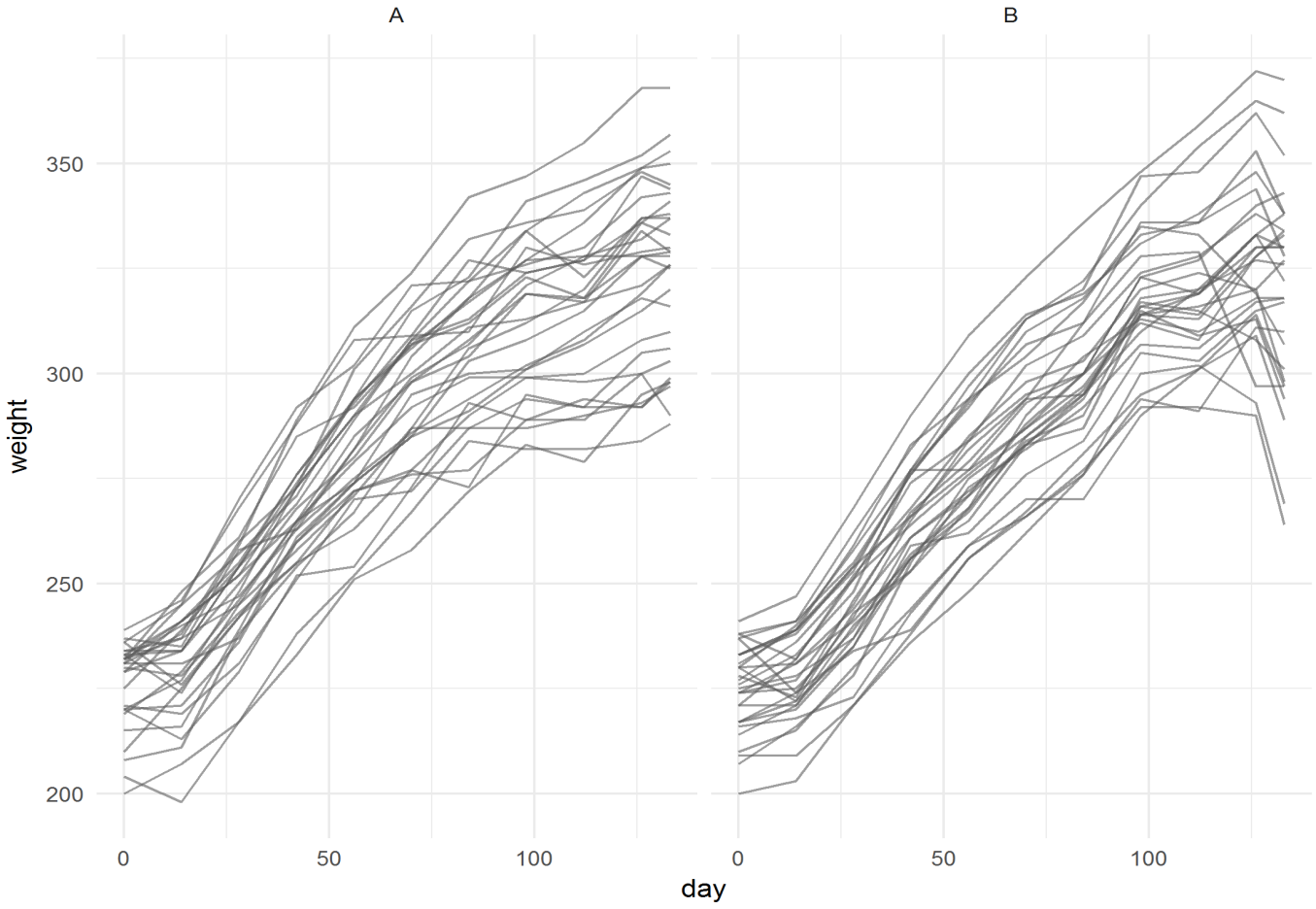


Figure 6.1: *Subject-specific weight curves over time for treatment groups A and B.*

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.

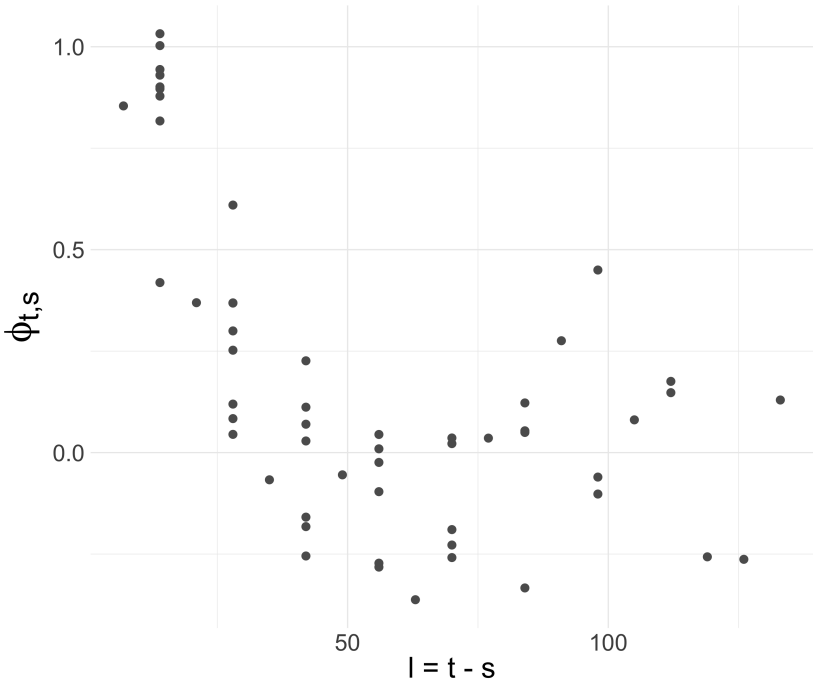
	day										
	0	14	28	42	56	70	84	98	112	126	133
0	1.00										
14	0.82	1.00									
28	0.76	0.91	1.00								
42	0.65	0.86	0.93	1.00							
56	0.63	0.83	0.89	0.93	1.00						
70	0.58	0.75	0.85	0.90	0.94	1.00					
84	0.51	0.64	0.75	0.80	0.85	0.92	1.00				
98	0.52	0.68	0.77	0.82	0.88	0.93	0.92	1.00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

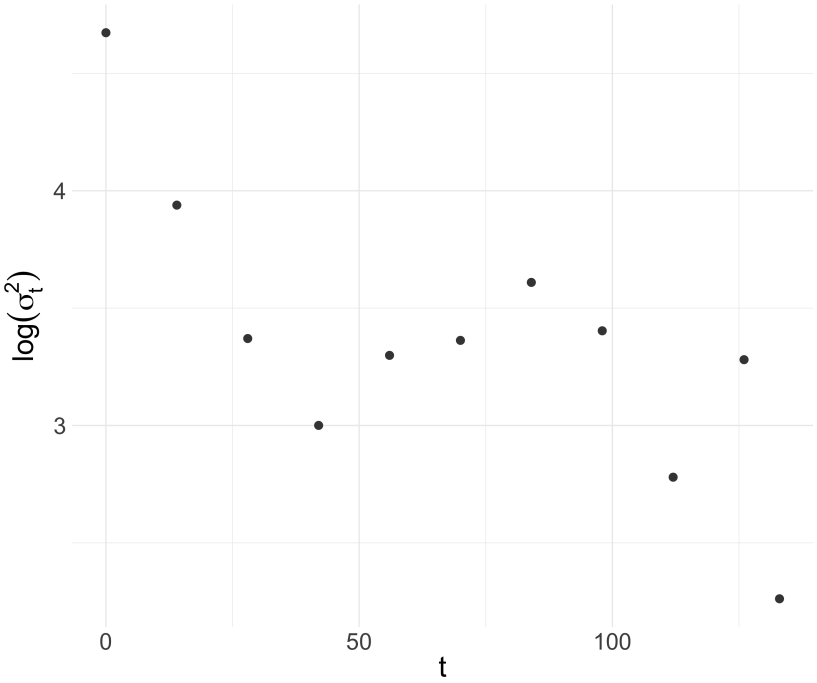
		day											
		0	14	28	42	56	70	84	98	112	126	133	
day	0	1											4.673
	14	1.00	1										3.939
	28	0.04	0.90	1									3.370
	42	-0.25	0.25	0.88	1								3.000
	56	-0.02	0.07	0.12	0.90	1							3.299
	70	0.04	-0.28	0.11	0.37	0.82	1						3.363
	84	0.12	-0.23	0.04	-0.16	0.08	1.03	1					3.610
	98	-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.*

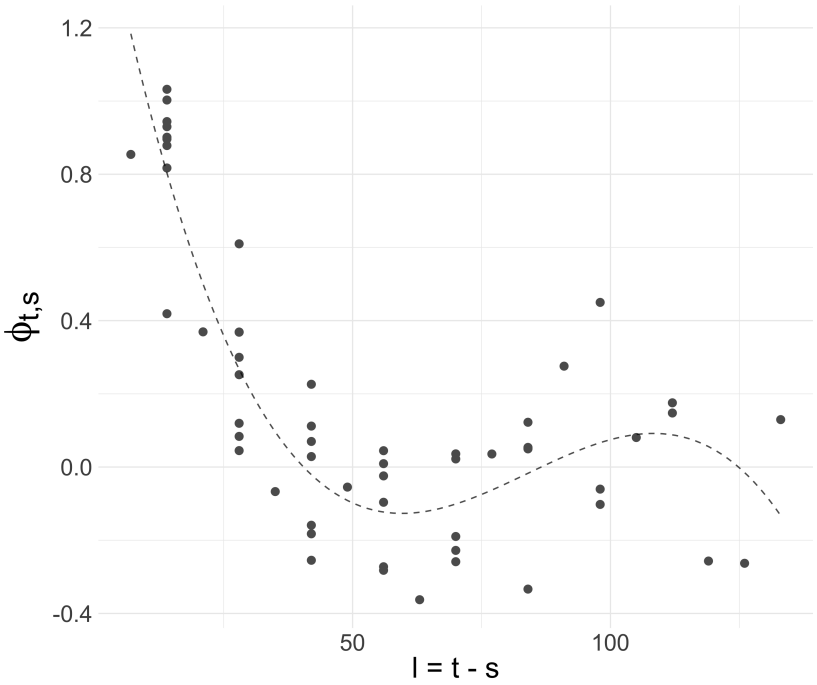
Analyzing the sample regressogram (Figure 6.2a) and sample innovation variogram (Figure 6.2b), Pourahmadi (1999) suggested that both sample generalized autoregressive parameters and the logarithms of the innovation variances can be characterized in terms of cubic functions of the lag only. They model

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

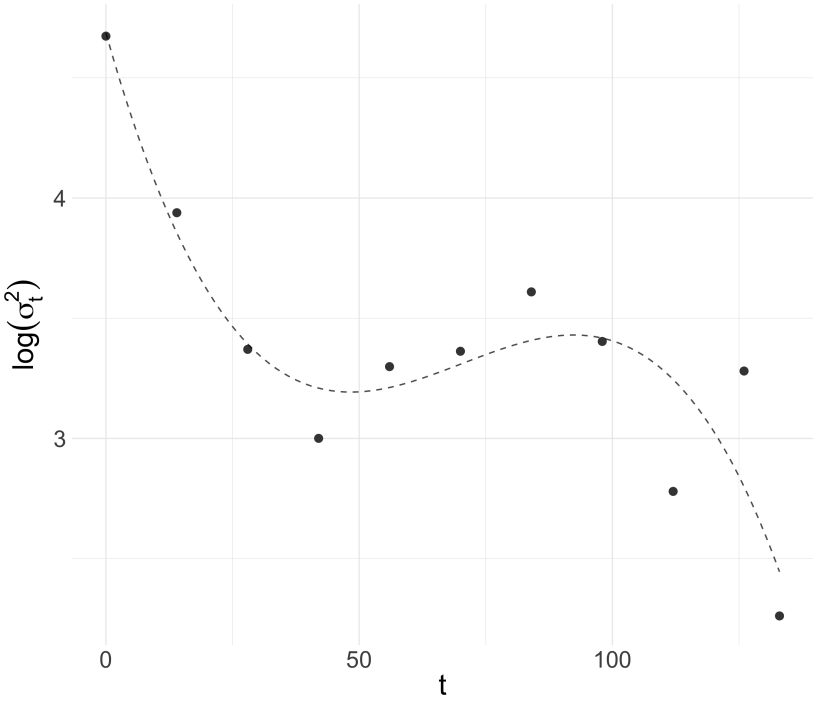
for $t = t_2, \ldots, t_{11}$ where

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

They estimate γ and ξ 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.*

First things first: 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 $m_i \times 1$ 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 $m_i \times q$ design matrix. f is the smooth function of t, and $t_i = \left(t_{i1}, \ldots, t_{i,m_i}\right)'$ is the $m_i \times 1$ 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

$$Z_i b_i = \alpha_i 1_{m_i},$$

so that the random effect corresponds to a subject-specific shift α_i, which are assumed to be independent and identically distributed $N(0, \sigma_\alpha^2)$ random variables. We assume that the $m_i \times 1$ 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

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

We take the estimators of f, $\alpha = (\alpha_1, \ldots, \alpha_N)'$ to minimize the penalized joint log likelihood

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

where $\text{Cov}\left(\alpha\right)=\Sigma_{\alpha}=\sigma_{\alpha}^{2}\text{I}$. 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.

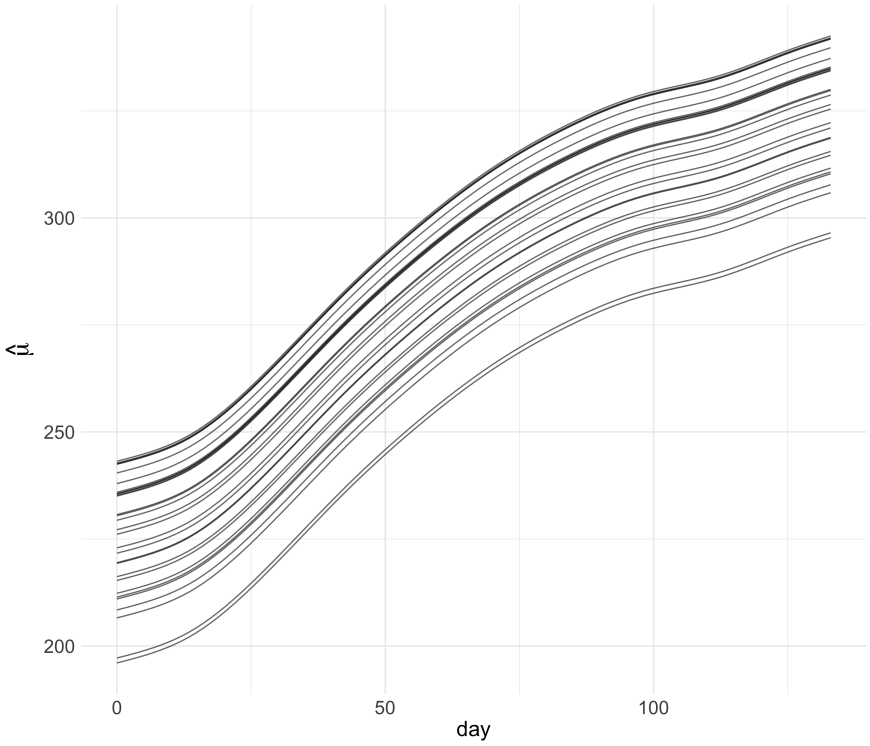


Figure 6.4: *Subject-specific fitted weight trajectories for cattle in treatment group A.*

Centering the data using the fitted mean, the residuals

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

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

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

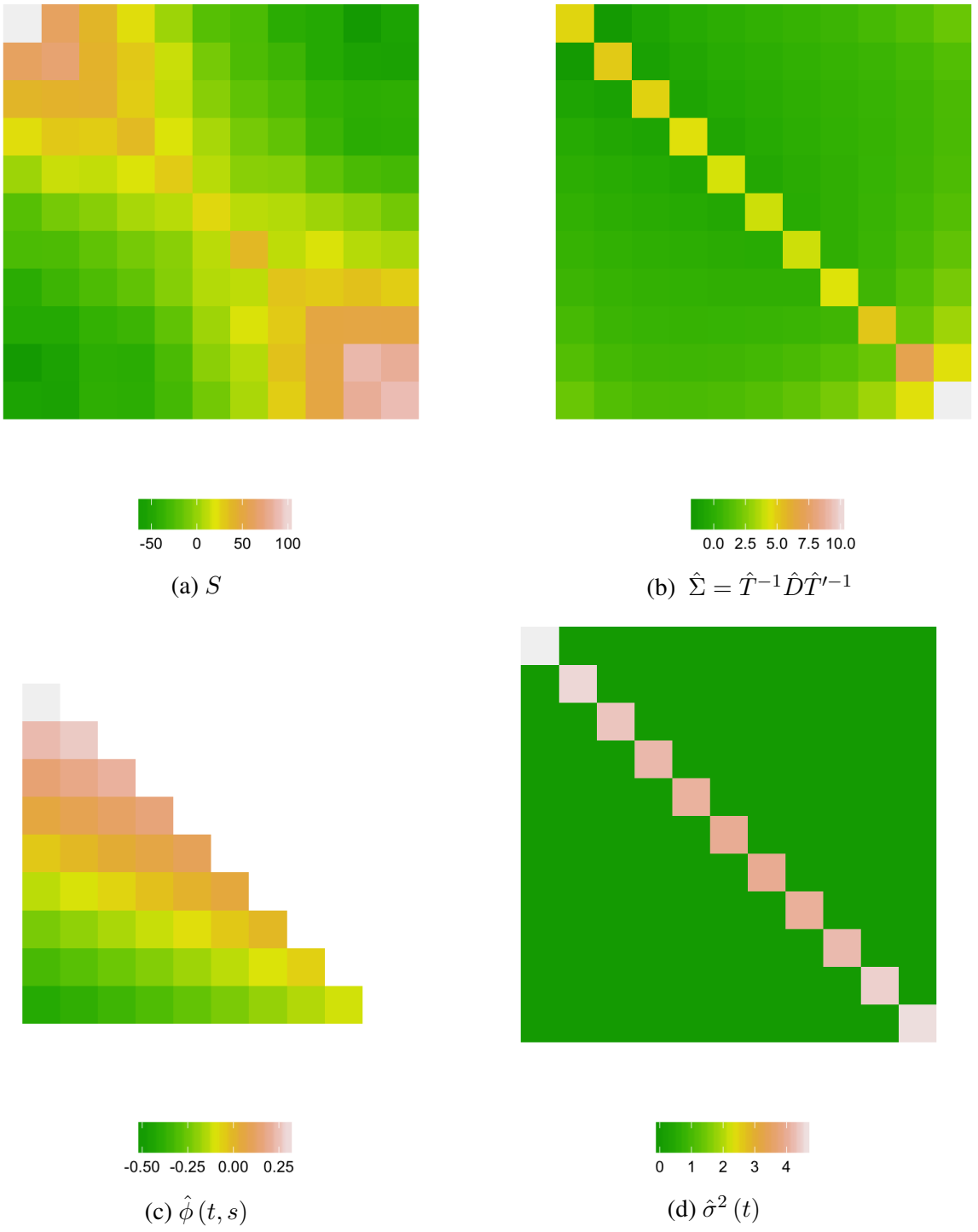
where ϵ is a mean zero gaussian process with unit variance.

Choice of penalty is critical for convergence of the iterative estimation of ϕ and $\log(\sigma_2)$. Pan and Pan (2017) concluded that the regressogram of empirical estimates of $\phi_{t,s}$ show consistent behaviour over $l = t - s$ for each value of t, indicating a lack of a strong functional component of m. 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 $\phi \in \mathcal{H} = \mathcal{H}_{[1]} \otimes \mathcal{H}_{[2]}$, where

$$\mathcal{H}_{[1]} = \left\{ \phi : \ddot{\phi} = 0 \right\} \oplus \left\{ \phi : \phi(0) = \dot{\phi}(0) = 0; \ \int_0^1 \ddot{\phi}^2 \, dx < \infty \right\}$$

$$\mathcal{H}_{[2]} = \left\{ \phi : \phi \propto 1 \right\} \oplus \left\{ \phi : \int_0^1 \phi \, dx = 0, \ \ \dot{\phi} \in \mathcal{L}_2 [0,1] \right\}$$

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 ϕ 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 $\phi(t, s)$ and innovation variance function $\sigma^2(t)$ 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.

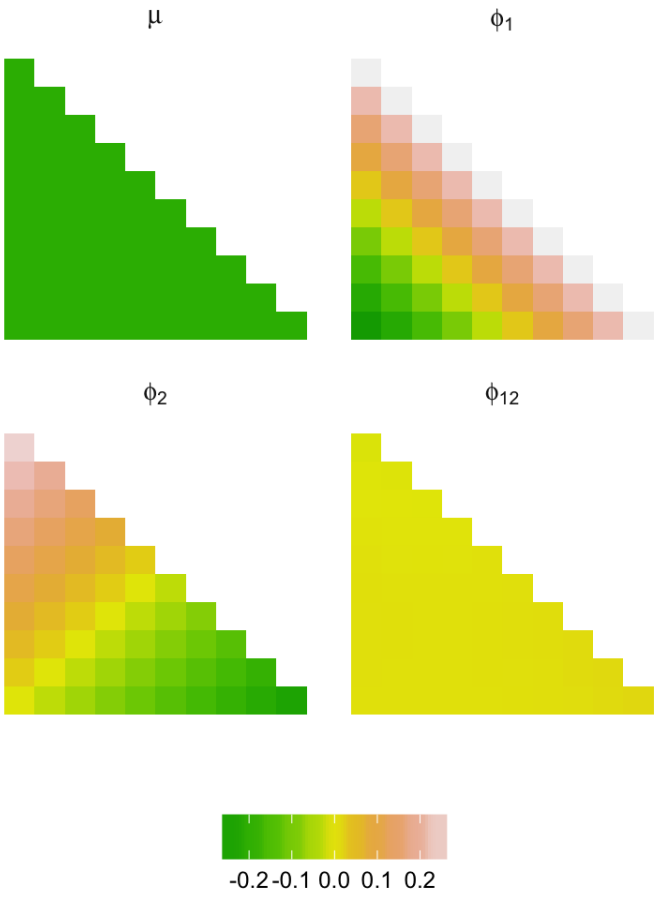


(a) S (b) $\hat{\Sigma} = \hat{T}^{-1}\hat{D}\hat{T}'^{-1}$

(c) $\hat{\phi}(t,s)$ (d) $\hat{\sigma}^2(t)$

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 (see (Klein, 1997), (Madsen, 2007)). Our estimated model instead captures the non-stationarity with both the log innovation variances as well as with $\phi_2]$, the functional component corresponding to the main effect of m. The size of the functional components (in terms of the squared 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, at 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

Our formulation of covariance estimation supplies a flexible framework which free of the impediment presented by the positive definite constraint and a statistically intuitive interpretation of the elements of a covariance matrix. Modeling the Cholesky decomposition rather than the covariance matrix itself allows us to reframe covariance estimation as a regression problem. By estimating the parameters of the corresponding regression model using bivariate smoothing, we naturally accommodate irregularly-spaced longitudinal data of varying within-subject sample sizes by without the need for data imputation methods.

We propose two representations of $(\phi, \log \sigma^2)$ the functional components of the Cholesky decomposition. The first leverages reproducing kernel Hilbert space methods to model the functional component corresponding to the generalized autoregressive

- functional component selection via the non-negative garrote
- In the case with P-splines, explore the performance of generalized additive models analogous to the SSANOVA models without an interaction term.
  - Under a tensor-product model $B = B_l \otimes B_m$, because the support of ϕ lies on the triangle $0 \leq s < t \leq 1$ rather than on a rectangle, the tensor product basis must be trimmed to omit any pairs of knots outside the domain. Without doing this, estimation of the

basis coefficients becomes very unstable. But trimming the basis removes components necessary for each of the marginal bases to possess the properties of a B-spline basis, so functional components corresponding to l and m are unidentifiable.

- Omission of the interaction term permits estimation by restricted maximum likelihood, where the tuning parameters can be interpreted as variances of random effects.
- Three-dimensional B-splines for smoothing over non-rectangular domains has been utilized in the field of graphics and computer vision, but hasn't received much attention in nonparametric statistical modeling - presumably due to how nasty the associated math is.

Appendix A: Chapter 2 Appendix

A.1 Proof of Theorem 3.1.1

Proof. Some detailed discussion of the components of the minimizer ϕ_λ is useful for the proof. Letting $\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_J$ with $\mathcal{H}_0 \perp \mathcal{H}_J$, with reproducing kernel

$$Q\left(\boldsymbol{v}, \boldsymbol{v}^*\right) = Q_0\left(\boldsymbol{v}, \boldsymbol{v}^*\right) + Q_J\left(\boldsymbol{v}, \boldsymbol{v}^*\right),$$

where Q_0 is the RK for $\mathcal{H}_0$ and Q_J is the RK for $\mathcal{H}_J$. Then

$$\begin{aligned}
\xi_i\left(\boldsymbol{v}\right) &= \langle \xi_i, Q_{\boldsymbol{v}} \rangle_{\mathcal{H}} \\
&= \langle P_J \psi_i, Q_{\boldsymbol{v}} \rangle_{\mathcal{H}} = \langle \psi_i, P_J Q_{\boldsymbol{v}} \rangle_{\mathcal{H}} \\
&= \langle \psi_i, Q_{J\boldsymbol{v}} \rangle_{\mathcal{H}} \\
&= L_i Q_{J\boldsymbol{v}}.
\end{aligned}$$

where $Q_{J\boldsymbol{v}}$ is the representer for the evaluation functional at $\boldsymbol{v}$ in $\mathcal{H}_J$. Since $\langle \psi_i - \xi_i, \xi_j \rangle = 0$,

$$\langle \xi_i, \xi_j \rangle_{\mathcal{H}} = \langle \psi_i, \xi_j \rangle_{\mathcal{H}}.$$

Therefore, we have that

$$\langle \xi_i, \xi_j \rangle_{\mathcal{H}} = L_i \xi_j = L_{i(\boldsymbol{v})} L_{j(\boldsymbol{v}^*)} Q_J\left(\boldsymbol{v}, \boldsymbol{v}^*\right).$$

Let the minimizer ϕ_λ be of the form

$$\phi_\lambda = \sum_{\nu=1}^{d_0} d_\nu \eta_\nu + \sum_{\boldsymbol{v}_i \in V} c_i \xi_i + \rho, \tag{A.1}$$

where $\rho \in \mathcal{H}_J$ is perpendicular to $\eta_1, \ldots, \eta_{d_0}, \xi_1, \ldots, \xi_{|V|}$. The properties of reproducing kernel Hilbert spaces give us that any element in $\mathcal{H}$ permits such a representation. Using that

$$\begin{aligned}
L_{ijk}\phi &= \langle \phi(\cdot), Q(\boldsymbol{v}_{ijk}, \cdot) \rangle_{\mathcal{H}} \\
&= \langle \sum_{\nu=1}^{d_0} d_\nu \eta_\nu(\cdot) + \sum_{\boldsymbol{v}_i \in V} c_i \xi_i + \rho(\cdot), Q(\boldsymbol{v}_{ijk}, \cdot) \rangle_{\mathcal{H}} \\
&= \langle \sum_{\nu=1}^{d_0} d_\nu \eta_\nu(\cdot) + \sum_{\boldsymbol{v}_i \in V} c_i \xi_i + \rho(\cdot), Q_0(\boldsymbol{v}_{ijk}, \cdot) \rangle_{\mathcal{H}} + \langle \sum_{\nu=1}^{d_0} d_\nu \eta_\nu(\cdot) + \sum_{\boldsymbol{v}_i \in V} c_i \xi_i + \rho(\cdot), Q_1(\boldsymbol{v}_{ijk}, \cdot) \rangle_{\mathcal{H}} \\
&= \langle \sum_{\nu=1}^{d_0} d_\nu \eta_\nu(\cdot), Q_0(\boldsymbol{v}_{ijk}, \cdot) \rangle_{\mathcal{H}} + \langle \sum_{\boldsymbol{v}_i \in V} c_i \xi_i, Q_0(\boldsymbol{v}_{ijk}, \cdot) \rangle_{\mathcal{H}} + \langle \rho(\cdot), Q_0(\boldsymbol{v}_{ijk}, \cdot) \rangle_{\mathcal{H}} \\
&\quad + \langle \sum_{\nu=1}^{d_0} d_\nu \eta_\nu(\cdot), Q_1(\boldsymbol{v}_{ijk}, \cdot) \rangle_{\mathcal{H}} + \langle \sum_{\boldsymbol{v}_i \in V} c_i \xi_i, Q_1(\boldsymbol{v}_{ijk}, \cdot) \rangle_{\mathcal{H}} + \langle \rho(\cdot), Q_1(\boldsymbol{v}_{ijk}, \cdot) \rangle_{\mathcal{H}},
\end{aligned}$$

where $\langle \rho(\cdot), Q_0(\boldsymbol{v}_{ijk}, \cdot) \rangle_{\mathcal{H}} = \langle \rho(\cdot), Q_1(\boldsymbol{v}_{ijk}, \cdot) \rangle_{\mathcal{H}} = 0$. Thus, substituting (A.1) for ϕ into the penalized sums of squares, the objective function (3.15) becomes

$$\sum_{i=1}^{N} \sum_{j=2}^{m_i} \sigma_{ij}^{-2} \left(y_{ij} - \sum_{k<j} (L_{ijk}\phi) y_{ik} \right)^2 + \lambda \left(|| \sum_{\boldsymbol{v}_i \in V} c_i \xi_i ||_{\mathcal{H}}^2 + ||\rho||_{\mathcal{H}}^2 \right),$$

which is obviously minimized when $||\rho||^2 = 0$.

□

Appendix B: Chapter 4

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 study 1

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.*

	M	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	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.*

	M	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	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.*

	M	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	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.*

	M	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	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	M	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{poly}$	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 study 2

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.*

M	% 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.*

M	% 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.*

M	% 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.*

M	% 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.*

M	% missing	$\Delta_1(\hat{\Sigma}^U_{SS})$		$\Delta_1(\hat{\Sigma}^{V^*}_{SS})$	
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 study 1

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	M	$\hat{\Sigma}^{ure}_{SS}$		$\hat{\Sigma}^{ure}_{PS}$		$\hat{\Sigma}_{oracle}$		$\hat{\Sigma}_{poly}$		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	M	$\hat{\Sigma}^{ure}_{SS}$		$\hat{\Sigma}^{ure}_{PS}$		$\hat{\Sigma}_{oracle}$		$\hat{\Sigma}_{poly}$		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.

Antoniadis, A. and J. Fan (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association 96*(455), 939–967.

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.

Chen, Z., M. Shi, G. W., and M. Tang (2011). Efficient semiparametric estimation via cholesky decomposition for longitudinal data. *Computational Statistics and Data Analysis 55*, 677–690.

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.

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.

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 ??-spline bases built better by ??-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.

Diggle, P. (2002). *Analysis of longitudinal data*. Oxford University Press.

Edgeworth, F. Y. (1892). Xxii. correlated averages. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 34*(207), 190–204.

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.

Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. CRC press.

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.

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

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.

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.

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

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.

Meinhausen, N. and P. Buhlmann (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics 34*(3), 1436–1462.

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.

Peng, J., P. Wang, N. Zhou, and J. Zhu (2012). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*.

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.

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.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association 88*(422), 486–494.

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*.

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. Asymptotic optimality and efficient computation of the leave-subject-out cross-validation: Supplementary materials. *arXiv preprint math.PR/0000000*.

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.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association 93*(441), 120–131.

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.

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.